Subtree Sizes in Recursive Trees and Binary Search Trees: Berry-Esseen Bounds and Poisson Approximations

Michael FUCHS*
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, 300
Taiwan

June 10, 2008

Abstract

We study the number of subtrees on the fringe of random recursive trees and random binary search trees whose limit law is known to be either normal or Poisson or degenerate depending on the size of the subtree. We introduce a new approach to this problem which helps us to further clarify this phenomenon. More precisely, we derive optimal Berry-Esseen bounds and local limit theorems for the normal range and prove a Poisson approximation result as the subtree size tends to infinity.

1 Introduction and Results

In this paper, we are interested in the quantity $X_{n,k}$ that counts the number of subtrees of size k on the fringe of certain random trees of size n. This number is an important tree characteristic carrying a lot of information about the shape of a tree. It can be considered as another kind of "profile", the latter notation usually reserved for the number of nodes at a fixed level. The profile attracted a lot of attention in recent research; for random recursive trees and random binary search trees see Drmota and Hwang [8] and Fuchs, Hwang, and Neininger [14]; for other classes of random trees see [9], [16], [17]. Apart from being an important shape parameter, $X_{n,k}$ has also practical relevance due to its close relationship with some fundamental quantities of trees arising in molecular biology and genetics; see Blum and François [4] and Rosenberg [18] for more details.

In this paper, we will focus on recursive trees and binary search trees as underlying classes of trees. Due to the similarity of our approach for those two families, we will mainly restrict our attention to recursive trees. One way to define these trees is as rooted, non-plane trees with nodes labelled by positive integers, where the labels of any path from the root to the leave form an increasing sequence. We consider the uniform random model on the class of recursive trees which assumes that every tree is equally likely. The resulting random tree is then called *random recursive tree*. A different way to define a random recursive tree is by a tree evolution process: first start with the root and then attach randomly nodes, where the

*Partially supported by National Science Council under the grant NSC-95-2115-M-009-017.

parent for an incoming node is chosen uniformly from the already existing nodes. Recursive trees have found many applications in diverse fields; see [14] and references therein.

The number of subtrees of random recursive trees seems to have appeared first in a paper of Aldous [1] who obtained the weak law of large numbers

$$\frac{X_{n,k}}{n} \longrightarrow \frac{1}{k(k+1)} \qquad \text{in probability.}$$

This result was then re-derived by Devroye in [7] with a different approach. Moreover, Devroye obtained the following expressions for mean value

$$\mu_{n,k} := \mathbb{E}(X_{n,k}) = \frac{n}{k(k+1)} \qquad (k < n)$$

and variance

$$\sigma^2_{n,k} := \mathbb{V}(X_{n,k}) = \frac{(2k^2-1)n}{k(k+1)^2(2k+1)} \qquad (2k < n)$$

and proved that for fixed k a central limit theorem holds

$$\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \to N(0,1), \qquad (n \to \infty).$$

Devroye's results were re-discovered in a recent paper by Feng, Mahmoud, and Su [11]. The methods of the latter authors were however completely different, namely, they applied the contraction method and Polya urns, whereas Devroye's approach was based on central limit theorems for m-dependent random variables.

The above form of the mean value motivates the subdivision of the range of k into the following three distinct subranges: (i) *subcritical* when $1 \le k = o(\sqrt{n})$; (ii) *critical* when $k \sim c\sqrt{n}$; and (iii) *supercritical* when $k < n$ and $k/\sqrt{n} \longrightarrow \infty$. Feng, Mahmoud, and Su posed in [11] the problem of extending the central limit theorem for fixed k to the whole subcritical range. This problem was then solved in a subsequent paper by Feng, Mahmoud, and Panholzer [10] (we will refer to this paper as Feng et al. throughout the current work). More precisely, they derived all possible limit laws of $X_{n,k}$ as k varies: in the subcritical range, the limit law of $(X_{n,k} - \mu_{n,k})/\sigma_{n,k}$ is standard normal; in the critical range, the limit law of $X_{n,k}$ is Poisson with parameter $1/c^2$; in the supercritical range, the limit law of $X_{n,k}$ is degenerate.

The method of proof proposed by Feng et al. rested on an exact expression for all factorial moments. In an unpublished paper, we independently obtained the same results by a completely different approach. More precisely, our approach was based on the crucial observation that all moments (centered or non-centered) satisfy the same type of recurrence. Proving asymptotic transfer theorems for the latter recurrence then provided a scheme for recursively obtaining first order asymptotics of all higher moments. The limit distribution is then identified via these asymptotics. Such a procedure, occasionally nicknamed "moment pumping", was successfully applied in a great variety of problems; see Chern, Fuchs, and Hwang [6] and references therein. The current situation is however more involved due to the dependence on two indices; see the above references for the profile, where a similar two-indexed situation was encountered.

One of the main aims of this paper and a companion paper is to demonstrate that our approach based on moments is very general in the sense that it applies to a great variety of problems and lends itself to refinements to give much deeper results.

First, the approach of Feng et al. can be used to work out the above result for random recursive trees and also to prove a corresponding result for random binary search trees, but it is not clear how to amend their method to other classes of random trees. Our approach however can be more straightforwardly adapted

to treat other classes of random trees. This will be done in a companion paper, where we will prove similar limit results for many other classes of random trees. These results then demonstrate that the above phenomenon exhibits great universality, thereby adding further importance to the number of subtrees as a fundamental tree characteristics.

In this paper, we will propose refinements of our recursive method above. These refinements will all be based on a method introduced by Hwang [15] for obtaining second phase change results in random m-ary search trees (for more information see Section 3). Our refined approaches will also exhibit some universality; we will demonstrate this by showing that they straightforwardly apply to random binary search trees, too. Technically, the main new challenges are again arising from the dependence of the problem on two indices.

We will use our refined methods to shed further light on the phenomenon discovered by Feng et al. In order to do so, we will prove two different type of results. First, in the subcritical range, we deduce the optimal Berry-Essen bound and a local limit theorem. In particular, from the Berry-Essen bound, we see that the convergence is getting weaker and weaker as k approaches $\sqrt{n}$, thereby identifying the latter as the critical point.

Theorem 1. *Let $k = k_n$ be a sequence with $1 \le k_n = o(\sqrt{n})$. Then,*

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \le x \right) - \Phi(x) \right| = \mathcal{O}\left(\frac{k}{\sqrt{n}} \right),$$

where the rate is (up to the implied constant) optimal.

Theorem 2. *Let $k = k_n$ be a sequence with $1 \le k_n = o(\sqrt{n})$. Then,*

$$P\left(X_{n,k} = \lfloor \mu_{n,k} + x\sigma_{n,k} \rfloor\right) = \frac{e^{-x^2/2}}{\sqrt{2\pi\sigma_{n,k}^2}} \left(1 + \mathcal{O}\left(\left(1 + |x|^3\right) \frac{k}{\sqrt{n}} \right) \right)$$

uniformly in $x = o(n^{1/6}/k^{1/3})$.

Our second explanation of the above result of Feng et al. was inspired by a study of predecessors in random mappings due to Baron, Drmota, and Mutafchiev [3]. First, observe that that the above expressions for mean value and variance imply

$$\mathbb{E}(X_{n,k}) \sim \mathbb{V}(X_{n,k}) \sim \frac{n}{k^2}$$

for $k = k_n < 2n$, where $k \to \infty$ as $n \to \infty$. This suggests that a Poisson approximation result should hold for $k \to \infty$ (with the latter range being optimal). This is in fact the case.

Theorem 3. *Let $k = k_n$, where $k < n$ and $k \to \infty$ as $n \to \infty$. Then, the total variation distance between $X_{n,k}$ and a Poisson random variable with rate $\mu_{n,k}$ tends to 0, i.e.,*

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \frac{1}{2} \sum_{l \ge 0} \left| P(X_{n,k} = l) - e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \right| \to 0, \qquad (n \to \infty).$$

The proof of the above result will rely on the local limit theorem for the normal range as well as a similar local limit theorem for the critical and supercritical range that will be obtained below. Apart from proving convergence to 0, the proof will also yield a rate which however is quite poor (see the remark at the end of Section 4 for further details).

We give a short sketch of the paper. In the next section, we give some general asymptotic transfer results and re-derive the above two explicit formulas for mean value and variance. In Section 3, we will introduce our first refined moment approach and use it to prove the above two results for the normal range. Section 4 is then dedicated to obtaining a local limit theorem for the non-normal range via a second refinement of the moment approach. This result together with the local limit theorem above will then be used to prove Theorem 3. In a final section, we will outline similar results for random binary search trees.

Notation. Throughout the paper, c_0 will denote a constant whose value might change from one occurrence to the next one.

2 Preliminaries

First, it is easy to see that $X_{n,k}$ satisfies the following recurrence

$$X_{n,k} \stackrel{d}{=} X_{I_n,k} + X^*_{n-I_n,k} - \mathbf{1}_{\{n-I_n=k\}} \qquad (k < n) \tag{1}$$

with $X_{n,k} = 0$ for $n < k$ and $X_{k,k} = 1$, where $I_n = \text{Unif}\{1, n-1\}$, $X^*_{n,k}$ is an independent copy of $X_{n,k}$, and the sequences $X_{n,k}$, $X^*_{n,k}$, and I_n are all independent.

Roughly speaking, this recurrence can be explained as follows: the number of subtrees of size k are counted by first counting the number of subtrees of size k in the left most subtree of the root and then adding the number of subtrees of size k in the remaining tree, where we have counted one subtree too much if the remaining tree itself is of size k; for a more detailed explanation we direct the reader to [11].

From this recurrence, we immediately obtain corresponding recurrences for the (centered or non-centered) moments of $X_{n,k}$, all of them having the following general shape

$$a_{n,k} = \frac{2}{n-1}\sum_{j=1}^{n-1} a_{j,k} + b_{n,k}, \qquad (k < n), \tag{2}$$

where $b_{n,k}$ is a suitable sequence, $a_{n,k} = 0$ for $n < k$, and $a_{k,k}$ is fixed.

For instance, for the mean value we get the above recurrence with $b_{n,k} = -1/(n-1)$ for $k < n$ and $a_{k,k} = 1$ and for the variance the "toll" sequence $b_{n,k}$ becomes

$$b_{n,k} = \frac{1}{n-1}\sum_{j=1}^{n-1} \left(\mu_{j,k} + \mu_{n-j,k} - \mu_{n,k} - \mathbf{1}_{\{j=n-k\}}\right)^2$$

and $a_{k,k} = 0$.

We start by solving recurrence (2).

Lemma 1. *For $k < l < n$,*

$$a_{n,k} = \frac{n}{l}a_{l,k} + 2n\sum_{l<j<n}\frac{b_{j,k}}{j(j+1)} + b_{n,k} - \frac{n(l-1)}{l(l+1)}b_{l,k} \tag{3}$$

$$= \frac{2n}{k(k+1)}a_{k,k} + 2n\sum_{k<j<n}\frac{b_{j,k}}{j(j+1)} + b_{n,k}. \tag{4}$$

Proof. Consider $(n-1)a_{n,k} - (n-2)a_{n-1,k}$ and iterate the resulting recurrence. ∎

Using the above solution of (2), we can re-derive the explicit formulas for mean value and variance obtained by Feng, Mahmoud, and Su.

First, for the mean value,

$$\begin{aligned}\mu_{n,k} &= \frac{2n}{k(k+1)}\mathbb{E}X_{k,k} - 2n\sum_{k<j<n}\frac{1}{(j-1)j(j+1)} - \frac{1}{n-1}\\ &= \frac{2n}{k(k+1)} - 2n\left(-\frac{1}{2(n-1)n} + \frac{1}{2k(k+1)}\right) - \frac{1}{n-1}\\ &= \frac{n}{k(k+1)},\end{aligned}$$

where this holds for $k < n$. Overall,

$$\mu_{n,k} = \begin{cases} n/(k(k+1)), & \text{if } k < n;\\ 1, & \text{if } k = n;\\ 0, & \text{if } k > n.\end{cases}$$

For the variance, slightly more work is necessary. Here, we are going to use (3) with $l = 2k+1$. First, for $n > 2k$

$$\begin{aligned}b_{n,k} &= \frac{1}{n-1}\sum_{j=1}^{n-1}\left(\mu_{j,k} + \mu_{n-j,k} - \mu_{n,k} - \mathbf{1}_{\{j=n-k\}}\right)^2\\ &= \frac{2}{n-1}\sum_{j<k}\left(\frac{n-j}{k(k+1)} - \frac{n}{k(k+1)}\right)^2 + \frac{1}{n-1}\left(\frac{n-k}{k(k+1)} + 1 - \frac{n}{k(k+1)}\right)^2\\ &\qquad + \frac{1}{n-1}\left(\frac{n-k}{k(k+1)} - \frac{n}{k(k+1)}\right)^2\\ &= \frac{3k^2-k+1}{3k(k+1)(n-1)} = \frac{c_k}{n-1}.\end{aligned}$$

Now, observe that all terms except the first one on the right hand side of (3) add to 0 for toll functions of the form $b_{n,k} = c_k/(n-1)$. So, for $n > 2k$,

$$\sigma^2_{n,k} = \frac{n}{2k+1}\sigma^2_{2k+1,k}.$$

It is easy to see that

$$X_{2k+1,k} = \begin{cases} 0, & \text{with probability } (k-1)(2k^2-1)/(2k^2(k+1));\\ 1, & \text{with probability } (2k^2-1)/(k^2(k+1));\\ 2, & \text{with probability } 1/(2k^2).\end{cases}$$

Thus,

$$\sigma^2_{2k+1,k} = \frac{2k^2-1}{k(k+1)^2}.$$

Plugging the latter into the above formula proves the claimed result. As for the remaining range of n, observe that for $k < n \leq 2k$, we have

$$X_{n,k} = \begin{cases} 0, & \text{with probability } 1 - n/(k(k+1)); \\ 1, & \text{with probability } n/(k(k+1)). \end{cases} \tag{5}$$

Consequently, for $k < n \leq 2k$,

$$\sigma_{n,k}^2 = \frac{n}{k(k+1)} \left(1 - \frac{n}{k(k+1)}\right).$$

Overall,

$$\sigma_{n,k}^2 = \begin{cases} (2k^2-1)n/(k(k+1)^2(2k+1)), & \text{if } n > 2k; \\ n/(k(k+1))(1 - n/(k(k+1))), & \text{if } k < n \leq 2k; \\ 0, & \text{if } n < k. \end{cases} \tag{6}$$

Subsequently, we will need the following simple transfer result.

Proposition 1. *Let $a_{k,k} = 0$.*

(i) Assume that $|b_{n,k}| \leq c_1 c_k$, where $c_1 > 0$ is a constant. Then, $|a_{n,k}| \leq 2c_1 c_k n/k$.

(ii) Assume that

$$|b_{n,k}| \leq c_1 (c_k n)^v,$$

where $c_1 > 0$ and $v > 1$. Then,

$$|a_{n,k}| \leq \frac{2c_1 v}{v-1} (c_k n)^v.$$

Proof. We first prove part (i). Therefore, we use the above lemma. Consequently,

$$|a_{n,k}| \leq 2c_1 c_k n \sum_{k<j<n} \frac{1}{j(j+1)} + c_1 c_k = \frac{2c_1 c_k n}{(k+1)} - c_1 c_k \leq \frac{2c_1 c_k n}{k}.$$

For the second part, we use a similar reasoning and obtain

$$|a_{n,k}| \leq 2c_1 c_k^v n \sum_{k<j<n} j^{v-2} + c_1 (c_k n)^v \leq \frac{2c_1 v}{v-1} (c_k n)^v.$$

This proves the claimed result. ∎

3 Subcritical Range: Berry-Esseen Bound and LLT

In this section, we are going to prove Theorem 1 and Theorem 2. Therefore, we will propose a new version of Hwang's refined method of moments which was introduced in [15] for proving second phase change results for random m-ary search trees; see also [6] and Bai, Hwang, and Tsai [2] for other applications of the latter method. As already explained in the introduction, the main new feature of the current situation is that we have to deal with a double-indexed recurrence. This will make the analysis much more involved. In particular, the crucial bound from Proposition 2 below must hold uniform in both k and n. Where the

case of n large compared with k resembles the situations encountered in previous studies, the remaining range of n and k has to be treated completely different.

Now, we will provide more details. Therefore, denote by

$$\phi_{n,k}(y) = e^{-\sigma_{n,k}^2 y^2/2}\mathbb{E}\left(e^{(X_{n,k}-\mu_{n,k})y}\right).$$

Then, (1) implies the following recurrence for $\phi_{n,k}$

$$\phi_{n,k}(y) = \frac{1}{n-1}\sum_{j=1}^{n-1}\phi_{j,k}(y)\phi_{n-j,k}(y)e^{\Delta_{j,n,k}y+\delta_{j,n,k}y^2}, \qquad (k<n)$$

with $\phi_{n,k}(y) = 1$ for $n \leq k$,

$$\Delta_{j,n,k} := \mu_{j,k} + \mu_{n-j,k} - \mu_{n,k} - \mathbf{1}_{\{j=n-k\}},$$

and

$$\delta_{j,n,k} := \left(\sigma_{j,k}^2 + \sigma_{n-j,k}^2 - \sigma_{n,k}^2\right)/2.$$

Next, denote by $\phi_{n,k}^{(m)}$ the m-th derivative of $\phi_{n,k}(y)$ at 0. From the above recurrence, we then obtain

$$\phi_{n,k}^{(m)} = \frac{2}{n-1}\sum_{j=1}^{n-1}\phi_{j,k}^{(m)} + \psi_{n,k}^{(m)}, \qquad (k<n)$$

with $\phi_{n,k}^{(m)} = 0, n \leq k$ and

$$\psi_{n,k}^{(m)} = \sum_{\substack{i_1+i_2+i_3+2i_4=m\\ 0\leq i_1,i_2<m}} \frac{m!}{i_1!i_2!i_3!i_4!}\cdot\frac{1}{n-1}\sum_{j=1}^{n-1}\phi_{j,k}^{(i_1)}\phi_{n-j,k}^{(i_2)}\Delta_{j,n,k}^{i_3}\delta_{j,n,k}^{i_4}.$$

Our main aim is to prove the following proposition.

Proposition 2. *For $n, k \geq 1$ and $m \geq 0$,*

$$|\phi_{n,k}^{(m)}| \leq m!A^m \max\left\{\left(\frac{n}{k^2}\right)^{m/3}, \frac{n}{k^2}\right\},$$

where A is a suitable constant.

We will first prove the claim for some small m. Therefore, note that it trivially holds for $m = 0, 1, 2$. Next, we consider $m = 3$. Here, (3) becomes

$$\phi_{n,k}^{(3)} = \frac{2}{n-1}\sum_{j=1}^{n-1}\phi_{j,k}^{(3)} + \psi_{n,k}^{(3)}, \qquad (k<n)$$

with $\phi_{n,k}^{(3)} = 0, n \leq k$ and

$$\psi_{n,k}^{(3)} = \frac{6}{n-1}\sum_{j=1}^{n-1}\Delta_{j,n,k}\delta_{j,n,k} + \frac{1}{n-1}\sum_{j=1}^{n-1}\Delta_{j,n,k}^3.$$

Now, observe that

$$\Delta_{j,n,k} = \begin{cases} 0, & \text{if } k < j < n-k; \\ \mathcal{O}(1), & \text{if } j = k; \\ \mathcal{O}\left(1/k\right), & \text{otherwise} \end{cases} \tag{7}$$

and

$$\delta_{j,n,k} = \begin{cases} 0, & \text{if } 2k < j < n-2k; \\ \mathcal{O}\left(1/k\right), & \text{otherwise.} \end{cases} \tag{8}$$

From those two estimates we obtain $\psi_{n,k}^{(3)} = \mathcal{O}(1/k)$, where the implied constant is absolute. Applying Proposition 1 then yields

$$\phi_{n,k}^{(3)} = \mathcal{O}\left(\frac{n}{k^2}\right).$$

This proves our assertion for $m = 3$.

Next, assume that we have already proved that

$$|\phi_{n,k}^{(m)}| \le m! A_0^m \frac{n}{k^2} \tag{9}$$

for all $n \le 2k^2$ and $m \ge 1$, where A_0 is a suitable constant. This actually is already half of the claim above and will be proved later on. To prove the other half, we have to show that

$$|\phi_{n,k}^{(m)}| \le m! A_1^m \left(\frac{n}{k^2}\right)^{m/3} \tag{10}$$

for all $n > 2k^2$ and $m \ge 0$, where A_1 is a suitable constant. W.l.o.g, we can assume that $A_1 \ge A_0$. In order to prove (10), we will use induction and (9).

We know already that (10) holds for $m = 0, 1, 2, 3$. Now, assume we have proved it for all m' with $m' < m$.

In order to prove the claim for m, we first consider $\psi_{n,k}^{(m)}$ and break it into three parts.

$$\begin{aligned} \psi_{n,k}^{(m)} &= \sum_{\substack{i_1+i_2+i_3+2i_4=m \\ 0 \le i_1, i_2 < m}} \sum_{2k^2 < j < n-2k^2} + \sum_{\substack{i_1+i_2+i_3+2i_4=m \\ 0 \le i_1, i_2 < m}} \sum_{j \le 2k^2} + \sum_{\substack{i_1+i_2+i_3+2i_4=m \\ 0 \le i_1, i_2 < m}} \sum_{j \ge n-2k^2} \\ &=: \Sigma_1 + \Sigma_2 + \Sigma_3. \end{aligned}$$

We start by treating the first sum. Note that due to (7) and (8) it simplifies to

$$\Sigma_1 = \sum_{i=1}^{m-1} \binom{m}{i} \frac{1}{n-1} \sum_{2k^2 < j < n-2k^2} \phi_{j,k}^{(i)} \phi_{n-j,k}^{(m-i)}$$

Using the induction hypotheses, the latter sum can be estimated as follows

$$\begin{aligned} |\Sigma_1| &\le m! A_1^m \sum_{i=1}^{m-1} \frac{1}{n-1} \sum_{2k^2 < j < n-2k^2} \left(\frac{j}{k^2}\right)^{i/3} \left(\frac{n-j}{k^2}\right)^{(m-i)/3} \\ &\le c_0 m! A_1^m \left(\frac{n}{k^2}\right)^{m/3} \sum_{i=1}^{m-1} \frac{\Gamma(i/3+1)\Gamma((m-i)/3+1)}{\Gamma(m/3+2)} \\ &\le c_0 (m-1)! A_1^m \left(\frac{n}{k^2}\right)^{m/3}, \end{aligned}$$

where the last step follows from Lemma 3 in [15].

As for the second sum, we first break it into the following two parts

$$\Sigma_2 = \sum_{\substack{i_1+i_2+i_3+2i_4=m\\ 0\le i_1,i_2<m}} \sum_{j\le 2k^2,\ j<n-2k^2} + \sum_{\substack{i_1+i_2+i_3+2i_4=m\\ 0\le i_1,i_2<m}} \sum_{j\le 2k^2,\ j\ge n-2k^2} =: \Sigma_{2,1}+\Sigma_{2,2}.$$

We will use the induction hypotheses and the two estimates (7) and (8) to bound the latter two parts separately. First,

$$\begin{aligned}
|\Sigma_{2,1}| &\le 2m!A_1^m \sum_{\substack{i_1+i_2+i_3+2i_4=m\\ 0\le i_1,i_2<m}} \frac{1}{i_3!i_4!}\frac{1}{n-1} \sum_{j\le 2k^2,\ j<n-2k^2} \left(\frac{n-j}{k^2}\right)^{i_2/3} C^{i_3}D^{i_4}\\
&\le c_0 m!A_1^m \sum_{i=0}^{m}\sum_{l=0}^{i} \left(\frac{n}{k^2}\right)^{l/3} \frac{1}{n-1} \sum_{1\le j<n} \left(1-\frac{j}{n}\right)^{l/3}\\
&\le c_0 m!A_1^m \sum_{i=0}^{m}\sum_{l=0}^{i} \frac{1}{l}\left(\frac{n}{k^2}\right)^{l/3} \le c_0(m-1)!A_1^m \left(\frac{n}{k^2}\right)^{m/3},
\end{aligned}$$

where the last line follows from $n > 2k^2$ (otherwise the sum would be 0 and the bound trivially holds). The second sum above we once more break into two parts

$$\Sigma_{2,2} = \sum_{\substack{i_1+i_2+i_3+2i_4=m\\ i_3\ne 0 \text{ or } i_4\ne 0}} \sum_{j\le 2k^2,\ j\ge n-2k^2} + \sum_{\substack{i_1+i_2=m\\ i_1,i_2\ge 1}} \sum_{j\le 2k^2,\ j\ge n-2k^2} =: \Sigma_{2,2,1}+\Sigma_{2,2,2}.$$

The first part, we crudely bound by

$$|\Sigma_{2,2,1}| \le c_0 m!A_0^m \frac{1}{k}\sum_{i=0}^{m}\sum_{l=0}^{i} 1 \le c_0 m^2 m!A_0^m \frac{1}{k}.$$

For the second part, we use (9) and obtain

$$|\Sigma_{2,2,2}| \le m!A_0^m \sum_{i=1}^{m-1} \frac{1}{n-1} \sum_{1\le j<n} \left(\frac{j}{k^2}\right)\left(\frac{n-j}{k^2}\right) \le c_0 m m!A_0^m \left(\frac{n}{k^2}\right)^2 \le c_0 m m!A_0^m \left(\frac{n}{k^2}\right)^{7/6},$$

where the last line follows from $n \le 4k^2$ (otherwise the sum would be 0).

The sum Σ_3 can be bounded similarly.

Overall, we obtain for $n > k$

$$|\psi_{n,k}^{(m)}| \le c_0 m!A_1^m \left(\frac{1}{m}\left(\frac{n}{k^2}\right)^{m/3} + m^2\frac{1}{k} + m\left(\frac{n}{k^2}\right)^{7/6}\right).$$

Applying Proposition 1 then yields

$$|\phi_{n,k}^{(m)}| \le c_0 m!A_1^m \left(\frac{1}{m-3}\left(\frac{n}{k^2}\right)^{m/3} + m^2\left(\frac{n}{k^2}\right) + m\left(\frac{n}{k^2}\right)^{7/6}\right).$$

For $n > 2k^2$ the latter in turn implies

$$|\phi_{n,k}^{(m)}| \le c_0\left(\frac{1}{m-3} + m^2 2^{-(m-3)/3} + m2^{-(2m-7)/6}\right) m!A_1^m \left(\frac{n}{k^2}\right)^{m/3} \le m!A_1^m \left(\frac{n}{k^2}\right)^{m/3},$$

where the last line holds for n large enough. Hence, by suitable tuning the constant A_1, the proof of (10) is finished.

Looking at (3), we see that we cannot prove (9) by the same approach. Hence, we will use a different (and more direct) proof. Since the claimed bound is for moderately small n, it is better to look at the factorial moments. Therefore, denote by $P_{n,k}(z) = \mathbb{E}\left(z^{X_{n,k}}\right)$. Then, (1) translates into

$$\begin{aligned} P_{n,k}(z) &= \frac{1}{n-1}\sum_{j=1}^{n-1} P_{j,k}(z)P_{n-j,k}(z)z^{\mathbf{1}_{\{j=n-k\}}} \\ &= \frac{1}{n-1}\sum_{j=1}^{n-1} P_{j,k}(z)P_{n-j,k}(z) - \frac{1}{n-1}\left(z-1\right)P_{n-k,k}(z), \end{aligned} \tag{11}$$

where $P_{n,k}(z) = 1$ for $n < k$ and $P_{k,k}(z) = z$. Next denote by $A_{n,k}^{(m)}$ the m-th derivative of $P_{n,k}(z)$ with respect to z evaluated at $z = 1$. The above recurrence in turn yields for $m \geq 2$

$$A_{n,k}^{(m)} = \frac{2}{n-1}\sum_{j=1}^{n-1} A_{j,k}^{(m)} + B_{n,k}^{(m)},$$

where $A_{n,k}^{(m)} = 0$ for $n \leq k$ and

$$B_{n,k}^{(m)} = \sum_{i=1}^{m-1}\binom{m}{i}\frac{1}{n-1}\sum_{j=1}^{n-1} A_{j,k}^{(i)}A_{n-j,k}^{(m-i)} - \frac{m}{n-1}A_{n-k,k}^{(m-1)}.$$

Now, we will use the same approach as above to prove the uniform bound

$$|A_{n,k}^{(m)}| \leq m!A^m\left(\frac{n}{k^2}\right)^m \tag{12}$$

with a suitable constant A and $m \geq 0$ (with the sole exception of $m = 1$ and $n = k$).

First observe that (12) trivially holds for $m = 0, 1$. Next, we look at $m = 2$. Here, we have

$$B_{n,k}^{(2)} = \frac{2}{n-1}\sum_{j=1}^{n-1} A_{j,k}^{(1)}A_{n-j,k}^{(1)} - \frac{2}{n-1}A_{n-k,k}^{(1)} = \mathcal{O}\left(\left(\frac{n}{k^2}\right)^2 + \frac{1}{n-1}\left(\frac{n}{k^2}\right)\right) = \mathcal{O}\left(\left(\frac{n}{k^2}\right)^2\right). \tag{13}$$

Note that in the above estimate, we have to be careful with $j = k$ and $j = n - k$: (i) if either $j \neq k$ or $j \neq n - k$, then we can exclude both cases from the above first sum and replace $-$ by $+$ in front of the second term; if $n = 2k$ then we can exclude the case $j = k$ from the first sum and completely drop the second term. Now, applying Proposition 1 yields (12) with $m = 2$.

For the general case, assume that the assertion holds for all m' with $m' < m$. To prove it for m, first consider

$$\begin{aligned} |B_{n,k}^{(m)}| &\leq m!A^m\sum_{i=1}^{n-1}\frac{1}{m-1}\sum_{j=1}^{n-1}\left(\frac{j}{k^2}\right)^i\left(\frac{n-j}{k^2}\right)^{m-i} + \frac{m!}{n-1}A^{m-1}\left(\frac{n-k}{k^2}\right)^{m-1} \\ &\leq c_0 m!A^m\left(\frac{n}{k^2}\right)^m\sum_{i=0}^{m}\frac{i!(m-i)!}{(m+2)!} + m!A^{m-1}\left(\frac{n}{k^2}\right)^m \\ &\leq c_0 m!A^m\left(\frac{n}{k^2}\right)^m\left(\frac{1}{m} + A^{-1}\right), \end{aligned}$$

where in the first estimate we again have to be careful with the cases $j = k$ and $j = n - k$ (but a similar remark as above holds) and the last line follows from Lemma 3 in [15]. Applying Proposition 1 then yields

$$|A_{n,k}^{(m)}| \le c_0 \left(\frac{1}{m-1} + A^{-1} \frac{m}{m-1} \right) m! A^m \left(\frac{n}{k^2} \right)^m \le m! A^m \left(\frac{n}{k^2} \right)^m,$$

where the last step follows for m and A large enough. Suitable tuning A completes the induction step.

Remark 1. Note that (12) is simpler than the bound previously obtained for $\phi_{n,k}^{(m)}$. This is essentially due to the more simpler nature of the toll sequence $B_{n,k}^{(m)}$ in this case. In particular, computing $\phi_{n,k}^{(3)}$ would reveal that a similar simple bound would not hold in the previous situation.

Moreover, we should mention that the following weaker result for $n \le ck^2$ was already obtained in Feng et al.

$$|A_{n,k}^{(m)}| \le c_0 (m-1)! \frac{n}{k^2}$$

for $m \ge 2$. Actually, this bound would be sufficient for us as well. The reason why we proved the above stronger bound is because the proof is more in the spirit of our paper and hence makes our paper more self-contained. Moreover, we will encounter a very similar situation in the next section, too.

Using the following well-known relation between moments and factorial moments

$$\mathbb{E} X_{n,k}^m = \sum_{i=1}^m S(n,i) A_{n,k}^{(i)},$$

where $S(n,i)$ are the Stirling numbers of second kind, we obtain for $k < n \le 2k^2$ and $m \ge 1$

$$\begin{aligned} |\mathbb{E} X_{n,k}^m| &\le \sum_{i=1}^m S(n,i) i! A^i \left(\frac{n}{k^2} \right)^i \\ &\le \frac{n}{k^2} (2A)^m \sum_{i=1}^m S(n,i) m(m-1) \cdots (m-i+1) \\ &= \frac{n}{k^2} (2A)^m m^m \le m! \bar{A}^m \frac{n}{k^2}, \end{aligned}$$

where the last line follows by the definition of Stirling numbers of second kind and Stirling's formula. The latter estimate then in turn implies for $k < n \le 2k^2$ and $m \ge 1$

$$\begin{aligned} |\mathbb{E}(X_{n,k} - \mu_{n,k})^m| &\le \sum_{i=1}^m \binom{m}{i} \mathbb{E} X_{n,k}^i \, (\mu_{n,k})^{m-i} + (\mu_{n,k})^m \\ &\le m! \bar{A}^m \frac{n}{k^2} \sum_{i=1}^m \frac{2^{m-i}}{(m-i)!} + \left(\frac{n}{k^2} \right)^m \\ &\le m! \tilde{A}^m \frac{n}{k^2}. \end{aligned}$$

Finally, we have

$$\phi_{n,k}^{(m)} = \sum_{i=0,\ i \text{ even}}^m \binom{m}{i} \frac{i!}{(i/2)!} \left(-\sigma_{n,k}^2/2 \right)^{i/2} \mathbb{E}(X_{n,k} - \mu_{n,k})^{m-i}.$$

Consequently, for $k < n \leq 2k^2$ and $m \geq 2$

$$|\phi_{n,k}^{(m)}| \leq m! \tilde{A}^m \frac{n}{k^2} \sum_{i=0,\ i \text{ even}}^{m} \frac{C^{i/2}}{(i/2)!} + D^m \left(\frac{n}{k^2}\right)^{m/2} \leq m! A_0^m \frac{n}{k^2}.$$

This concluded the proof of (10).

Overall, our proof of Proposition 2 is finished.

Next, we will apply the latter proposition to deduce the following two results for the characteristic function of $X_{n,k}$

$$\varphi_{n,k}(y) := \mathbb{E}\left(e^{iy(X_{n,k}-\mu_{n,k})/\sigma_{n,k}}\right) = e^{-iy\mu_{n,k}/\sigma_{n,k}} P_{n,k}\left(e^{iy/\sigma_{n,k}}\right).$$

Proposition 3. *Let $k = k_n$ with $1 \leq k = o(\sqrt{n})$.*

(i) For n large enough,

$$\varphi_{n,k}(y) = e^{-y^2/2}\left(1 + \mathcal{O}\left(|y|^3 \frac{k}{\sqrt{n}}\right)\right)$$

uniformly for y with $|y| \leq \epsilon n^{1/6}/k^{1/3}$, where $\epsilon > 0$ is sufficiently small.

(ii) For n large enough and $|y| \leq \pi\sigma_{n,k}$,

$$|\varphi_{n,k}(y)| \leq e^{-\epsilon y^2/2}$$

where $\epsilon > 0$ is sufficiently small.

Proof. The first part follows from Taylor series expansion and Proposition 2; see [15].

So, we just have to concentrate on the second part. Here, we will prove a slightly more general result: for $n \geq 3, 1 \leq k < n$ and $|y| \leq \pi$, we have

$$|P_{n,k}\left(e^{iy}\right)| \leq e^{-\epsilon y^2(n/k^2+c/k)} \tag{14}$$

with (dependent) constants ϵ and c that will be chosen below. From this, the above claim is then immediate.

In order to prove the latter result, first observe that from (11)

$$|P_{n,k}\left(e^{iy}\right)| \leq \frac{1}{n-1} \sum_{j=1}^{n-1} |P_{j,k}\left(e^{iy}\right)||P_{n-j,k}\left(e^{iy}\right)|.$$

We will establish our claim by induction on n. Note that for $n > 3k$ the induction step is deduced from the above recurrence as follows

$$\begin{aligned}
|P_{n,k}\left(e^{iy}\right)| &\leq \frac{2}{n-1} \sum_{j \leq k} |P_{j,k}\left(e^{iy}\right)||P_{n-j,k}\left(e^{iy}\right)| + \frac{1}{n-1} \sum_{k<j<n-k} |P_{j,k}\left(e^{iy}\right)||P_{n-j,k}\left(e^{iy}\right)| \\
&\leq e^{-\epsilon y^2(n/k^2+c/k)} \left(\frac{2}{n-1} \sum_{j \leq k} e^{\epsilon y^2 j/k^2} + \left(1 - \frac{2k}{n-1}\right) e^{-\epsilon y^2 c/k}\right) \\
&\leq e^{-\epsilon y^2(n/k^2+c/k)} \left(\frac{2k}{n-1} e^{\epsilon y^2(1/k+1/k^2)} + \left(1 - \frac{2k}{n-1}\right) e^{-\epsilon y^2 c/k}\right) \\
&\leq e^{-\epsilon y^2(n/k^2+c/k)} \left(\frac{2k}{n-1}\left(1 + c_0 \epsilon y^2/k\right) + \left(1 - \frac{2k}{n-1}\right)\left(1 - c_1 \epsilon y^2 c/k\right)\right),
\end{aligned}$$

where c_0, c_1 are suitable absolute constants (here, we need that ϵc is small; see below). Now, choose $c > 2c_0/c_1$. Then,

$$|P_{n,k}\left(e^{iy}\right)| \leq e^{-\epsilon y^2(n/k^2+c/k)}\left(1-\epsilon y^2\frac{(n-1-2k)cc_1-2kc_0}{k(n-1)}\right) \leq e^{-\epsilon y^2(n/k^2+c/k)}.$$

This establishes the induction step for $n > 3k$. As for the remaining range, we first consider $n \leq 2k$. Here, we can directly work with (5) and obtain

$$\begin{aligned}|P_{n,k}\left(e^{iy}\right)|^2 &= \left(\left(1-\frac{n}{k(k+1)}\right)+\cos(y)\frac{n}{k(k+1)}\right)^2+\sin^2(y)\frac{n^2}{k^2(k+1)^2}\\ &= 1+2\sigma_{n,k}^2\left(\cos y-1\right) \leq e^{-2\epsilon_0 y^2\sigma_{n,k}^2},\end{aligned}$$

where ϵ_0 is chosen so small that $\cos(y)-1 \leq -\epsilon_0 y^2$ for $|y| \leq \pi$. Thus,

$$|P_{n,k}\left(e^{iy}\right)| \leq e^{-\epsilon_0 y^2\sigma_{n,k}^2} \leq e^{-(\epsilon_0/(6+6c))y^2(n/k^2+c/k)}.$$

So, with $\epsilon := \epsilon_0/(6+6c)$, the above claim is established (note that ϵc is also small as required). The final range of $2k < n \leq 3k$ can be treated by a similar (but slightly more complicated) computation, where for this range, we have

$$X_{n,k} = \begin{cases} 0, & \text{with probability } 1+\mu_{n,k}(\mu_{n,k}-3)/2+\sigma_{n,k}^2/2;\\ 1, & \text{with probability } \mu_{n,k}(2-\mu_{n,k})-\sigma_{n,k}^2;\\ 2, & \text{with probability } \mu_{n,k}(\mu_{n,k}-1)/2+\sigma_{n,k}^2/2.\end{cases}$$

Of course, the above chosen ϵ has then to be adjusted accordingly.

As for $n = 3, 1 \leq k < 3$, note that these cases are already covered by the previous arguments. This concludes the induction proof and hence (14) is established. Finally, we remark that (14) becomes wrong for all other possible choices of n and k. ∎

Theorem 1 and Theorem 2 follow now from the latter proposition by standard tools; see [15].

4 Poisson Approximation

In this section, we will prove Theorem 3. The proof will be based on the local limit theorem from the supercritical range together with the following result.

Proposition 4. *For $k < n$ and $n \to \infty$,*

$$P(X_{n,k} = l) = e^{-\mu_{n,k}}\frac{(\mu_{n,k})^l}{l!}+\mathcal{O}\left(\frac{n}{k^3}\right)$$

uniformly in l.

This result will be proved with yet another variation of Hwang's refined method of moments. Therefore, denote by

$$\tilde{\phi}_{n,k}(z) = e^{-\mu_{n,k}(z-1)}\mathbb{E}\left(z^{X_{n,k}}\right) = e^{-\mu_{n,k}(z-1)}P_{n,k}(z),$$

where we use here the convention that $\mu_{k,k} = 0$ (this will simplify the proof below). Then, (11) can be rewritten to

$$\tilde{\phi}_{n,k}(z) = \frac{1}{n-1}\sum_{j=1}^{n-1} \tilde{\phi}_{j,k}(z)\tilde{\phi}_{n-j,k}(z)e^{\Lambda_{j,n,k}(z-1)} - \frac{1}{n-1}(z-1)\tilde{\phi}_{n-k,k}(z)e^{\lambda_{n,k}(z-1)},$$

where $\tilde{\phi}_{n,k}(z) = 1$ for $n < k$, $\tilde{\phi}_{k,k} = z$, and

$$\Lambda_{j,n,k} = \mu_{j,k} + \mu_{n-j,k} - \mu_{n,k}, \qquad \lambda_{n,k} = \mu_{n-k,k} - \mu_{n,k}.$$

Next, let $\tilde{\phi}_{n,k}^{(m)}$ be the m-th derivative of $\tilde{\phi}_{n,k}(z)$ with respect to z evaluated at $z = 1$. Then, the above recurrence in turn implies for $m \geq 2$

$$\tilde{\phi}_{n,k}^{(m)} = \sum_{j=1}^{n-1} \tilde{\phi}_{j,k}^{(m)} + \tilde{\psi}_{n,k}^{(m)},$$

where $\tilde{\phi}_{n,k}^{(m)} = 0$ for $n \leq k$ and

$$\tilde{\psi}_{n,k}^{(m)} = \sum_{\substack{i_1+i_2+i_3=m \\ 0\leq i_1,i_2<m}} \binom{m}{i_1, i_2, i_3} \frac{1}{n-1}\sum_{j=1}^{n-1} \tilde{\phi}_{j,k}^{(i_1)}\tilde{\phi}_{n-j,k}^{(i_2)}\Lambda_{j,n,k}^{i_3} - \frac{m}{n-1}\sum_{i=0}^{m-1}\binom{m-1}{i}\tilde{\phi}_{n-k,k}^{(i)}\lambda_{n,k}^{m-1-i}.$$

We will use a similar method as in the previous section to obtain the following uniform bound.

Proposition 5. *For all $n > k$ and $m \geq 0$,*

$$|\tilde{\phi}_{n,k}^{(m)}| \leq m! A^m \left(\frac{n}{k^3}\right)^{m/2},$$

where A is a suitable constant.

Note that the latter bound (once proved) will be hold as well for $n \leq k$ with the only exception being $n = k$ and $m = 1$. So, the situation here is very similar to the situation encountered in the proof of (12). This is also to main reason for setting $\mu_{k,k} = 0$ in the definition of $\tilde{\phi}_{n,k}(z)$.

As for the proof of the above bound, we will proceed by induction. Note that our claim trivially holds for $m = 0, 1$. We next consider the case $m = 2$. Here, direct computation yields for $n > k$

$$\tilde{\phi}_{n,k}^{(2)} = \sigma_{n,k}^2 - \mu_{n,k}$$

Now, we use (6). First, for $n > 2k$,

$$\tilde{\phi}_{n,k}^{(2)} = \frac{n}{2k+1}\tilde{\phi}_{2k+1,k}^{(2)} = -\frac{n(3k+2)}{k(k+1)^2(2k+1)} = \mathcal{O}\left(\frac{n}{k^3}\right).$$

Next, for $k < n \leq 2k$,

$$\tilde{\phi}_{n,k}^{(2)} = \frac{-n^2}{k^2(k+1)^2} = \mathcal{O}\left(\frac{n}{k^3}\right).$$

Overall, the claim is proved for $n = 2$.

Now, assume that it holds for all $m' < m$. We want to show that it holds for m as well. Therefore, observe that $1/k \leq (n/k^3)^{1/2}$. Using this together with the induction hypothesis, we obtain

$$
\begin{aligned}
|\tilde{\psi}^{(m)}_{n,k}| &= \Big| \sum_{\substack{i_1+i_2+i_3=m \\ 0\leq i_1,i_2<m}} \binom{m}{i_1,i_2,i_3} \frac{1}{n-1} \sum_{j=1}^{n-1} \tilde{\phi}^{(i_1)}_{j,k} \tilde{\phi}^{(i_2)}_{n-j,k} \Lambda^{i_3}_{j,n,k} - \frac{m}{n-1} \sum_{i=0}^{m-1} \binom{m-1}{i} \tilde{\phi}^{(i)}_{n-k,k} \lambda^{m-1-i}_{n,k} \Big| \\
&\leq m! \sum_{\substack{i_1+i_2+i_3=m \\ 0\leq i_1,i_2<m}} \frac{1}{i_3!} \frac{1}{(n-1)} A^{i_1+i_2} \sum_{j=1}^{n-1} \left(\frac{j}{k^3}\right)^{i_1/2} \left(\frac{n-j}{k^3}\right)^{i_2/2} \left(\frac{C}{k}\right)^{i_3} \\
&\qquad + \frac{m!}{n-1} A^{m-1} \sum_{i=0}^{m-1} \frac{1}{(m-1-i)!} \left(\frac{n}{k^3}\right)^{i/2} \left(\frac{D}{k}\right)^{m-1-i} \\
&\leq c_0 \left(m! \left(\frac{n}{k^3}\right)^{m/2} \sum_{i=0}^{m} A^i \sum_{l=0}^{i} \frac{\Gamma(l/2+1)\Gamma((i-l)/2+1)}{\Gamma(i/2+2)} + m! A^{m-1} \left(\frac{n}{k^3}\right)^{m/2} \right) \\
&\leq c_0 m! A^m \left(\frac{n}{k^3}\right)^{m/2} \left(\frac{1}{m} + \frac{1}{A}\right),
\end{aligned}
$$

where the last line follows from Lemma 3 in [15]. Note that in the first estimate above we have to be careful with the case where the induction hypothesis does not hold (see the remark below Proposition 5); however a similar cancellation as explained below (13) in the previous section takes place. Now, we can apply Proposition 1 and obtain

$$
|\tilde{\phi}^{(m)}_{n,k}| \leq c_0 \left(\frac{1}{m-2} + \frac{m}{m-2} \frac{1}{A} \right) m! A^m \left(\frac{n}{k^3}\right)^{m/2} \leq m! A^m \left(\frac{n}{k^3}\right)^{m/2},
$$

where the last estimate holds for m and A large enough. By suitable tuning A the proof of the proposition is finished.

From the previous proposition, we can now deduce Proposition 4.

Proof of Proposition 4. First, observe that the assertion is trivial for $k \leq cn^{1/3}$ with $c > 0$. Hence, we can restrict ourselves to k with $k \geq cn^{1/3}$, where c is large.

Next, from Proposition 5, we get

$$
\begin{aligned}
P_{n,k}(z) - e^{\mu_{n,k}(z-1)} &= e^{\mu_{n,k}(z-1)} \left(\tilde{\phi}_{n,k}(z) - 1 \right) \\
&= e^{\mu_{n,k}(z-1)} \mathcal{O}\left(\sum_{m=2}^{\infty} \frac{|\tilde{\phi}^{(m)}_{n,k}|}{m!} |z-1|^m \right) \\
&= e^{\mu_{n,k}(z-1)} \mathcal{O}\left(\sum_{m=2}^{\infty} \left(A \frac{\sqrt{n}}{k^{3/2}} |z-1| \right)^m \right) \\
&= e^{\mu_{n,k}(z-1)} \mathcal{O}\left(\frac{n}{k^3} \right),
\end{aligned}
$$

where the last line holds for all z with $|z-1| \leq \epsilon k^{3/2}/\sqrt{n}$ with ϵ suitable small. Finally, by making the above c large enough, we see that the above expansion holds uniformly in z with $|z| = 1$.

Next, we apply Cauchy's formula. Hence,

$$\begin{aligned} P(X_{n,k}=l) &= \frac{1}{2\pi i}\int_{|z|=1} P_{n,k}(z) z^{-l-1}\mathrm{d}z \\ &= \frac{1}{2\pi i}\int_{|z|=1} e^{\mu_{n,k}(z-1)} z^{-l-1}\left(1+\mathcal{O}\left(\frac{n}{k^3}\right)\right)\mathrm{d}z \\ &= e^{-\mu_{n,k}}\frac{(\mu_{n,k})^l}{l!}+\mathcal{O}\left(\frac{n}{k^3}\right) \end{aligned}$$

as it was claimed. ∎

Now, we are ready to prove our main result.

Proof of Theorem 3. First, consider k with $k \geq n^{2/5}$. Then, break the sum in the total variation distance into two parts

$$\sum_{l\geq 0}\left|P(X_{n,k}=l)-e^{-\mu_{n,k}}\frac{(\mu_{n,k})^l}{l!}\right| = \sum_{|l-\mu_{n,k}|\leq \eta\sqrt{\mu_{n,k}}} |\cdots| + \sum_{|l-\mu_{n,k}|>\eta\sqrt{\mu_{n,k}}} |\cdots| =: \Sigma_1+\Sigma_2.$$

The second part can be easily estimated by Tschebyscheff's inequality

$$\Sigma_2 \leq \frac{\sigma_{n,k}^2}{\eta^2\mu_{n,k}}+\frac{1}{\eta^2} = \mathcal{O}\left(\eta^{-2}\right).$$

For the first sum, we use Proposition 4 and obtain

$$\Sigma_1 = \mathcal{O}\left(\eta\sqrt{\mu_{n,k}}\frac{n}{k^3}\right) = \mathcal{O}\left(\eta n^{3/2}k^{-4}\right).$$

Now, choose $\eta = k^{4/3}n^{-1/2}$. Then,

$$d_{TV}(X_{n,k},\mathrm{Po}(\mu_{n,k})) = \mathcal{O}\left(\eta^{-2}+\eta n^{3/2}k^{-4}\right) = \mathcal{O}\left(nk^{-8/3}\right) = \mathcal{O}\left(n^{-1/15}\right)$$

which proves the result for $k \geq n^{2/5}$.

Next, we consider k with $n^{1/5} \leq k < n^{2/5}$ and choose $\eta = n^{1/12}k^{-1/6}$. We again use the above partition, where Σ_2 is estimated as above. As for bounding Σ_1, we use Theorem 2 together with the following two expansions

$$e^{-\mu_{n,k}}\frac{(\mu_{n,k})^l}{l!} = \frac{1}{\sqrt{2\pi\mu_{n,k}}}\exp\left(-\frac{(l-\mu_{n,k})^2}{2\mu_{n,k}}\right)\left(1+\mathcal{O}\left(\eta\mu_{n,k}^{-1}\right)\right)$$

and

$$\frac{1}{\sqrt{2\pi\sigma_{n,k}^2}}\exp\left(-\frac{(l-\mu_{n,k})^2}{2\sigma_{n,k}^2}\right) = \frac{1}{\sqrt{2\pi\mu_{n,k}}}\exp\left(-\frac{(l-\mu_{n,k})^2}{2\mu_{n,k}}\right)\left(1+\mathcal{O}\left(\eta^2k^{-1}\right)\right),$$

where the latter expansions hold uniformly for all l with $|l-\mu_{n,k}| \leq \eta\sqrt{\mu_{n,k}}$. Consequently,

$$\Sigma_1 = \mathcal{O}\left(\eta^3k^{-1}+\eta^4n^{-1/2}k\right) = \mathcal{O}\left(n^{-1/20}\right).$$

Overall, we obtain

$$d_{TV}(X_{n,k},\mathrm{Po}(\mu_{n,k})) = \mathcal{O}\left(n^{-1/20}+\eta^{-2}\right) = \mathcal{O}\left(n^{-1/20}\right).$$

For the final range of k with $k < n^{1/5}$, we choose $\eta = k^{1/5}$. Then, similar as above, we obtain

$$d_{TV}(X_{n,k},\mathrm{Po}(\mu_{n,k})) = \mathcal{O}\left(k^{-2/5}+n^{-7/50}\right).$$

Hence, our result is established. ∎

Remark 2. As already mentioned in the introduction, the above proof also gives a rate for the convergence of the total variation distance to 0. However, the rate is quite poor and can be further improved by incorporating estimates for higher moments. We will restrain ourselves from doing this here and postpone the issue of deriving sharp bounds for the total variation distance to the related work [5] which discusses applications of the techniques of this paper to random trees arising from molecular biology and genetics.

5 Binary Search Trees

In Feng et al. a similar result as described in the introduction was derived for random binary search trees as well; see [7], Feng, Miao, Su [12], Flajolet, Gourdon, Martinez [13] for earlier and/or related results in this direction.

A random binary search tree is recursively build from a random permutation of the sequence of records $1, \dots, n$ as follows: place the first record into the root and direct the other records either to the left or right subtree according to whether the record is smaller or larger then the record stored in the root; proceed like this to recursively build the left and right subtree. Binary search trees are a fundamental data structure in computer science and they have found numerous applications; see [14] and references therein.

Again, denote by $X_{n,k}$ the number of subtrees of size k on the fringe of a random binary search of size n. Then, $X_{n,k}$ satisfies the following recurrence

$$X_{n,k} \stackrel{d}{=} X_{I_n,k} + X^*_{n-1-I_n,k} \qquad (k < n)$$

with $X_{n,k} = 0$ for $n < k$ and $X_{k,k} = 1$, where $I_n = \text{Unif}\{0, n-1\}$, $X^*_{n,k}$ is an independent copy of $X_{n,k}$, and the sequences $X_{n,k}$, $X^*_{n,k}$ and I_n are all independent.

The latter recurrence can be similarly explained as the corresponding recurrence for recursive trees; see the paragraph below (1) in Section 2.

Now, the above recurrence again implies that all (centered or non-centered) moments satisfy a recurrence of the following shape

$$a_{n,k} = \frac{2}{n} \sum_{j=0}^{n-1} a_{j,k} + b_{n,k},$$

where $b_{n,k}$ is a suitable sequence (again called "toll sequence"), $a_{n,k} = 0$ for $n < k$ and $a_{k,k}$ is fixed. Using the same method that was already applied to (2) the above recurrence can easily be solved

$$a_{n,k} = \frac{2(n+1)}{(k+1)(k+2)} a_{k,k} + 2(n+1) \sum_{k<j<n} \frac{b_{j,k}}{(j+1)(j+2)} + b_{n,k},$$

where $n > k$.

We see already here that things are very similar as in the previous situation of random recursive trees. So, our methods will run through with only minor modifications and yield similar results for random binary search trees, too. It should be pointed out that this is quite different to the situation encountered in Feng et al. whose approach applied to random binary search trees was technically much more involved than when applied to random recursive trees.

Due to the similarities to random recursive trees, we do not give any details and instead only state the final results. The reader should have no difficulties to use the tools introduced in the previous sections to work out full proofs.

First for the mean value and variance, we obtain the following explicit formulas

$$\mu_{n,k} := \mathbb{E}(X_{n,k}) = \frac{2(n+1)}{(k+1)(k+2)}, \qquad (n > k)$$

and

$$\sigma^2_{n,k} := \mathbb{V}(X_{n,k}) = \frac{2k(4k^2+5k-3)(n+1)}{(k+1)(k+2)^2(2k+1)(2k+3)}, \qquad (n > 2k+1).$$

Our results in this situation then read as follows.

Theorem 4. *(i) (Berry-Esseen bound) Let $k = k_n$ be a sequence with $1 \le k_n = o(\sqrt{n})$. Then,*

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \le x\right) - \Phi(x) \right| = \mathcal{O}\left(\frac{k}{\sqrt{n}}\right),$$

where the rate is (up to the implied constant) optimal.

(ii) (Local limit theorem) Let $k = k_n$ be a sequence with $1 \le k_n = o(\sqrt{n})$. Then,

$$P\left(X_{n,k} = \lfloor \mu_{n,k} + x\sigma_{n,k} \rfloor\right) = \frac{e^{-x^2/2}}{\sqrt{2\pi\sigma^2_{n,k}}} \left(1 + \mathcal{O}\left(\left(1+|x|^3\right)\frac{k}{\sqrt{n}}\right)\right)$$

uniformly in $x = o(n^{1/6}/k^{1/3})$.

(iii) (Poisson approximation) Let $k = k_n$, where $k < n$ and $k \to \infty$ as $n \to \infty$. Then, the total variation distance between $X_{n,k}$ and a Poisson random variable with rate $\mu_{n,k}$ tends to 0, i.e.,

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \frac{1}{2}\sum_{l\ge 0} \left| P(X_{n,k} = l) - e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \right| \to 0, \qquad (n \to \infty).$$

6 Conclusion

In this paper, we introduced a new approach to the limit law of the number of subtrees on the fringe of random trees. Our new approach is based on the method of moments and its refinement. Compared to previous approaches, our method is capable of yielding much deeper results and it applies more uniformly to different types of random trees. We demonstrated here the validity of the first claim by considerable refining recent results of Feng, Mahmound, and Panholzer. In particular, our results further explain why the number of subtrees exhibits the phenomenon discovered by the latter authors.

Our method is likely to have many more applications. In particular, other classes of random trees seem to be treatable by our approach in a similar fashion, too. This will be postponed to future work.

Acknowledgments

We thank Hsien-Kuei Hwang for many helpful suggestions and comments. Moreover, we are grateful to Alois Panholzer for pointing out the paper of Baron, Drmota, and Mutafchiev which inspired us to prove the Poisson approximation result. Finally, remarks and comments of the referee are acknowledged as well.

References

[1] D. J. Aldous (1991). Asymptotic fringe distributions for general families of random trees, *Annals of Applied Probability*, **1**, 228-266.

[2] Z.-D. Bai, H.-K. Hwang, T.-H. Tsai (2003). Berry-Esseen bounds for the number of maxima in planar regions, *Electronic Journal of Probability*, **8**, 26 pages.

[3] G. Baron, M. Drmota, L. Mutafchiev (1996). Predecessors in random mappings, *Combinatorics, Probability and Computing*, **5**, 317-335.

[4] M. G. B. Blum and O. François (2005). Minimal clade size and external branch length under the neutral coalescent. *Advances in Applied Probability*, **37**, 647-662.

[5] H. Chang and M. Fuchs (2008). Limit theorems for patterns in phylogenetic trees, manuscript.

[6] H.-H. Chern, M. Fuchs, H.-K. Hwang (2007). Phase changes in random point quadtrees, *ACM Transactions on Algorithms* **3**, 51 pages.

[7] L. Devroye (1991). Limit laws for local counters in random binary search trees, *Random Structures and Algorithms*, **2**, 303-315.

[8] M. Drmota and H.-K. Hwang (2005). Profiles of random trees: correlation and width of random recursive trees and binary search trees, *Advances in Applied Probability*, **37**, 321-341.

[9] M. Drmota, S. Janson, R. Neininger (2008). A functional limit theorem for the profile of search trees, *Annals of Applied Probability*, **18**, 288-333.

[10] Q. Feng, H. Mahmoud, A. Panholzer (2008). Phase changes in subtree varieties in random recursive trees and binary search trees, *SIAM Journal on Discrete Mathematics*, **22**, 160-184.

[11] Q. Feng, H. Mahmoud, C. Su (2007). On the variety of subtrees in a random recursive tree, Technical report, The George Washington University, Washington, DC.

[12] Q. Feng, B. Miao, C. Su (2006). On the subtrees of binary search trees, *Chinese Journal of Applied Probability and Statistics*, **22**, 304-310.

[13] P. Flajolet, X. Gourdon, C. Martinez (1997). Patterns in random binary search trees, *Random Structures and Algorithms*, **11**, 223-244.

[14] M. Fuchs, H.-K. Hwang, R. Neininger (2006). Profiles of random trees: Limit theorems for random recursive trees and binary search trees, *Algorithmica*, **46**, 367-407.

[15] H.-K. Hwang (2003). Second phase changes in random m-ary search trees and generalized quicksort: convergence rates, *Annals of Probability*, **31**, 609-629.

[16] H.-K. Hwang (2007). Profiles of random trees: plane-oriented recursive trees, *Random Structures and Algorithms*, **30**, 380-413.

[17] H.-K. Hwang, P. Nicodème, G. Park, W. Szpankowski (2008), Profiles of tries, submitted.

[18] N. A. Rosenberg (2006). The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. *Annals of Combinatorics*, **10**, 129-146.