

On the distribution of subtree orders of a tree*

Dimbinaina Ralaivaosaona, Stephan Wagner[†]

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

Received 17 December 2015, accepted 15 March 2017, published online 22 July 2017

Abstract

We investigate the distribution of the number of vertices of a randomly chosen subtree of a tree. Specifically, it is proven that this distribution is close to a Gaussian distribution in an explicitly quantifiable way if the tree has sufficiently many leaves and no long branchless paths. We also show that the conditions are satisfied asymptotically almost surely for random trees. If the conditions are violated, however, we exhibit by means of explicit counterexamples that many other (non-Gaussian) distributions can occur in the limit. These examples also show that our conditions are essentially best possible.

Keywords: Subtrees, normal distribution, homeomorphically irreducible trees, random trees.

Math. Subj. Class.: 05C05

1 Introduction

By a subtree of a tree, we mean any nonempty connected subgraph; obviously, such a subgraph is again a tree. The distribution of the number of vertices of a randomly chosen subtree of a tree was first studied by Jamison in two papers [5, 6], in which he investigates the average subtree order of a tree, i.e. the mean number of vertices of a subtree. Among his main results is the fact that the average order of subtrees of an n -vertex tree is at least $(n+2)/3$, with equality only for the path. The problems that Jamison proposed in his papers received considerable attention recently [4, 14, 16], as did other aspects of subtrees in trees, specifically extremal problems, whose study was initiated by Székely and Wang [12, 13]. Jamison's question whether the average order is always at least $n/2$ for homeomorphically irreducible trees, i.e. trees without vertices of degree 2, was only answered (affirmatively) very recently by Vince and Wang [14], who also showed that the average subtree order of such a tree is less than $3n/4$.

*We would like to thank an anonymous referee for many useful comments.

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

E-mail addresses: naina@sun.ac.za (Dimbinaina Ralaivaosaona), swagner@sun.ac.za (Stephan Wagner)

Many other of Jamison's questions remain open to date. A question of his that was also discussed in the 2011 edition of the Combinatorics REGS [7] reads as follows:

Question 1.1. Given a tree T of order n , let $s_k(T)$ denote the number of subtrees of order k . When is it true that the numbers $s_2(T), \dots, s_n(T)$ form a unimodal list (weakly increasing at first, then weakly decreasing)? In particular, is it unimodal when T has no vertices of degree 2?

It should be noted here that $s_1(T) = n$ and $s_2(T) = n - 1$ for every tree T of order n , so $s_1(T)$ cannot be included if a unimodal list is to be obtained. The question seems to be fairly hard, and we will not actually answer it in this paper. However, we provide a related result: if a tree has sufficiently many leaves and no long branchless paths (this will be made more precise later), then the distribution of the subtree orders is close to a Gaussian distribution in an explicitly quantifiable way. In particular, this is the case for trees without vertices of degree 2. Moreover, the conditions we impose are usually satisfied: for random trees, they are valid asymptotically almost surely.

Asymptotic normality of the distribution does of course not imply unimodality, nor the other way around, but the two are clearly connected, so our result provides evidence that the answer to Question 1.1 might be affirmative. It should also be pointed out that our main result parallels a classic theorem of Godsil [3] on matchings: if $G_1, G_2, \dots$ is a sequence of graphs, then the distribution of the size of matchings in G_n (suitably renormalised) converges to a Gaussian distribution, provided that the variance tends to infinity. See [8] for a recent extension.

Godsil's theorem is based on properties of the matching polynomial, in particular the fact that all its zeros are real. Indeed, it is well known that a polynomial with positive coefficients and only real zeros has log-concave (and thus unimodal) coefficients, so Question 1.1 could be answered affirmatively if all zeros of the polynomial

$$\sum_{k=2}^n s_k(T)u^k$$

were real for every T . This “subtree polynomial” was already considered by Jamison himself in [5]. More recently, Yan and Yeh [18] studied a weighted version, and Martin et al. [9] considered a bivariate generalisation involving the number of leaves.

Unfortunately, the subtree polynomial does not have real roots only as the matching polynomial does, so the situation for subtrees of trees turns out to be more intricate than for matchings of graphs. As a simple concrete counterexample, consider the star S_4 with four vertices: we have

$$\sum_{k=1}^4 s_k(S_4)u^k = 4u + 3u^2 + 3u^3 + u^4,$$

a polynomial with two non-real roots. Even if the first coefficient is removed, we get

$$\sum_{k=2}^4 s_k(S_4)u^k = 3u^2 + 3u^3 + u^4,$$

which still has two non-real roots.

However, we obtain a central limit theorem for the distribution of subtree orders analogous to Godsil's theorem under some technical conditions. Our approach is of a rather different nature, and we hope that it might also prove useful to deal with other problems, such as a conjecture of Alavi, Malde, Schwenk and Erdős [1] concerning the independence polynomial of trees that parallels Question 1.1. Our main theorem can be stated as follows:

Theorem 1.2. *Let $T_1, T_2, \dots$ be a sequence of trees such that $|T_n|$, i.e. the number of vertices of T_n , goes to infinity, the proportion of leaves among all vertices is bounded below by a positive constant, and the length of the longest branchless path in T_n is at most $|T_n|^{1/2-\epsilon}$ for some fixed ϵ (and sufficiently large n). Then the order distribution of the subtrees of T_n (suitably renormalised) converges weakly to a Gaussian distribution.*

It is easy to find both examples and counterexamples for the normal distribution: for instance, if T_n is an n -vertex star, then the distribution of the subtree orders is essentially a binomial distribution, which converges to a Gaussian law. On the other hand, if one considers the sequence of n -vertex paths, then the limit distribution is quite different. This and other examples and counterexamples will be discussed in Section 2, where we also show that the technical conditions of Theorem 1.2 are indeed important and also essentially best possible.

The main part of the paper is organised as follows: we first obtain some auxiliary results and prove two versions of our main theorem (a central and a local limit theorem, see Theorem 4.8 and Theorem 4.10 respectively) for rooted trees in Section 4 before passing on to unrooted trees in Section 5. Rooted trees are more accessible because one can use a recursive approach, and we will see that an appropriate root can always be chosen in such a way that most subtrees contain the root. In Section 6, we will see that in the “generic” case of random trees, the conditions of our main theorem are satisfied, so that the Gaussian distribution is indeed the typical limit distribution of subtree orders.

Notation. Throughout this paper, we make frequent use of the Vinogradov symbol $\ll$ interchangeably with the $\mathcal{O}$ -notation: $f(T) \ll g(T)$ or $f(T) = \mathcal{O}(g(T))$ means that $f(T) \leq Kg(T)$ for a suitable positive constant K and all (sufficiently large) trees T . If further variables are included in an $\mathcal{O}$ -term, the constant K is always independent of those, unless mentioned otherwise.

2 Examples and counterexamples

For a tree T , we let $\mathcal{S}(T)$ denote the set of all subtrees of T , i.e. all connected induced subgraphs of T . The polynomial associated with this set, which we call the *subtree polynomial* of T , is denoted by $S(T, u)$:

$$S(T, u) = \sum_{\tau \in \mathcal{S}(T)} u^{|\tau|}.$$

The total number of subtrees is clearly $S(T, 1)$, for which we will simply write $S(T)$. Our goal will be to prove central and local limit theorems for the coefficients of this polynomial. Note also that $S_u(T, 1) = \frac{\partial}{\partial u} S(T, u)|_{u=1}$ is the total number of vertices in T 's subtrees, so $S_u(T, 1)/S(T)$ is the mean subtree order. Likewise, the variance is given by

$$\frac{S_{uu}(T, 1) + S_u(T, 1)}{S(T)} - \left(\frac{S_u(T, 1)}{S(T)} \right)^2. \quad (2.1)$$

Before we get to the proof of the main theorem, let us briefly discuss some examples and counterexamples to illustrate its statement.

2.1 The star

If $T = S_n$ is a star of order n , then every subtree either consists of the centre and an arbitrary set of leaves, or it is a single leaf. Thus we have

$$S(T, u) = nu + \sum_{k=2}^n \binom{n-1}{k-1} u^k$$

and in particular $S(T) = 2^{n-1} + n - 1$. We see that the distribution of subtree orders is essentially a binomial distribution, which gives us a Gaussian distribution in the limit.

2.2 The path

The distribution of subtree orders of a path P_n turns out to be quite different: every subtree is again a path and uniquely characterised by its endpoints. We obtain

$$S(P_n, u) = \sum_{k=1}^n (n - k + 1) u^k.$$

If we divide the subtree orders by n and take the limit, we obtain a distribution whose density is given by $f(t) = 2(1 - t)$ on the interval $[0, 1]$.

The examples that we consider in the following are all constructed by suitably combining paths and stars. Depending on how this is done, a variety of different limit distributions can be obtained. Of course, there does not even have to be a limit distribution at all: this is not the case, for example, if we consider a sequence of trees of increasing orders, alternating between paths and stars.

2.3 The broom

The simplest possible combination of a star and a path is the broom, consisting of a path of k vertices and ℓ leaves attached to one of its ends (the “centre” of the broom, denoted v in Figure 1). Here, the limit as $k + \ell \rightarrow \infty$ depends very much on the relative sizes of k and ℓ . If k is fixed, then there is very little difference to a star, and we obtain a Gaussian limit distribution. On the other hand, if ℓ is fixed, then we have essentially the same order distribution as for a path (and exactly the same in the limit). As soon as ℓ grows faster than $\log_2 k$, almost all subtrees contain the broom’s centre v (i.e., the proportion of such subtrees tends to 1). This is because there are $k2^\ell$ subtrees containing it, as opposed to $\mathcal{O}(k^2 + \ell)$ not containing it.

Subtrees containing the centre v have a distribution that is a convolution of a binomial distribution (stemming from the leaves attached to v) and a discrete uniform distribution (stemming from the path). In the limit, the distribution with greater variance dominates. Since the variances are of order k^2 and ℓ respectively, we have three phases:

- (i) $k^2/\ell \rightarrow 0$: the leaves dominate, and a suitably renormalised version of the order distribution converges to a normal distribution.
- (ii) $k^2/\ell \rightarrow a > 0$: the limit distribution is a convolution of a (continuous) uniform distribution and a Gaussian distribution.

- (iii) $k^2/\ell \rightarrow \infty$ (but $k/2^\ell \rightarrow 0$): the long path dominates, and the renormalised order distribution converges to a uniform distribution.

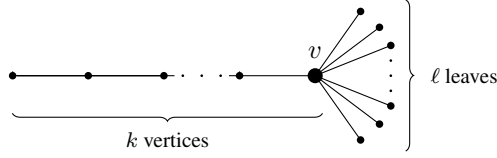


Figure 1: The broom.

2.4 The extended star

Figure 2 shows an extended star, obtained by attaching d (≥ 3) paths of length k to a common vertex v . For fixed d , we obtain (by the same argument as in the previous example) a convolution of d uniform distributions in the limit as $k \rightarrow \infty$. As soon as d also tends to infinity, however, the limit is Gaussian again (showing that the conditions of Theorem 1.2 are important, but not strictly necessary).

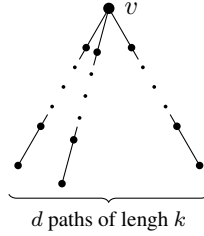


Figure 2: The extended star.

2.5 A discontinuous limit distribution

By suitably choosing the parameters of a double-star (see Figure 3), we can even obtain a discontinuous limit distribution. Such a tree consists of a path of length k and leaves attached to the two endpoints v_1 and v_2 (ℓ and r leaves, respectively). We set $\ell = 3n$, $r = n + c$ for some constant c , and $k = 2^n$. The same argument that we used for the broom shows that almost all subtrees contain v_1 in this case. The probability that v_2 is contained as well is easily found to be $2^c/(1 + 2^c)$ in the limit. In this case, the subtree order is $2^n + \mathcal{O}(n)$. Otherwise, it essentially follows a discrete uniform distribution on the interval $[1, 2^n]$ (the leaves attached to v_1 only playing a minor role). So if we divide the subtree orders by 2^n , we obtain a limit distribution that is a mix of the uniform distribution on $[0, 1]$ and a point measure at 1, which means that its distribution function has a discontinuity at 1.

We remark that another choice of parameters is interesting as well: if we set $\ell = r = 3n$ and $k = 2^n$, then almost all subtrees contain both v_1 and v_2 (and the probability that this is not the case is as low as $\mathcal{O}(4^{-n})$). Thus the subtree order distribution is essentially a convolution of two binomial distributions, and the variance is $\mathcal{O}(n)$. This shows that

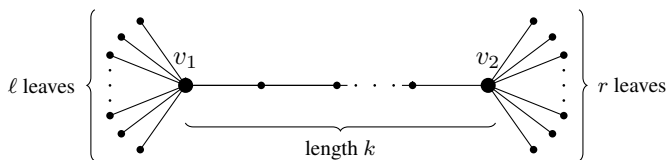


Figure 3: The double-star.

the variance of the subtree order distribution can be as low (in order of magnitude) as the logarithm of the order of the underlying tree, and we conjecture that it cannot be less, i.e. that (2.1) is bounded below by $K \log |T|$ for some positive constant K . On the other hand, the order of magnitude of the variance can be as high as $|T|^2$, as the example of the path shows.

2.6 Short branchless paths are insufficient

The two conditions of Theorem 1.2 (short branchless paths, many leaves) ensure that the trees T_n are not too “path-like”. However, as we exhibit now, neither of the two conditions suffices on its own to ensure a Gaussian limit distribution. The broom is a simple example showing that even a proportion of leaves tending to 1 may not be enough: if we choose k and ℓ such that $\ell = ak^2$ for some fixed constant a , then we obtain a convolution Gaussian-Uniform in the limit rather than a pure normal distribution. This example also explains why $\sqrt{|T_n|}$ is the threshold for the length of branchless paths.

Finding a counterexample that satisfies the condition on paths, but does not have sufficiently many leaves, is a little bit more complicated. It can be constructed as follows (see Figure 4): fix positive constants α, β, γ such that $\beta < \alpha < \frac{1}{2}$, $\alpha + \gamma = 1$ and $2\alpha > \beta + \gamma$. Start with a central vertex v , which is connected to $\ell + 1 = \lfloor n^\gamma \rfloor$ vertices $w_0, w_1, w_2, \dots, w_\ell$ by paths of length $\lfloor n^\alpha \rfloor$. To each of these vertices except w_0 , we attach $\lfloor n^\beta \rfloor$ leaves. Note that the order of the resulting tree T_n is $|T_n| \sim n^\alpha \cdot n^\gamma = n$, so that there are no branchless paths of length $|T_n|^{1/2-\epsilon}$ if $\epsilon < \frac{1}{2} - \alpha$ and n is sufficiently large. On the other hand, the number of leaves is $L(T_n) \sim n^\beta \cdot n^\gamma = o(n)$ (note, however, that the exponent $\beta + \gamma$ can be made arbitrarily close to 1 with an appropriate choice of α, β, γ).

The limit distribution is not Gaussian in this case: the same argument that we used in previous examples shows that $v, w_1, w_2, \dots, w_\ell$ and thus also the paths connecting them are part of almost all subtrees. The remaining “random” part is the same as for a broom consisting of a path of length (approximately) n^α and (approximately) $n^{\beta+\gamma}$ leaves. Since $2\alpha > \beta + \gamma$ by our choice, we are in the situation where the limit distribution as $n \rightarrow \infty$ is uniform.

3 Preliminary results

Before we start with the actual proof of our main result, let us fix some notation and prove some auxiliary inequalities.

3.1 Definitions and notation

Most of the time, we will be working with rooted trees, since they allow for a recursive approach. Thus we first define an analogue of the polynomial $S(T, u)$ for rooted trees.

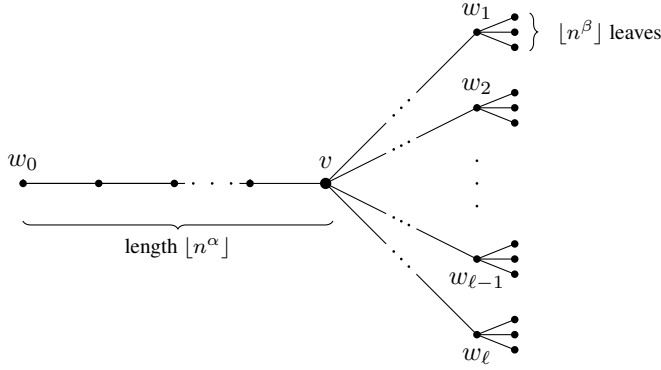


Figure 4: The final counterexample.

Consider a tree T with root v_0 , and let $\mathcal{S}^\bullet(T)$ be the set of all subtrees of T containing v_0 . The generating polynomial for subtrees containing the root is denoted by $S^\bullet(T, u)$:

$$S^\bullet(T, u) = \sum_{\tau \in \mathcal{S}^\bullet(T)} u^{|\tau|}.$$

The main reason for considering this polynomial is the fact that it can be computed recursively from the root branches. For a vertex v of T , we let $T(v)$ be the branch of T rooted at v (consisting of v and all its descendants). Suppose that $v_1, v_2, \dots, v_d$ are the root's children. It is not hard to see that $S^\bullet(T, u)$ satisfies the following recursive formula:

$$S^\bullet(T, u) = S^\bullet(T(v_0), u) = u \prod_{j=1}^d (1 + S^\bullet(T(v_j), u)). \quad (3.1)$$

This follows from the fact that a subtree of T that contains the root v_0 induces either the empty tree or a subtree that contains v_j in the branch $T(v_j)$ for each v_j .

Notation. For the convenience of the reader we list some further notation that is used throughout this paper:

- (i) $\mathcal{L}(T)$ and $L(T)$ are the set and the number of leaves, respectively.
- (ii) $\mathcal{I}(T)$ and $I(T)$ are the set and the number of interior vertices, respectively.
- (iii) By a branchless path or 2-path, we mean a path in which all vertices, except for the endpoints, have degree 2. We let $P(T)$ denote the maximum length of a 2-path of T .

Moreover, we use $c_0, c_1, c_2, \dots$ to denote absolute constants (that do not depend on the specific tree or any of its parameters).

3.2 Two inequalities

We begin with the following simple but useful lemma, which provides two inequalities that will be used repeatedly in the following section.

Lemma 3.1. *If T is a rooted tree with $|T| \geq 2$, then*

$$S^\bullet(T) \geq 2^{L(T)} \text{ and } L(T) \geq \frac{|T|}{2P(T)}.$$

Proof. Every subset A of $\mathcal{L}(T)$ gives rise to a subtree obtained as the union of all paths connecting the leaves in A to the root. If A is empty, we take the subtree consisting only of the root as the corresponding subtree. This proves the first inequality.

For the proof of the second inequality, we let $V_2(T)$ be the number of non-root vertices of degree 2 and $V_{\geq 3}(T)$ the number of non-root vertices of degree at least 3. Consider all maximal 2-paths (not containing the root as an inner vertex in case that the root has degree 2). To each such path, we can uniquely associate its endpoint that is further away from the root, which is either a leaf or a (non-root) vertex of degree at least 3. Thus there are $L(T) + V_{\geq 3}(T)$ such paths. Since the total number of edges, which is $|T| - 1$, is at most $P(T)$ times the number of maximal 2-paths, we obtain

$$(L(T) + V_{\geq 3}(T))P(T) \geq |T| - 1.$$

On the other hand, the handshake lemma gives us

$$2(L(T) + V_2(T) + V_{\geq 3}(T)) = 2(|T| - 1) \geq L(T) + 2V_2(T) + 3V_{\geq 3}(T) + 1,$$

the final 1 being the trivial lower bound for the root degree. Thus $L(T) \geq V_{\geq 3}(T) + 1$, and consequently

$$2L(T)P(T) \geq (L(T) + V_{\geq 3}(T) + 1)P(T) \geq (L(T) + V_{\geq 3}(T))P(T) + 1 \geq |T|,$$

which is equivalent to the second inequality in the statement of the lemma. $\square$

4 Rooted trees

4.1 The moment generating function

In order to prove the central limit theorem for the order distribution of subtrees, we study the associated moment generating function, first only for rooted trees. Note that

$$\frac{S^\bullet(T, u)}{S^\bullet(T, 1)} = \frac{S^\bullet(T, u)}{S^\bullet(T)} = \frac{1}{S^\bullet(T)} \sum_{\tau \in \mathcal{S}^\bullet} u^{|\tau|}$$

is the probability generating function for the order of a randomly chosen subtree of T that contains the root. Likewise,

$$\frac{S^\bullet(T, e^t)}{S^\bullet(T)}$$

is the moment generating function. For our purposes, it turns out to be useful to consider an auxiliary function, denoted $F(T, t)$. We define it recursively by $F(T, t) = \log(1 + e^t)$ if $|T| = 1$ and

$$F(T, t) = \sum_j F(T(v_j), t) + f(T, t), \quad (4.1)$$

where

$$f(T, t) = t + \log \left(1 + \frac{1}{S^\bullet(T, e^t)} \right). \quad (4.2)$$

Here and in the following, $\log$ will always denote the principal branch of the logarithm. In view of (3.1), we have

$$1 + S^\bullet(T, e^t) = e^{F(T, t)}, \quad (4.3)$$

as can be seen by a simple induction. As a first step, we show that $S^\bullet(T, e^t)$ is bounded away from 0 if t is sufficiently small, so that we can actually take the logarithm in (4.2).

Lemma 4.1. *There exist absolute constants $\delta > 0$ and $c_0 > 0$ with the following property: if T is a tree such that the lengths of the 2-paths of T are all bounded above by some positive integer P (which can be a function of T), then we have*

$$|1 + S^\bullet(T, e^t)| \geq e^{c_0 L(T)} \quad (4.4)$$

whenever $|t| \leq \frac{\delta}{P}$. Moreover, the function $f(T, t)$ as defined in (4.2) is analytic in the disk defined by the inequality $|t| \leq \frac{\delta}{P}$.

Remark 4.2. It is important to bear in mind that t is complex in this context. If we were to consider only real values of t , it would e.g. be trivial that $|1 + S^\bullet(T, e^t)| = 1 + S^\bullet(T, e^t) > 1$.

Proof. We will show that the statements of the lemma hold for the following explicit constants:

$$\delta = 0.001 \quad \text{and} \quad c_0 = 0.012.$$

So we assume throughout this proof that δ is as defined above. We show first that the inequality (4.4) implies analyticity of the function $f(T, t)$ in (4.2). Note that

$$1 + \frac{1}{S^\bullet(T, e^t)} = \left(1 - \frac{1}{1 + S^\bullet(T, e^t)}\right)^{-1}.$$

If $|1 + S^\bullet(T, e^t)| \geq e^{c_0 L(T)} \geq e^{c_0}$, then $1 - \frac{1}{1 + S^\bullet(T, e^t)}$ lies inside the disk with centre 1 and radius e^{-c_0} . Thus the reciprocal $\left(1 - \frac{1}{1 + S^\bullet(T, e^t)}\right)^{-1}$ lies inside the disk with centre $\frac{1}{1 - e^{-2c_0}}$ and radius $\frac{e^{-c_0}}{1 - e^{-2c_0}}$. The principal branch of the logarithm is an analytic function inside this disk, so $f(T, t)$ is analytic.

Now we prove (4.4). Let P be an arbitrary positive integer, and let T be a tree such that no 2-path of T has length greater than P . The lemma is satisfied for $|T| = 1$: in this case, we have $S^\bullet(T, e^t) = e^t$, and it is easily verified that

$$|1 + S^\bullet(T, e^t)| = |1 + e^t| \geq 1 + e^{-\delta} > e^{c_0}$$

holds for $|t| \leq \delta$.

If $2 \leq |T| \leq 12P$, then for every subtree τ of T we have $|t||\tau| \leq 12\delta$. It follows that $\operatorname{Re}(e^{t|\tau|}) \geq e^{-12\delta} \cos(12\delta)$. Therefore,

$$\begin{aligned} \left|1 + \sum_{\tau \in \mathcal{S}^\bullet(T)} e^{t|\tau|}\right| &\geq 1 + \sum_{\tau \in \mathcal{S}^\bullet(T)} \operatorname{Re}(e^{t|\tau|}) \\ &\geq e^{-12\delta} \cos(12\delta) (1 + S^\bullet(T)). \end{aligned} \quad (4.5)$$

Applying Lemma 3.1 to estimate the right side of (4.5), we have

$$\begin{aligned} |1 + S^\bullet(T, e^t)| &= \left| 1 + \sum_{\tau \in S^\bullet(T)} e^{t|\tau|} \right| \geq e^{-12\delta} \cos(12\delta) 2^{L(T)} \\ &\geq e^{c_0} 2^{L(T)-1} \geq e^{c_0} e^{c_0(L(T)-1)} = e^{c_0 L(T)}. \end{aligned}$$

So the proof is complete in this case, and we assume from now on that $|T| > 12P$.

For each vertex v of T , we define

$$m(v, t) = |1 + S^\bullet(T(v), e^t)|.$$

Let $v_1, v_2, \dots$ be v 's children. Using (3.1), we find that for $|t| \leq \frac{\delta}{P}$,

$$m(v, t) \geq e^{-\delta/P} \prod_j m(v_j, t) - 1. \quad (4.6)$$

Let A be the set of vertices in “small branches”, defined formally as the set of all vertices w of T for which $|T(w)| \leq 12P$. Thus for every $w \in A$, the bound in (4.5) applies to the branch $T(w)$, and we have

$$m(w, t) = \left| 1 + \sum_{\tau \in S^\bullet(T(w))} e^{t|\tau|} \right| \geq e^{-12\delta} \cos(12\delta) (1 + S^\bullet(T(w))). \quad (4.7)$$

We define $m_0 = 2e^{-12\delta} \cos(12\delta) \approx 1.976$, so we can deduce from (4.7) that for every $w \in A$

$$m(w, t) \geq m_0. \quad (4.8)$$

The rest of the proof is divided into two parts: in the first part, we prove that $m(v, t)$ cannot be too small when v is outside of A . In the second part, we use recursion (4.6) to complete the proof of (4.4).

Part 1: We claim that $m(v, t) \geq 3P$ for all $v \in T \setminus A$.

Assume that the claim is not true, and let $w \in T \setminus A$ be a counterexample (i.e., $m(w, t) < 3P$) with maximum distance from the root. In addition, let $w_0 = w, w_1, \dots, w_r$ be the longest sequence of vertices (possibly, $r = 0$) such that none of these vertices lies in A , w_{j+1} is w_j 's only child for $0 \leq j < r$, and w_r has either more than one child or a single child that lies in A . Now consider two different cases:

- (i) Suppose that all of w_r 's children, which we denote by $x_1, x_2, \dots, x_d$, lie in A . Since $w_r \notin A$, we have $|T(w_r)| > 12P$, so at least one of these children is the root of a branch of order at least $12P/d$. Without loss of generality, $|T(x_1)| \geq 12P/d$. We have

$$m(x_1, t) \geq \frac{m_0}{2} (1 + S^\bullet(T(x_1))) \geq \frac{m_0}{2} (1 + |T(x_1)|) \geq \frac{m_0}{2} \cdot \frac{12P}{d}$$

by (4.5); the inequality $S^\bullet(T(x_1)) \geq |T(x_1)|$ simply follows from the fact that we can associate the path from the root, which is also a subtree, to each vertex. Moreover, we know that $m(x_2, t), \dots, m(x_d, t) \geq m_0$ by (4.8). Now (4.6) gives us

$$m(w_r, t) \geq e^{-\delta/P} \cdot \frac{m_0}{2} \cdot \frac{12P}{d} \cdot m_0^{d-1} - 1 = 6e^{-\delta/P} \cdot \frac{m_0^d}{d} \cdot P - 1.$$

Using the numerical values of δ and m_0 , one easily verifies that $m_0^d \geq d$ for all $d \geq 1$ and $6e^{-\delta/P} \geq 6e^{-\delta} \geq \frac{11}{2}$. Hence,

$$m(w_r, t) \geq \frac{11P}{2} - 1 \geq \frac{9P}{2}.$$

(ii) Otherwise, w_r has at least 2 children $x_1, x_2, \dots, x_d$, at least one of which (without loss of generality, x_1) does not lie in A . By our choice of w as a counterexample to our claim with maximum distance from the root, we have $m(x_1, t) \geq 3P$. Moreover, $m(x_j, t) \geq \min(3P, m_0) = m_0$ for all other children (the lower bound $3P$ applies if $x_j \notin A$, the lower bound m_0 otherwise). It follows that

$$m(w_r, t) \geq e^{-\delta/P} \cdot 3P \cdot m_0^{d-1} - 1 \geq 3m_0 e^{-\delta/P} \cdot P - 1 \geq \frac{11}{2}P - 1.$$

Again, we obtain

$$m(w_r, t) \geq \frac{9P}{2}.$$

Now note that $w_0, w_1, \dots, w_r$ is a branchless path, so that $r \leq P$ by definition. We apply (4.6) repeatedly to $w_{r-1}, w_{r-2}, \dots, w_0 = w$ to obtain

$$m(w_0, t) \geq (e^{-\delta/P})^r m(w_r, t) - \sum_{k=0}^{r-1} e^{-\delta k/P} \geq e^{-\delta} m(w_r, t) - P \geq \left(\frac{9e^{-\delta}}{2} - 1\right)P.$$

The last expression is greater than $3P$ by our choice of δ , and we reach a contradiction. So the claim is proven.

Part 2: Now we complete the proof of (4.4). Taking the logarithm of inequality (4.6), we obtain

$$\log m(v, t) + \log \left(1 + \frac{1}{m(v, t)}\right) \geq \sum_j \log m(v_j, t) - \frac{\delta}{P}. \quad (4.9)$$

Note that the set A can be written as a disjoint union of the vertex sets of certain trees $T(y_1), T(y_2), \dots$ rooted at $y_1, y_2, \dots$. Iterating (4.9) from the root v_0 to the vertices $y_1, y_2, \dots$ and applying (4.7) and Lemma 3.1 yields

$$\begin{aligned} \log m(v_0, t) &\geq \sum_j \log m(y_j, t) - \sum_{v \in T \setminus A} \log \left(1 + \frac{1}{m(v, t)}\right) - \frac{\delta|T \setminus A|}{P} \\ &\geq \sum_j \log \left(\frac{m_0}{2} S^\bullet(T(y_j))\right) - \sum_{v \in T \setminus A} \log \left(1 + \frac{1}{m(v, t)}\right) - \frac{\delta|T \setminus A|}{P} \\ &\geq \sum_j \log \left(\frac{m_0}{2} 2^{L(T(y_j))}\right) - \sum_{v \in T \setminus A} \log \left(1 + \frac{1}{m(v, t)}\right) - \frac{\delta|T \setminus A|}{P}. \end{aligned}$$

Furthermore, since $m_0 < 2$ and the trees $T(y_1), T(y_2), \dots$ contain all leaves of T , we have

$$\begin{aligned} \sum_j \log \left(\frac{m_0}{2} 2^{L(T(y_j))}\right) &\geq \sum_j \log \left(m_0^{L(T(y_j))}\right) \\ &= \log(m_0) \sum_j L(T(y_j)) = \log(m_0)L(T). \end{aligned}$$

Now recall that $m(v, t) \geq 3P$ for all $v \notin A$, which gives us

$$\begin{aligned} \sum_{v \in T \setminus A} \log \left(1 + \frac{1}{m(v, t)} \right) + \frac{\delta |T \setminus A|}{P} &\leq \left(\log \left(1 + \frac{1}{3P} \right) + \frac{\delta}{P} \right) |T \setminus A| \\ &\leq \left(\frac{1}{3} + \delta \right) \frac{|T|}{P}. \end{aligned}$$

Putting these bounds together, we obtain

$$\log m(v_0, t) \geq \log(m_0)L(T) - \left(\frac{1}{3} + \delta \right) \frac{|T|}{P}.$$

From Lemma 3.1, we know that

$$L(T) \geq \frac{|T|}{2P}.$$

Hence, we finally have

$$\log |1 + S^\bullet(T, t)| = \log m(v_0, t) \geq \left(\log(m_0) - \frac{2}{3} - 2\delta \right) L(T).$$

The proof of (4.4) is completed by applying the exponential function on both sides of the latter inequality and by noting that the constant $\log(m_0) - \frac{2}{3} - 2\delta$ is greater than $c_0 = 0.012$ (defined at the beginning of the proof). $\square$

We have shown that $f(T, t)$ and consequently $F(T, t)$ can be regarded as complex analytic functions in a disk around zero, so $F(T, t)$ admits a Taylor expansion near zero, which we are now going to investigate further. By (4.3), we have

$$\mu(T) = \frac{d}{dt} F(T, t) \Big|_{t=0} = \frac{S_u^\bullet(T)}{1 + S^\bullet(T)}$$

and

$$\sigma^2(T) = \frac{d^2}{dt^2} F(T, t) \Big|_{t=0} = \frac{S_{uu}^\bullet(T)}{1 + S^\bullet(T)} + \mu(T) - \mu^2(T),$$

where we use $S_u^\bullet(T)$ as a shorthand for $S_u^\bullet(T, 1) = \frac{d}{du} S^\bullet(T, u) \Big|_{u=1}$ in the same way as $S^\bullet(T)$, and $S_{uu}^\bullet(T)$ is defined analogously for the second derivative. The intuition behind the notation $\mu(T)$ and $\sigma^2(T)$ is that these two quantities are essentially the average order of subtrees in $S^\bullet(T)$ and the variance respectively, if we include an additional dummy subtree of order 0 in the count (compare also the considerations at the beginning of Section 2). This is asymptotically irrelevant and simplifies the following calculations.

For the rest of this section, we let γ be a fixed positive real number, let P be a positive integer that represents an upper bound on the length of all 2-paths in T , and set

$$\Delta = \frac{\delta}{2^{P+1+\gamma}},$$

where $\delta = 0.001$ is as defined in the proof of Lemma 4.1. It also follows from Lemma 4.1 that for every vertex v in T , the function $F(T(v), t)$ is analytic in the disk centred at zero with radius 2Δ . So we can define the quantity

$$r(T) = \sup_{0 < |t| \leq \Delta} \left| \frac{F(T, t) - F(T, 0) - \mu(T)t - \sigma^2(T)\frac{t^2}{2}}{t^3} \right|,$$

which represents the error in the second-order Taylor approximation of $F(T, t)$. Then by definition, for $|t| \leq \Delta$ we have

$$F(T, t) = F(T, 0) + \mu(T)t + \sigma^2(T)\frac{t^2}{2} + \mathcal{O}(r(T)|t|^3). \quad (4.10)$$

Next, we estimate the quantities $\sigma^2(T)$ and $r(T)$. Note first that $\sigma^2(T)$ satisfies the following additive relation that one can easily deduce from its definition and (4.1):

$$\sigma^2(T) = \frac{S^\bullet(T)}{1 + S^\bullet(T)} \sum_j \sigma^2(T(v_j)) + \frac{\mu(T)^2}{S^\bullet(T)}. \quad (4.11)$$

Moreover, the recursion (4.1) also yields

$$\begin{aligned} F(T, t) - F(T, 0) - \mu(T)t - \sigma^2(T)\frac{t^2}{2} &= \\ &= \sum_j \left(F(T(v_j), t) - F(T(v_j), 0) - \mu(T(v_j))t - \sigma^2(T(v_j))\frac{t^2}{2} \right) \\ &\quad + f(T, t) - f(T, 0) - f'(T, t)t - f''(T, t)\frac{t^2}{2}, \end{aligned}$$

so by the triangle inequality

$$r(T) \leq \sum_j r(T(v_j)) + \sup_{0 < |t| \leq \Delta} \left| \frac{f(T, t) - f(T, 0) - f'(T, 0)t - f''(T, 0)\frac{t^2}{2}}{t^3} \right|,$$

and since

$$f(T, t) - f(T, 0) - f'(T, 0)t - f''(T, 0)\frac{t^2}{2} = \int_0^t \int_0^u \int_0^v f'''(T, w) dw dv du,$$

we have

$$r(T) \leq \sum_j r(T(v_j)) + \frac{1}{6} \sup_{|t| \leq \Delta} |f'''(T, t)|. \quad (4.12)$$

As in the proof of Lemma 4.1, we will now iterate (4.11) and (4.12) along the tree to obtain a lower estimate for $\sigma^2(T)$ and an upper estimate for $r(T)$. To this end, we introduce a (now slightly different) notion of “small branches” again: we let B be the set of all vertices w for which $|T(w)| \leq P^{1+\gamma}$. Our first lemma gives an upper estimate for $r(T)$.

Lemma 4.3. *We have*

$$r(T) \ll |T| + \sum_{v \in \mathcal{I}(T) \cap B} \frac{|T(v)|^3}{S^\bullet(T(v))}. \quad (4.13)$$

Proof. Iterating (4.12), we have

$$r(T) \ll L(T) + \sum_{v \in \mathcal{I}(T) \setminus B} \sup_{|t| \leq \Delta} |f'''(T(v), t)| + \sum_{v \in \mathcal{I}(T) \cap B} \sup_{|t| \leq \Delta} |f'''(T(v), t)|.$$

The term $L(T)$ on the right side bounds the contribution from the leaves. We now consider two cases each estimating one of the sums above:

- (i) We first look at the case that $v \notin B$. Cauchy's integral formula yields, for $|t| \leq \Delta$,

$$f'''(T(v), t) = \frac{3!}{2\pi i} \oint_{\mathcal{C}(t, \Delta)} \frac{f(T(v), z) - z}{(z - t)^4} dz,$$

where $\mathcal{C}(t, \Delta)$ is the circle centred at t with radius Δ . The integral representation of $f'''(T(v), t)$ gives us the bound

$$\begin{aligned} |f'''(T(v), t)| &\leq 6\Delta^{-3} \sup_{z \in \mathcal{C}(t, \Delta)} |f(T(v), z) - z| \\ &= 6\Delta^{-3} \sup_{z \in \mathcal{C}(t, \Delta)} \left| \log \left(1 + \frac{1}{S^\bullet(T(v), e^z)} \right) \right|. \end{aligned}$$

Hence,

$$\begin{aligned} \sup_{|t| \leq \Delta} |f'''(T(v), t)| &\leq 6\Delta^{-3} \sup_{|z| \leq 2\Delta} \left| \log \left(1 + \frac{1}{S^\bullet(T(v), e^z)} \right) \right| \\ &\leq 6\Delta^{-3} \sup_{|z| \leq \frac{\delta}{P}} \left| \log \left(1 + \frac{1}{S^\bullet(T(v), e^z)} \right) \right|. \end{aligned}$$

Now we can apply Lemma 4.1 to estimate $|S^\bullet(T(v), e^z)|$ for $|z| \leq \frac{\delta}{P}$ (recall that $|S^\bullet(T(v), e^z)|$ is bounded below by a constant greater than 1 in this case). We obtain

$$\sup_{|t| \leq \Delta} |f'''(T(v), t)| \ll \Delta^{-3} e^{-c_0 L(T(v))}.$$

The assumption $v \notin B$ implies $|T(v)| > P^{1+\gamma}$. In addition, we know that the lengths of all branchless paths in $T(v)$ are bounded above by P since $T(v)$ is a branch of T , so by Lemma 3.1 we have

$$L(T(v)) \geq \frac{|T(v)|}{2P} > \frac{1}{2} P^\gamma.$$

Therefore,

$$\sup_{|t| \leq \Delta} |f'''(T(v), t)| \ll \Delta^{-3} e^{-\frac{c_0}{2} P^\gamma} \ll P^{3(1+\gamma)} e^{-\frac{c_0}{2} P^\gamma},$$

which is bounded above by a constant (that depends on our choice of γ , but is independent of P). Thus

$$\sum_{v \in \mathcal{I}(T) \setminus B} \sup_{|t| \leq \Delta} |f'''(T(v), t)| \ll |T|. \quad (4.14)$$

- (ii) If $v \in \mathcal{I}(T) \cap B$, then the function $f(T(v), z)$ is analytic in the closed disk centred at zero with a slightly larger radius $\frac{\delta}{|T(v)|}$ (this is greater than Δ since $v \in B$, i.e. $|T(v)| \leq P^{1+\gamma}$ by definition). To see this, we can use the same argument that gave us (4.5): for $|z| \leq \frac{\delta}{|T(v)|}$, we have

$$|S^\bullet(T(v), e^z)| \geq e^{-\delta} \cos(\delta) S^\bullet(T(v)), \quad (4.15)$$

which in turn is strictly greater than 1 by the choice we have made for δ and by the fact that $S^\bullet(T(v)) \geq 2$ since $v \in \mathcal{I}(T)$. Now for any t such that $|t| \leq \Delta$, let $\mathcal{C}(t, R)$ be the circle centred at t with radius $R = \frac{\delta}{2|T(v)|}$. Note that $\mathcal{C}(t, R)$ lies in the region of analyticity of the function $f(T(v), z)$, since if $z \in \mathcal{C}(t, R)$, we have

$$|z| \leq |t| + |z - t| \leq \frac{\delta}{2P^{1+\gamma}} + \frac{\delta}{2|T(v)|} \leq \frac{\delta}{|T(v)|}.$$

Thus, by Cauchy's integral formula, we have

$$f'''(T(v), t) = \frac{3!}{2\pi i} \oint_{\mathcal{C}(t, R)} \frac{f(T(v), z) - z}{(z - t)^4} dz,$$

from which we deduce the bound

$$|f'''(T(v), t)| \leq 48 \delta^{-3} |T(v)|^3 \sup_{z \in \mathcal{C}(t, R)} |f(T(v), z) - z|.$$

The right side can be estimated using (4.15):

$$\sup_{z \in \mathcal{C}(t, R)} |f(T(v), z) - z| = \sup_{z \in \mathcal{C}(t, R)} \left| \log \left(1 + \frac{1}{S^\bullet(T(v), e^z)} \right) \right| \ll \frac{1}{S^\bullet(T(v))}$$

uniformly for $|t| \leq \Delta$. Therefore, we obtain

$$\sup_{|t| \leq \Delta} |f'''(T(v), t)| \ll \frac{|T(v)|^3}{S^\bullet(T(v))} \quad (4.16)$$

for $v \in \mathcal{I}(T) \cap B$.

The lemma follows by combining (4.14) and (4.16). $\square$

Let $\mathcal{P}(v)$ denote the set of all vertices on the path in T from v to the root v_0 (excluding v , but including v_0). We define

$$\eta(v) = \begin{cases} 1 & \text{if } v = v_0, \\ \prod_{w \in \mathcal{P}(v)} \frac{S^\bullet(T(w))}{1 + S^\bullet(T(w))} & \text{otherwise.} \end{cases}$$

Lemma 4.4. *Suppose that $L(T) \geq \lambda|T|$ for some fixed constant $\lambda > 0$. We have*

$$\sigma^2(T) \gg |T| + \sum_{v \in \mathcal{I}(T) \cap B} \eta(v) \frac{|T(v)|^2}{S^\bullet(T(v))}. \quad (4.17)$$

The implied constant only depends on λ .

Proof. Iterating (4.11) (and noting that $\sigma^2(T) = \frac{1}{4} > 0$ if $|T| = 1$), we obtain

$$\begin{aligned} \sigma^2(T) &\gg \sum_{v \in \mathcal{L}(T)} \eta(v) + \sum_{v \in \mathcal{I}(T)} \eta(v) \frac{\mu(T(v))^2}{S^\bullet(T(v))} \\ &\geq \sum_{v \in \mathcal{L}(T)} \eta(v) + \sum_{v \in \mathcal{I}(T) \cap B} \eta(v) \frac{\mu(T(v))^2}{S^\bullet(T(v))}. \end{aligned}$$

It was shown by Jamison in [5] that the average cardinality of a subtree containing the root of a rooted tree of order n is at least $(n+1)/2$, so

$$\frac{S_u^\bullet(T(v))}{S^\bullet(T(v))} \geq \frac{|T(v)|+1}{2},$$

which implies that

$$\mu(T(v)) = \frac{S_u^\bullet(T(v))}{S^\bullet(T(v))} \cdot \frac{1}{1 + S^\bullet(T(v))^{-1}} \geq \frac{|T(v)|+1}{2} \cdot \frac{1}{1 + |T(v)|^{-1}} = \frac{|T(v)|}{2}.$$

So it remains to show that

$$\sum_{v \in \mathcal{L}(T)} \eta(v) \gg |T|. \quad (4.18)$$

To this end, we define a set of “exceptional branches” in such a way that $\eta(v)$ is bounded below by an explicit constant unless v lies in one of these branches. Choose two constants $\beta \in (0, 1)$ and $K > (\lambda/2)^{-1/\beta}$, and let $z_1, z_2, \dots, z_M$ be the vertices that satisfy

$$L(T(z_j)) \leq |T(z_j)|^{1-\beta} \text{ and } |T(z_j)| \geq K$$

and are closest to the root with this property (in the sense that no vertex on the path from the root to z_j satisfies both inequalities). We set $E_j = T(z_j)$ and let E be the union of all E_j . Now take any leaf v that does not lie in E , and let v' be its ancestor closest to the root that satisfies $|T(v')| < K$ (possibly, $v' = v$). Now we split the product that defines $\eta(v)$ as follows:

$$\begin{aligned} \eta(v) &= \prod_{w \in \mathcal{P}(v)} \frac{1}{1 + S^\bullet(T(w))^{-1}} \\ &= \prod_{w \in \mathcal{P}(v) \setminus \mathcal{P}(v')} \frac{1}{1 + S^\bullet(T(w))^{-1}} \prod_{w \in \mathcal{P}(v')} \frac{1}{1 + S^\bullet(T(w))^{-1}}. \end{aligned}$$

There are at most K vertices in $\mathcal{P}(v) \setminus \mathcal{P}(v')$ since the set $\mathcal{P}(v) \setminus \mathcal{P}(v')$ lies entirely in $T(v')$. In addition, for every w we have the trivial bound $1 + S^\bullet(T(w))^{-1} \leq 2$. Therefore,

$$\prod_{w \in \mathcal{P}(v) \setminus \mathcal{P}(v')} \frac{1}{1 + S^\bullet(T(w))^{-1}} \geq 2^{-K}.$$

Furthermore, for every vertex w on the path from the root to v' , we must have $|T(w)| \geq K$ by the choice of v' , and $L(T(w)) > |T(w)|^{1-\beta}$ since v does not lie in E . Recall from Lemma 3.1 that $S^\bullet(T(w)) \geq 2^{L(T(w))}$. Hence we have

$$\begin{aligned} \eta(v) &\geq 2^{-K} \prod_{w \in \mathcal{P}(v')} \left(1 + 2^{-L(T(w))}\right)^{-1} \\ &\geq 2^{-K} \prod_{w \in \mathcal{P}(v')} \left(1 + 2^{-|T(w)|^{1-\beta}}\right)^{-1} \\ &\geq 2^{-K} \prod_{j \geq K} \left(1 + 2^{-j^{1-\beta}}\right)^{-1}. \end{aligned}$$

Note that the infinite product converges. So we can deduce that $\eta(v)$ is bounded below by a constant that only depends on β and K unless $v \in E$. Consequently,

$$\sum_{v \in \mathcal{L}(T)} \eta(v) \gg |\mathcal{L}(T) \setminus E|. \quad (4.19)$$

We will see that E cannot contain more than half of the leaves. We may assume that E is non-empty, for otherwise this statement is trivial. So let us assume that

$$\sum_{j=1}^M L(E_j) > \frac{L(T)}{2} \geq \frac{\lambda}{2} |T|.$$

By the definition of the branches $E_1, E_2, \dots, E_M$, this gives us

$$\sum_{j=1}^M |E_j|^{1-\beta} \geq \frac{\lambda}{2} |T|.$$

On the other hand, since $E_1, E_2, \dots, E_M$ are pairwise disjoint, we also have

$$\sum_{j=1}^M |E_j| \leq |T|.$$

Since we are assuming that E is non-empty, we have $M \neq 0$. Hence, by Jensen's inequality,

$$\frac{\lambda}{2} |T| \leq \sum_{j=1}^M |E_j|^{1-\beta} \leq M \left(\frac{\sum_{j=1}^M |E_j|}{M} \right)^{1-\beta} \leq M \left(\frac{|T|}{M} \right)^{1-\beta}.$$

It follows that

$$M \geq (\lambda/2)^{1/\beta} |T|.$$

On the other hand, each E_j contains at least K vertices, so we have

$$|T| \geq \sum_{j=1}^M |E_j| \geq MK.$$

Combining the last two inequalities, we obtain

$$K \leq (\lambda/2)^{-1/\beta}, \quad (4.20)$$

which contradicts the choice of K . This means that $|E| \leq L(T)/2$, so (4.19) finally yields

$$\sum_{v \in \mathcal{L}(T)} \eta(v) \gg L(T) \gg |T|,$$

which completes the proof. Note that the implied constant does indeed only depend on λ (and our choice of β and K , which was arbitrary). $\square$

To make use of the previous lemma, we also need to bound $\eta(v)$ from below for $v \in \mathcal{I}(T) \cap B$, which is achieved by the following lemma:

Lemma 4.5. *For every vertex $v \in T$ and every vertex $v' \in \mathcal{P}(v)$, we have*

$$\eta(v) \geq \eta(v') \frac{|T(v)|}{2|T(v')|}.$$

Proof. The statement is void if v is the root v_0 , so we assume from now on that v is not the root. Let $v' = w_0, w_1, w_2, \dots, w_k = v$ be the vertices of the path connecting v' and v (which form part of the path connecting v_0 and v). By definition, we have

$$\frac{\eta(v)}{\eta(v')} = \prod_{j=0}^{k-1} \frac{S^\bullet(T(w_j))}{1 + S^\bullet(T(w_j))}.$$

Clearly, $S^\bullet(T(w_j)) \geq 1 + S^\bullet(T(w_{j+1}))$ for $j = 0, 1, \dots, k-1$; iterating further, we obtain

$$S^\bullet(T(w_j)) \geq k - j + S^\bullet(T(v)).$$

So we have, for $j = 0, 1, \dots, k-1$,

$$\frac{S^\bullet(T(w_j))}{1 + S^\bullet(T(w_j))} \geq \frac{k - j + S^\bullet(T(v))}{1 + k - j + S^\bullet(T(v))},$$

and it follows that

$$\frac{\eta(v)}{\eta(v')} \geq \prod_{j=0}^{k-1} \frac{k - j + S^\bullet(T(v))}{1 + k - j + S^\bullet(T(v))} = \frac{1 + S^\bullet(T(v))}{1 + k + S^\bullet(T(v))} \geq \frac{S^\bullet(T(v))}{k + S^\bullet(T(v))}.$$

Now we consider two cases:

(i) First, if $S^\bullet(T(v)) \geq k$ then

$$\frac{\eta(v)}{\eta(v')} \geq \frac{S^\bullet(T(v))}{k + S^\bullet(T(v))} \geq \frac{1}{2} \geq \frac{|T(v)|}{2|T(v')|}.$$

(ii) Otherwise, if $S^\bullet(T(v)) < k$, then

$$\frac{\eta(v)}{\eta(v')} \geq \frac{S^\bullet(T(v))}{2k} \geq \frac{|T(v)|}{2|T(v')|}.$$

The last inequality holds because $S^\bullet(T(v)) \geq |T(v)|$ and $|T(v')| > k$ (the latter since $T(v')$ contains the $k+1$ vertices $w_0, w_1, \dots, w_k$).

Hence, the lemma follows. $\square$

The bound on $\eta(v)$ is now used to bound $r(T)$ in terms of $\sigma^2(T)$.

Lemma 4.6. *Suppose that $L(T) \geq \lambda|T|$ for a fixed constant $\lambda > 0$. We have*

$$r(T) \ll P^{1+\gamma} \sigma^2(T).$$

The constant implied in this estimate depends on λ and γ , but nothing else, in particular not on P .

Proof. Recall that B consists of all vertices w for which $T(w) \leq P^{1+\gamma}$. We write B as the disjoint union of branches $T(y_1), T(y_2), \dots$. If v lies on the path connecting the root v_0 and one of the y_j , then by definition we have

$$|T(v)| > P^{1+\gamma}.$$

By Lemma 3.1, this implies

$$L(T(v)) \geq \frac{|T(v)|}{2P} \geq \frac{|T(v)|}{2|T(v)|^{1/(1+\gamma)}} = \frac{1}{2}|T(v)|^{\gamma/(1+\gamma)}.$$

Using this inequality, we can argue as in the proof of (4.19) that $\eta(y_j)$ is bounded below by an absolute constant for every j . Applying Lemma 4.5, we deduce that for $v \in T(y_j)$,

$$\eta(v) \gg \frac{|T(v)|}{|T(y_j)|} \geq \frac{|T(v)|}{P^{1+\gamma}}.$$

Therefore,

$$\begin{aligned} \sum_{v \in \mathcal{I}(T) \cap B} \eta(v) \frac{|T(v)|^2}{S^\bullet(T(v))} &= \sum_j \sum_{v \in \mathcal{I}(T(y_j))} \eta(v) \frac{|T(v)|^2}{S^\bullet(T(v))} \\ &\gg P^{-1-\gamma} \sum_{v \in \mathcal{I}(T) \cap B} \frac{|T(v)|^3}{S^\bullet(T(v))}. \end{aligned}$$

The desired statement now follows from Lemma 4.3 and Lemma 4.4. $\square$

As a consequence of Lemma 4.6, we now obtain the required information on the Taylor expansion of $F(T, t)$.

Proposition 4.7. *Let $\delta = 0.001$ be as previously defined, and let $\lambda, \gamma > 0$ be fixed constants. If $L(T) \geq \lambda|T|$, then we have*

$$F(T, t) = F(T, 0) + \mu(T)t + \sigma^2(T)\frac{t^2}{2} + \mathcal{O}(P(T)^{1+\gamma}\sigma^2(T)|t|^3) \quad (4.21)$$

for $|t| \leq \frac{\delta}{2P(T)^{1+\gamma}}$, where the constant implied in the $\mathcal{O}$ -term only depends on λ and γ .

Proof. This statement follows directly from Lemma 4.6 and (4.10). $\square$

4.2 Central limit theorem

We are now ready to prove the central limit theorem for the order distribution of subtrees.

Theorem 4.8. *Let $T_1, T_2, \dots$ be a sequence of rooted trees such that $|T_n| \rightarrow \infty$ as $n \rightarrow \infty$ and the following two conditions are satisfied for all sufficiently large n :*

- (i) $P(T_n) \leq |T_n|^{\frac{1}{2}-\epsilon}$ for some constant $\epsilon > 0$,
- (ii) $L(T_n) \geq \lambda|T_n|$ for some constant $\lambda > 0$.

Then the distribution of the random variable $X_n^\bullet$, defined as the order of a randomly chosen subtree of T_n containing the root, is asymptotically Gaussian. More precisely, if $\Phi_n^\bullet(x)$ denotes the distribution function of the renormalised random variable

$$Y_n^\bullet = \frac{X_n^\bullet - \mu(T_n)}{\sigma(T_n)},$$

then we have the following estimate for the speed of convergence:

$$\sup_{x \in \mathbb{R}} \left| \Phi_n^\bullet(x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \right| = \mathcal{O}(|T_n|^{-\alpha}) \quad (4.22)$$

for every positive constant $\alpha < \epsilon/3$. The constant implied in the $\mathcal{O}$ -term only depends on α and λ .

Proof. For ease of notation, we drop the dependence on n . Recall that the moment generating function of $X^\bullet = X_n^\bullet$ is

$$\mathbb{E}(e^{tX^\bullet}) = \frac{S^\bullet(T, e^t)}{S^\bullet(T)}.$$

Instead of working directly with $X^\bullet$, we use the modified random variable $X^* = X_n^*$ that also includes an empty dummy subtree. The moment generating function of this random variable is given by

$$\mathbb{E}(e^{tX^*}) = \frac{1 + S^\bullet(T, e^t)}{1 + S^\bullet(T)},$$

and if $Y^* = Y_n^* = (X_n^* - \mu(T_n))/\sigma(T_n)$ is the associated renormalised random variable, it is easy to see that the distribution functions $\Phi^\bullet$ of $Y^\bullet$ and Φ^* of Y^* differ only by very little:

$$|\Phi^\bullet(x) - \Phi^*(x)| \leq \frac{1}{1 + S^\bullet(T)} \quad (4.23)$$

for all $x \in \mathbb{R}$, so it is sufficient to prove the estimate for Φ^* instead of $\Phi^\bullet$. The condition $L(T) \geq \lambda|T|$ implies

$$\sigma^2(T) \gg |T|$$

by Lemma 4.4, in particular $\sigma(T) \rightarrow \infty$ as $|T| \rightarrow \infty$. The moment generating function of the renormalised random variable Y^* is

$$\begin{aligned} \mathbb{E}(e^{tY^*}) &= e^{-\mu(T)t/\sigma(T)} \mathbb{E}(e^{tX^*/\sigma(T)}) \\ &= \exp\left(-\frac{\mu(T)t}{\sigma(T)} + F\left(T, \frac{t}{\sigma(T)}\right) - F(T, 0)\right). \end{aligned}$$

The expansion in Proposition 4.7 gives us

$$F\left(T, \frac{t}{\sigma(T)}\right) = F(T, 0) + \frac{\mu(T)}{\sigma(T)}t + \frac{t^2}{2} + \mathcal{O}\left(\frac{P(T)^{1+\gamma}}{\sqrt{|T|}}|t|^3\right)$$

and thus

$$\mathbb{E}(e^{tY^*}) = \exp\left(\frac{t^2}{2} + \mathcal{O}\left(\frac{P(T)^{1+\gamma}}{\sqrt{|T|}}|t|^3\right)\right) \quad (4.24)$$

if $|t| \leq \frac{\delta\sigma(T)}{2P(T)^{1+\gamma}}$. Note that we can choose γ freely here (the choice affects the $\mathcal{O}$ -constant, though). The condition $P(T) \leq |T|^{\frac{1}{2}-\epsilon}$ allows us to choose γ in such a way that

$$\frac{P(T)^{1+\gamma}}{\sqrt{|T|}} \rightarrow 0.$$

Therefore,

$$\mathbb{E}\left(e^{tY^*}\right) \longrightarrow e^{t^2/2}$$

for any fixed t as $n \rightarrow \infty$, which would already prove a central limit theorem. For the precise error estimate, we use the following Berry-Esseen type inequality [10, Theorem 5.1]:

$$\sup_{x \in \mathbb{R}} \left| \Phi^*(x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \right| \leq c_1 \int_{-M}^M \left| \frac{\varphi_T(t) - e^{-t^2/2}}{t} \right| dt + \frac{c_2}{M}$$

for certain absolute constants c_1, c_2 , where

$$\varphi_T(t) = \int_{-\infty}^{\infty} e^{ity} d\Phi^*(y) = \mathbb{E}\left(e^{itY^*}\right).$$

In view of (4.24), we have

$$\left| \varphi_T(t) - e^{-t^2/2} \right| \ll |t|^3 e^{-t^2/2} \frac{P(T)^{1+\gamma}}{\sqrt{|T|}}$$

if $|t|^3 = \mathcal{O}\left(\sqrt{|T|}/P(T)^{1+\gamma}\right)$. Therefore,

$$\sup_{x \in \mathbb{R}} \left| \Phi^*(x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \right| = \mathcal{O}\left(\frac{P(T)^{1+\gamma}}{\sqrt{|T|}} + \frac{1}{M}\right)$$

for any M satisfying $M^3 = \mathcal{O}\left(\sqrt{|T|}/P(T)^{1+\gamma}\right)$. We choose

$$M = \left(\frac{\sqrt{|T|}}{P(T)^{1+\gamma}} \right)^{1/3}$$

and γ in such a way that

$$\frac{1}{M} = \left(\frac{P(T)^{1+\gamma}}{\sqrt{|T|}} \right)^{1/3} \leq \left(\frac{|T|^{(1+\gamma)(1/2-\epsilon)}}{\sqrt{|T|}} \right)^{1/3} = |T|^{\gamma(1-2\epsilon)/6-\epsilon/3} \leq |T|^{-\alpha}.$$

Note finally that the difference between $\Phi^\bullet(x)$ and $\Phi^*(x)$ is uniformly bounded above by $S^\bullet(T)^{-1}$ in view of (4.23). Since $S^\bullet(T) \geq |T| \geq |T|^\alpha$, this completes the proof. $\square$

4.3 Local limit theorem

Now that we have established a central limit theorem, it is natural to ask whether a local limit theorem for single coefficients of $S^\bullet(T, u)$ also holds. To be precise, given a sequence of rooted trees $T_1, T_2, \dots$ satisfying both properties of Theorem 4.8, can we give an estimate for the number of subtrees of order k , for values of k around the mean $\mu(T_n)$? In this section, we show that it is indeed possible to obtain such a result. Before we come to the proof, an estimate for $|S^\bullet(T, u)|$ when u lies on the unit circle is required. This is precisely what we state in the next lemma.

Lemma 4.9. *Let $\lambda, \gamma > 0$ be fixed constants, and suppose that $L(T) \geq \lambda|T|$. There exist constants δ_1, c_3, c_4 depending on λ, γ such that, with*

$$\Delta_1 = \frac{\delta_1}{2P(T)^{1+\gamma}},$$

we have

$$\frac{|1 + S^\bullet(T, e^{it})|}{1 + S^\bullet(T)} \leq \begin{cases} e^{-c_3 t^2 \sigma(T)^2} & \text{if } t \in [-\Delta_1, \Delta_1], \\ e^{-c_4 t^2 |T|} & \text{for all } t \in [-\pi, \pi]. \end{cases}$$

Proof. The bound corresponding to $|t| \leq \Delta_1$ follows easily from Proposition 4.7 for sufficiently small $\delta_1 \leq \delta (= 0.001)$. Thus it suffices to prove the second bound.

Recall that we have

$$S^\bullet(T, e^{it}) = \prod_{j=1}^d (1 + S^\bullet(T(v_j), e^{it}))$$

if $v_1, v_2, \dots, v_d$ are the root's children, and consequently

$$|1 + S^\bullet(T, e^{it})| \leq 1 + \prod_{j=1}^d |1 + S^\bullet(T(v_j), e^{it})|. \quad (4.25)$$

This motivates the definition of a polynomial $R(T, x)$ (for positive real x) that is similar to $S^\bullet(T, u)$: it is given by $R(T, x) = x$ for $|T| = 1$ and the recursion

$$R(T(v), x) = 1 + \prod_{j=1}^d R(T(v_j), x). \quad (4.26)$$

In view of (4.25), we have

$$|1 + S^\bullet(T, e^{it})| \leq R(T, |1 + e^{it}|) \quad (4.27)$$

and $1 + S^\bullet(T, 1) = 1 + S^\bullet(T) = R(T, 2)$. Note that $R(T, x)$ is a polynomial of degree $L(T)$ with positive coefficients. Therefore, it is a strictly increasing function of x , and it admits the trivial lower bound

$$R(T, x) \geq x^{L(T)} \quad (4.28)$$

for all positive x . We also define the function $G(T, x) = \log(R(T, x))$, which satisfies the recurrence

$$G(T, x) = \sum_{j=1}^d G(T(v_j), x) - \log \left(1 - \frac{1}{R(T, x)} \right), \quad (4.29)$$

where $G(T, x) = \log x$ (and thus $G'(T, x) = x^{-1}$) if T only has one vertex. In order to estimate $S^\bullet(T, e^{it})$ by means of (4.27), we establish a bound for the difference $G(T, 2) - G(T, x)$ for x in the interval $[\sqrt{2}, 2]$. By the mean value theorem, there exists some $y \in [x, 2]$ such that

$$G(T, 2) - G(T, x) = (2 - x)G'(T, y).$$

It is not hard to see from (4.26) that the derivative $G'(T, y)$ satisfies

$$G'(T, y) = \frac{R(T, y) - 1}{R(T, y)} \sum_{j=1}^d G'(T(v_j), y). \quad (4.30)$$

We essentially use the same argument as in the proof of Lemma 4.4 to bound $G'(T, y)$ from below. Iterating (4.30) starting from the root of T down to the leaves, we obtain, with

$$\xi(v, y) = \begin{cases} 1 & \text{if } v \text{ is the root of } T, \\ \prod_{w \in \mathcal{P}(v)} \frac{R(T(w), y) - 1}{R(T(w), y)} & \text{otherwise,} \end{cases}$$

that

$$G'(T, y) = y^{-1} \sum_{v \in \mathcal{L}(T)} \xi(v, y).$$

Recall that we are assuming $x \in [\sqrt{2}, 2]$ and thus also $y \in [\sqrt{2}, 2]$. Since $R(T(v), y) \geq y^{L(T(v))} \geq 2^{L(T(v))/2}$, the same argument that gave us (4.18) now yields

$$G'(T, y) \geq \frac{1}{2} \sum_{v \in \mathcal{L}(T)} \xi(v, y) \gg \sum_{v \in \mathcal{L}(T)} \xi(v, y) \gg |T|.$$

This implies that there exists a positive constant c_5 such that

$$G(T, x) - G(T, 2) \leq c_5(x - 2)|T|.$$

Equivalently, if $\sqrt{2} \leq x \leq 2$, then

$$\frac{R(T, x)}{R(T, 2)} \leq e^{c_5(x-2)|T|}. \quad (4.31)$$

To complete the proof, recall that (by (4.27)) $|1 + S^\bullet(T, e^{it})|$ is bounded above by $R(T, |1 + e^{it}|)$ while $R(T, 2) = 1 + S^\bullet(T)$. For $|t| \leq \pi/2$, we have $|1 + e^{it}| \geq \sqrt{2}$ and

$$|1 + e^{it}| - 2 = 2(\cos \frac{t}{2} - 1) \leq -\frac{2}{\pi^2} t^2,$$

thus

$$\frac{|1 + S^\bullet(T, e^{it})|}{1 + S^\bullet(T)} \leq \frac{R(T, |1 + e^{it}|)}{R(T, 2)} \leq e^{-(2c_5/\pi^2)t^2|T|} \leq e^{-c_4 t^2 |T|}$$

if we choose $c_4 \leq 2c_5/\pi^2$. For the case that $|t| \geq \pi/2$, we simply note that $R(T, x)$ is an increasing function of x , so that

$$\frac{|1 + S^\bullet(T, e^{it})|}{1 + S^\bullet(T)} \leq \frac{R(T, |1 + e^{it}|)}{R(T, 2)} \leq \frac{R(T, \sqrt{2})}{R(T, 2)} \leq e^{-c_5(2-\sqrt{2})|T|} \leq e^{-c_4 t^2 |T|}$$

if we choose $c_4 \leq (2 - \sqrt{2})c_5/\pi^2$. This completes the proof. $\square$

Now we have all required ingredients for a local limit theorem. In the following, we let $s_k^\bullet(T)$ denote the number of subtrees of order k in T that contain the root, so that

$$S^\bullet(T, u) = \sum_{k=1}^{|T|} s_k^\bullet(T) u^k.$$

Theorem 4.10. *Suppose that the sequence $T_1, T_2, \dots$ of rooted trees satisfies the conditions of Theorem 4.8. If $k = \mu(T_n) + x\sigma(T_n)$, then we have*

$$\frac{s_k^\bullet(T_n)}{S^\bullet(T_n)} \sim \frac{e^{-x^2/2}}{\sqrt{2\pi\sigma(T_n)}},$$

uniformly for x in any fixed compact interval as $n \rightarrow \infty$.

Proof. Once again, we drop the index n for convenience. By Cauchy's integral formula, the number $s_k^\bullet(T)$ can be expressed as

$$s_k^\bullet(T) = \frac{1}{2\pi i} \oint_{C(0,1)} \left(1 + S^\bullet(T, z)\right) \frac{dz}{z^{k+1}},$$

where $C(0, 1)$ is the unit circle. If we set $z = e^{it}$, then we obtain

$$s_k^\bullet(T) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(1 + S^\bullet(T, e^{it})\right) e^{-ikt} dt.$$

Choose $\gamma > 0$ and $\kappa > 0$ in such a way that $\gamma/2 + 3\kappa < \epsilon$, and set $M = |T|^\kappa/\sigma(T)$. We split the integral into two parts: the central part

$$\frac{1}{2\pi} \int_{-M}^M \left(1 + S^\bullet(T, e^{it})\right) e^{-ikt} dt,$$

and the rest. Recall that we are assuming $P(T) \leq |T|^{1/2-\epsilon}$ and that we have already established $\sigma(T)^2 \gg |T|$. Since

$$\frac{\Delta_1}{M} = \frac{\delta_1 \sigma(T)}{2P(T)^{1+\gamma}|T|^\kappa} \gg |T|^{1/2-\kappa-(1/2-\epsilon)(1+\gamma)} \gg |T|^{\epsilon-\gamma/2-3\kappa}$$

is greater than 1 for sufficiently large $|T|$, we have $M \leq \Delta_1 = \frac{\delta_1}{2P(T)^{1+\gamma}}$, so we can apply Proposition 4.7, which gives us, for $|t| \leq M$,

$$\begin{aligned} 1 + S^\bullet(T, e^{it}) &= \exp(F(T, it)) \\ &= \exp\left(F(T, 0) + i\mu(T)t - \sigma^2(T)\frac{t^2}{2} + \mathcal{O}(|T|^{3\kappa+(1/2-\epsilon)(1+\gamma)-1/2})\right) \\ &= \exp\left(F(T, 0) + i\mu(T)t - \sigma^2(T)\frac{t^2}{2}\right) \left(1 + \mathcal{O}(|T|^{-(\epsilon-\gamma/2-3\kappa)})\right). \end{aligned}$$

We plug in $k = \mu(T) + x\sigma(T)$ and obtain

$$\begin{aligned} \frac{1}{2\pi} \int_{-M}^M \left(1 + S^\bullet(T, e^{it})\right) e^{-ikt} dt &= \frac{1}{2\pi} \int_{-M}^M e^{F(T, 0) - ix\sigma(T)t - \sigma^2(T)t^2/2} dt \\ &\quad + \mathcal{O}\left(|T|^{-(\epsilon-\gamma/2-3\kappa)} \int_{-M}^M e^{F(T, 0) - \sigma^2(T)t^2/2} dt\right). \end{aligned}$$

Since we have

$$\begin{aligned} \int_{-M}^M e^{F(T,0) - ix\sigma(T)t - \sigma^2(T)t^2/2} dt &= \frac{\sqrt{2\pi}}{\sigma(T)} e^{F(T,0) - x^2/2} + \mathcal{O}\left(\int_M^\infty e^{F(T,0) - \sigma^2(T)t^2/2} dt\right) \\ &= \frac{\sqrt{2\pi}}{\sigma(T)} e^{F(T,0) - x^2/2} + \mathcal{O}\left(e^{F(T,0) - \sigma^2(T)M^2/2}\right) \\ &= \frac{\sqrt{2\pi}}{\sigma(T)} e^{F(T,0) - x^2/2} + \mathcal{O}\left(e^{F(T,0) - |T|^{2\kappa}/2}\right), \end{aligned}$$

and $e^{F(T,0)} = 1 + S^\bullet(T)$, we end up with

$$\frac{1}{2\pi} \int_{-M}^M \left(1 + S^\bullet(T, e^{it})\right) e^{-ikt} dt = S^\bullet(T) \frac{e^{-x^2/2}}{\sqrt{2\pi}\sigma(T)} \left(1 + \mathcal{O}(|T|^{-(\epsilon-\gamma/2-3\kappa)})\right).$$

For the remaining integrals, where $|t| \geq M$, we use the estimates from Lemma 4.9. For $|t| \leq \Delta_1 = \frac{\delta_1}{2P(T)^{1+\gamma}}$, they give us

$$\frac{|1 + S^\bullet(T, e^{it})|}{1 + S^\bullet(T)} \leq e^{-c_3 M^2 \sigma(T)^2} = e^{-c_3 |T|^{2\kappa}},$$

and for $|t| \geq \Delta_1$, we get

$$\frac{|1 + S^\bullet(T, e^{it})|}{1 + S^\bullet(T)} \leq e^{-c_4 \Delta_1^2 |T|} \leq e^{-\delta_1^2 c_4 |T|^{1-(1-2\epsilon)(1+\gamma)}/4} \leq e^{-\delta_1^2 c_4 |T|^{2\epsilon-\gamma}/4}.$$

Since these decay faster than any power of T , the parts of the integral for which $|t| \geq M$ will only contribute to the error term. In summary, we have

$$\frac{s_k^\bullet(T)}{S^\bullet(T)} = \frac{e^{-x^2/2}}{\sqrt{2\pi}\sigma(T_n)} \left(1 + \mathcal{O}(|T|^{-(\epsilon-\gamma/2-3\kappa)})\right),$$

which completes the proof. $\square$

Remark 4.11. Theorem 4.10 provides a positive answer to Question 1.1 in an asymptotic sense for large rooted trees (and as we will see in the following section, also unrooted trees) without vertices of degree 2, since both technical conditions are trivially satisfied in this case.

5 Unrooted trees

Now that we have established both a central and a local limit theorem for the number of subtrees containing the root of a rooted tree, we would like to carry the results over to unrooted trees as well. This is achieved by means of the following lemma, which guarantees the existence of a vertex that is contained in most subtrees:

Lemma 5.1. *For every tree T , there exists a vertex v of T such that the proportion of subtrees of T that do not contain v is at most $|T|2^{-L(T)/2}$.*

Proof. Let v be a vertex that minimises the sum of the distances to all leaves, i.e. the expression $\sum_{w \in \mathcal{L}(T)} d(v, w)$ attains its minimum (this is called a “leaf centroid” in [17],

in analogy to the centroid). Let $T_1, T_2, \dots, T_k$ be the branches of T , rooted at v , and $v_1, v_2, \dots, v_k$ the corresponding neighbours of v . The important observation about v is that none of the branches can contain more than half of the leaves: if T_j contains more than $L(T)/2$ leaves, then we have

$$\sum_{w \in \mathcal{L}(T)} d(v, w) > \sum_{w \in \mathcal{L}(T)} d(v_j, w),$$

since $d(v_j, w) = d(v, w) - 1$ if w is in T_j , and $d(v_j, w) = d(v, w) + 1$ otherwise. This would contradict the choice of v .

Let τ be a subtree of T that does not contain v . It must then be completely contained in some branch T_j . It has a unique vertex closest to v , which we denote by w . We can associate $2^{|\mathcal{L}(T) \cap (T \setminus T_j)|} \geq 2^{L(T)/2}$ subtrees to τ that contain v , obtained by adding the path from w to v as well as all non-leaves not contained in T_j and any subset of the $|\mathcal{L}(T) \cap (T \setminus T_j)|$ leaves that do not lie in T_j . Finally, we root the resulting subtrees at w .

Let the total number of subtrees of T be denoted by $S(T)$ and the number of those subtrees not containing v by $S^\circ(T)$. The construction above yields at least $2^{L(T)/2}$ rooted subtrees of T associated with every subtree τ that does not contain v . The original tree τ can be recovered uniquely from such a tree σ : it consists of the root w of σ and all vertices for which the unique path from v passes through w . Thus our construction is an injection to the set of rooted subtrees of T (whose cardinality is clearly at most $|T|S(T)$), and we obtain the inequality

$$S^\circ(T) \cdot 2^{L(T)/2} \leq |T| \cdot S(T),$$

from which the statement of the lemma follows. $\square$

Our main theorem now follows immediately both in the central and local version:

Theorem 5.2. *Let $T_1, T_2, \dots$ be a sequence of trees such that $|T_n| \rightarrow \infty$ as $n \rightarrow \infty$ and the following two conditions are satisfied:*

- (i) $P(T_n) \leq |T_n|^{\frac{1}{2}-\epsilon}$ for some constant $\epsilon > 0$,
- (ii) $L(T_n) \geq \lambda|T_n|$ for some constant $\lambda > 0$.

Then the distribution of the random variable X_n , defined as the order of a randomly chosen subtree of T_n , is asymptotically Gaussian. More precisely, if $\Phi_n(x)$ denotes the distribution function of the renormalised random variable

$$Y_n = \frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}},$$

then we have the following estimate for the speed of convergence:

$$\sup_{x \in \mathbb{R}} \left| \Phi_n(x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \right| = \mathcal{O}(|T_n|^{-\alpha}), \quad (5.1)$$

for any positive constant $\alpha < \epsilon/3$. The constant implied in the $\mathcal{O}$ -term only depends on α and λ . Moreover, if $k = \mathbb{E}(X_n) + x\sqrt{\mathbb{V}(X_n)} \in \mathbb{N}$, then we have the local limit theorem

$$\mathbb{P}(X_n = k) \sim \frac{e^{-x^2/2}}{\sqrt{2\pi\mathbb{V}(X_n)}},$$

uniformly for x in any fixed compact interval.

Proof. As in the proofs of Theorem 4.8 and Theorem 4.10, we suppress the dependence on n for ease of notation. Choose v as in Lemma 5.1, and let $X^{(v)}$ be the random variable defined as the order of a randomly selected subtree of T containing v . By Lemma 5.1, the total variation distance between the two random variables $X = X_n$ and $X^{(v)}$, which is defined as

$$\sup_A |\mathbb{P}(X^{(v)} \in A) - \mathbb{P}(X \in A)|,$$

is $\mathcal{O}(|T|/2^{L(T)/2})$. In view of our assumption on the number of leaves, this goes to 0 even at an exponential rate. Letting $\mu(T)$ and $\sigma^2(T)$ be defined as before for the tree T rooted at v , it is also easy to see by the same argument that $\mathbb{E}(X) = \mu(T) + \mathcal{O}(1)$ and $\mathbb{V}(X) = \sigma^2(T) + \mathcal{O}(1)$ (in fact, both error terms can be made exponentially small). The two statements now follow directly from Theorem 4.8 and Theorem 4.10. $\square$

6 Random trees

The technical conditions of Theorems 4.8, 4.10 and 5.2 are not satisfied for all possible sequences of trees, but they do hold for “generic” (randomly chosen) trees. In fact, it was shown in [11] that the length of the longest branchless path of a random labelled tree of order n is concentrated around $\log n$ for large n (with a limit distribution of double exponential type), and the number of leaves of a random labelled tree of order n is concentrated around n/e (with a Gaussian limit distribution, see e.g. [2, Section 3.2.1]). Analogous statements (with different constants) hold for other families of random trees (e.g. random plane trees, random binary trees).

If T_n denotes a random labelled tree of order n for $n = 1, 2, \dots$, then a simple application of the Borel-Cantelli Lemma shows that the conditions of Theorem 5.2 with arbitrary $\epsilon < \frac{1}{2}$ and $\lambda < \frac{1}{e}$ are satisfied for all but finitely many T_j almost surely (for both conditions, it is not difficult to obtain bounds for the probability that they are not satisfied that go to 0 faster than any power of n). Thus we obtain the following theorem:

Theorem 6.1. *Let $T_1, T_2, \dots$ be a sequence of uniformly random labelled trees, where the order of T_n is n , let X_n denote the order of a randomly chosen subtree of T_n , and let Φ_n be the distribution function of the renormalised random variable*

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}}.$$

We have

$$\sup_{x \in \mathbb{R}} \left| \Phi_n(x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \right| \rightarrow 0$$

as $n \rightarrow \infty$ almost surely.

Informally, this means that the distribution of subtree orders is close to a Gaussian distribution for almost all trees. We remark that the average subtree order of a random labelled tree T_n of order n was shown to follow a Gaussian limit distribution itself (see [15] for details).

References

- [1] Y. Alavi, P. J. Malde, A. J. Schwenk and P. Erdős, The vertex independence sequence of a graph is not constrained, *Congr. Numerantium* **58** (1987), 15–23, http://www.mta.renyi.hu/~p_erdos/1987-33.pdf.

- [2] M. Drmota, *Random Trees*, Springer, Vienna, 2009, doi:10.1007/978-3-211-75357-6.
- [3] C. D. Godsil, Matching behaviour is asymptotically normal, *Combinatorica* **1** (1981), 369–376, doi:10.1007/bf02579458.
- [4] J. Haslegrave, Extremal results on average subtree density of series-reduced trees, *J. Combin. Theory Ser. B* **107** (2014), 26–41, doi:10.1016/j.jctb.2014.02.003.
- [5] R. E. Jamison, On the average number of nodes in a subtree of a tree, *J. Combin. Theory Ser. B* **35** (1983), 207–223, doi:10.1016/0095-8956(83)90049-7.
- [6] R. E. Jamison, Monotonicity of the mean order of subtrees, *J. Combin. Theory Ser. B* **37** (1984), 70–78, doi:10.1016/0095-8956(84)90046-7.
- [7] R. E. Jamison, Mean size of subtrees of a tree, 2011, REGS in Combinatorics (University of Illinois), <http://www.math.uiuc.edu/~west/regs/meantree.html>.
- [8] J. Kahn, A normal law for matchings, *Combinatorica* **20** (2000), 339–391, doi:10.1007/pl00009835.
- [9] J. L. Martin, M. Morin and J. D. Wagner, On distinguishing trees by their chromatic symmetric functions, *J. Combin. Theory Ser. A* **115** (2008), 237–253, doi:10.1016/j.jcta.2007.05.008.
- [10] V. V. Petrov, *Limit Theorems of Probability Theory*, volume 4 of *Oxford Studies in Probability*, The Clarendon Press, New York, 1995.
- [11] H. Prodinger and S. Wagner, Bootstrapping and double-exponential limit laws, *Discrete Math. Theor. Comput. Sci.* **17** (2015), 123–144, <https://www.dmtcs.org/dmtcs-ojs/index.php/dmtcs/article/view/2781.1.html>.
- [12] L. A. Székely and H. Wang, On subtrees of trees, *Adv. in Appl. Math.* **34** (2005), 138–155, doi:10.1016/j.aam.2004.07.002.
- [13] L. A. Székely and H. Wang, Binary trees with the largest number of subtrees, *Discrete Appl. Math.* **155** (2007), 374–385, doi:10.1016/j.dam.2006.05.008.
- [14] A. Vince and H. Wang, The average order of a subtree of a tree, *J. Combin. Theory Ser. B* **100** (2010), 161–170, doi:10.1016/j.jctb.2009.05.006.
- [15] S. Wagner, Additive tree functionals with small toll functions and subtrees of random trees, in: N. Broutin and L. Devroye (eds.), *23rd International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12)*, volume AQ of *DMTCS Proc.*, pp. 67–80, 2012, <https://hal.inria.fr/hal-01197234/>.
- [16] S. Wagner and H. Wang, On the local and global means of subtree orders, *J. Graph Theory* **81** (2016), 154–166, doi:10.1002/jgt.21869.
- [17] H. Wang, Centroid, leaf-centroid, and internal-centroid, *Graphs Combin.* **31** (2015), 783–793, doi:10.1007/s00373-013-1401-1.
- [18] W. Yan and Y.-N. Yeh, Enumeration of subtrees of trees, *Theor. Comput. Sci.* **369** (2006), 256–268, doi:10.1016/j.tcs.2006.09.002.