J. Math. Biol. (2010) 60:481–512
DOI 10.1007/s00285-009-0275-6

Mathematical Biology

Limit theorems for patterns in phylogenetic trees

Huilan Chang · Michael Fuchs

Received: 10 September 2008 / Revised: 12 April 2009 / Published online: 23 May 2009

Abstract Studying the shape of phylogenetic trees under different random models is an important issue in evolutionary biology. In this paper, we propose a general framework for deriving detailed statistical results for patterns in phylogenetic trees under the Yule–Harding model and the uniform model, two of the most fundamental random models considered in phylogenetics. Our framework will unify several recent studies which were mainly concerned with the mean value and the variance. Moreover, refined statistical results such as central limit theorems, Berry–Esseen bounds, local limit theorems, etc., are obtainable with our approach as well. A key contribution of the current study is that our results are applicable to the whole range of possible sizes of the pattern.

Keywords Phylogenetic trees · Patterns · Moments · Central limit theorems · Poisson approximations

Mathematics Subject Classification (2000) 92B05 · 05C05 · 60F05

1 Introduction

Phylogenetic trees are the standard tool in evolutionary biology for depicting the ancestor history of a set of given species (or taxa); see page 117 in Darwin (1859). Consequently, their properties have been extensively studied. In particular, the investigation of the probabilistic behavior of parameters related to the shape of phylogenetic trees under different random models has evolved into a major issue in recent decades.

H. Chang · M. Fuchs (✉)
Department of Applied Mathematics, National Chiao Tung University,
Hsinchu 300, Taiwan
e-mail: mfuchs@math.nctu.edu.tw

The reason for this is multi-fold: such results can be used in statistical tests, e.g., for testing the appropriateness of a random model; they enhance our understanding of the process that generates the data; they yield conclusions about possible outcomes of evolution; etc.; for further motivation we refer the reader to Blum et al. (2006a), Mooers and Heard (1997), Mooers and Heard (2002), Rosenberg (2006) and references therein.

First, we will make precise the notation of phylogenetic trees. Throughout this paper, phylogenetic trees will be binary trees, where the external nodes represent the species and the internal nodes represent their ancestors. Moreover, all trees considered will be rooted meaning that we assume that the set of species has a common ancestor. Finally, branch lengths will be ignored, i.e., we will just look at the topology of the tree. So, the family of phylogenetic trees of size n is the family of plane, rooted, unlabelled binary trees with n external nodes (and consequently $n-1$ internal nodes).

Next, we will equip the family of phylogenetic trees of size n with a random model. In this paper, we will consider the two most fundamental random models of evolutionary biology: the *Yule–Harding model* and the *uniform model*; see Semple and Steel (2003).

First, the Yule–Harding model (see Harding 1971 and Yule 1924) is defined by a tree evolution process: the tree grows by choosing at random one of the leaves and replacing it by a cherry (an internal node with two children); we stop when a tree with n external nodes is constructed. This is the top-down construction of a phylogenetic tree of size n under the Yule–Harding model. Alternatively, a bottom-up construction can be used as well: start with n external nodes and successively choose a random pair and coalesce the two nodes; stop when only one node (the root) is left. It is easy to see that the random models arising from these two constructions are the same. Moreover, it is easy to see as well that the Yule–Harding model is the same as the permutation model of binary search trees considered in computer science; see Blum et al. (2006b).

The usage of the Yule–Harding model is well motivated since it provides an easy way of mimicking how the data might have evolved over time. Note, however, that it does not assign the same probability to every phylogenetic tree of size n. This motivates a second model which does assign the same probability and is hence called *uniform model* (or PDA model see Blum et al. 2006a). Although less motivated from a practical point of view, the usage of this model is justified as well by a couple of theoretical results; see Aldous (1991) and McKenzie and Steel (1991). Moreover, this model has also been investigated in computer science, where it is called the *Catalan model*; see Fill and Kapur (2004) and references therein. The name comes from the fact that the number of phylogenetic trees of size n is given by the $n-1$-st Catalan number, subsequently denoted by C_{n-1} (a proof of this can be found in standard textbook on enumerative combinatorics such as Stanley 1997, Stanley 1999 and Flajolet and Sedgewick 2009).

This paper will be concerned with a study of statistical properties of the occurrence of patterns in phylogenetic trees under the above two random models. Here, the word "pattern" will be used in a rather broad sense, namely, any subset of the set of all phylogenetic trees of a fixed size k will be considered as pattern. Moreover, throughout this work, we will fix the notation $X_{n,k}$ to denote the number of occurrences of this pattern in a random phylogenetic tree of size n.

Table 1 Values of p_k

Pattern	Yule–Harding model	Uniform model
k-pronged nodes	1	1
k-caterpillars	$2^{k-2}/(k-1)!$	$2^{k-2}/C_{k-1}$
Nodes with minimal clade size k	$2/(k-1)$	$2C_{k-2}/C_{k-1}$

For both random models above, $X_{n,k}$ satisfies

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

where $X_{n,k}$, $X^*_{n,k}$, and I_n are independent random variables, $X^*_{n,k}$ has the same distribution as $X_{n,k}$, and I_n is the size of the left subtree of the root. This distributional recurrence is nothing more than the mathematical formulation of the trivial observation that the number of occurrences of a pattern is the sum of the number of occurrences of the pattern in the left and in the right subtree of the root. The initial conditions of this recurrence are given by $X_{n,k} = 0$ for $n < k$ and $X_{k,k}$ is a Bernoulli random variable with success probability p_k that equals the probability that a phylogenetic tree of size k belongs to our pattern. In order to avoid ambiguity, we will assume that $p_k > 0$ throughout this work. Note that this probabilistic description of $X_{n,k}$ already implies that stochastic properties of $X_{n,k}$ just depend on p_k and not on the specific pattern.

In order to make the above more lucid, we recall some of the patterns previously considered in literature. The first and most straightforward pattern is the set of all trees of size k. The root of such a subtree in a phylogenetic tree of size n is called *k-pronged node*; see Rosenberg (2006) and McKenzie and Steel (2000) for the special case of $k = 2$. Hence, in this situation, $X_{n,k}$ counts the number of such nodes. The second pattern is the set of *k-caterpillars* also considered in Rosenberg (2006). Here, the pattern consists of phylogenetic trees of size k that have an internal node which is descendent of all other internal nodes. A final pattern is given by the set of all phylogenetic trees of size k with either left or right subtree of the root empty. Here, $X_{n,k}$ is the same as the number of taxa with *minimal clade size* k if $k \geq 3$; see Blum and François (2005). A related parameter was also considered in computer science; see Drmota et al. (2008b). The probabilities p_k for these three patterns under the above two random models are easily obtained and collected in Table 1.

The aim of this paper is to study moments of $X_{n,k}$ as well as more refined properties such as limit laws, rates of convergence, local limit theorems, etc. Therefore, we will use the setting of (1). Consequently, our setting will contain all three cases discussed above as special cases. As for k-pronged nodes and k-caterpillars, mean and variance were derived in Rosenberg (2006) under the Yule–Harding model by a bottom-up approach. We will re-derive these results using (1). So, in contrast to Rosenberg (2006), our approach will be top-down. We will see that our approach is technically easier and also allows us to derive higher moments and more refined properties. Here, we should mention that for k-pronged nodes our results were already sketched in Fuchs (2008); see also Feng et al. (2008) for related results. As for the uniform model, only

results on k-pronged nodes with $k = 2$ have been obtained before; see McKenzie and Steel (2000). Mean value and variance of the number of nodes with minimal clade size k have been derived in Blum and François (2005) under the Yule–Harding model. Moreover, in the latter paper, the authors also derived a central limit theorem of $X_{n,k}$ for fixed k. Again, we will re-derive all those results and add many new ones as well as prove corresponding results under the uniform model.

Before we start to explain our results in more details, it should be mentioned that the behavior of $X_{n,k}$ for fixed k is well understood. Here, a detailed description of the stochastic properties of $X_{n,k}$ follows from results in computer science; see Flajolet et al. (1997), Hwang (2003), Hwang and Neininger (2002) for the Yule–Harding model and Flajolet and Sedgewick (2009) for the uniform model. However, from an application point of view, results which hold uniformly in k are more desirable. So, one of the main contributions of this paper is to provide results where k is allowed to grow with n. Proving such results will involve multivariate asymptotics which in recent years has evolved into a major topic in analytic combinatorics; see Drmota and Hwang (2005), Drmota et al. (2008a), Fuchs et al (2007), Hwang (2007), Hwang (2008) and Pemantle (2000), Pemantle and Wilson (2002), Pemantle and Wilson (2004), Pemantle and Wilson (2008).

Now, we will discuss our findings in more details. For the sake of simplicity, we will choose the number of nodes with minimal clade size k as a guiding example. For our general results, we refer the reader to Sects. 2 and 3.

First, we consider the Yule–Harding model. Here, we have the following results for mean value and variance.

Theorem 1 *For $k < n$,*

$$\mathbb{E}(X_{n,k}) = \frac{4n}{(k-1)k(k+1)}$$

and

$$\mathrm{Var}(X_{n,k}) = \begin{cases} \frac{4(4k^4-27k^2+11)n}{(k-1)^2k(k+1)^2(2k-1)(2k+1)}, & \text{if } n > 2k; \\ \frac{4(4k^3-32k+20)}{(k-1)^2(k+1)^2(2k-1)}, & \text{if } n = 2k; \\ \frac{4(k^3-k-4n)n}{(k-1)^2k^2(k+1)^2}, & \text{if } k < n < 2k. \end{cases}$$

In particular, for $k < n$ and $k \to \infty$,

$$\mathbb{E}(X_{n,k}) \sim \mathrm{Var}(X_{n,k}) \sim \frac{4n}{k^3}, \quad (n \to \infty).$$

Moreover, the first order asymptotic of all higher moments will be derived as well. This will then allow us to study limit laws. Note that for fixed k, a central limit theorem follows from previous results; see Hwang and Neininger (2002). We will show that the central limit theorem continues to hold for some range with $k \to \infty$. Moreover, we will derive the Berry–Esseen bound as well.

Theorem 2 *Let* $3 \le k = o(\sqrt[3]{n})$. *Then,*

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{X_{n,k} - \mathbb{E}(X_{n,k})}{\sqrt{\mathrm{Var}(X_{n,k})}} \le x \right) - \Phi(x) \right| = \mathcal{O}\left(\frac{k^{3/2}}{\sqrt{n}} \right),$$

where $\Phi(x)$ *denotes the distribution function of the standard normal random variable.*

The above range will turn out to be the largest possible range for which a central limit theorem holds. Hence, the normal distribution just provides a good approximation for k small. From a practical point of view, this is quite unsatisfactory. Therefore, we will show that approximating by a Poisson random variable works well in a much larger range and is hence more desirable.

Theorem 3 *Let* $k < n$ *and* $k \to \infty$. *Then,*

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

More precisely, we have

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mathbb{E}(X_{n,k}))) = \begin{cases} \mathcal{O}\left(1/k^{2\alpha/(3\alpha+1)}\right), & \text{if } n \ge k^3; \\ \mathcal{O}\left(n/k^{3+2\alpha}\right), & \text{if } n < k^3, \end{cases}$$

where $\alpha = 2m/(2m+1)$ *with a fixed (but arbitrary)* $m \ge 1$.

So, only for very small k one should use the normal distribution as an approximation. For the remaining range, a Poisson random variable yields a better approximation.

As for the proofs of these results, we will use the elementary approach (in the sense that complex analysis is avoided) from Fuchs (2008); for more details see Sect. 2.

Now, we turn to the uniform model. Here, we will prove similar results as above. First, for mean value and variance, we have the following theorem.

Theorem 4 (i) *For constant* k,

$$\mathbb{E}(X_{n,k}) = \frac{2C_{k-2}}{4^{k-1}} n + \mathcal{O}(1), \quad (n \to \infty),$$

and

$$\mathrm{Var}(X_{n,k}) = \left(\frac{2C_{k-2}}{4^{k-1}} - \frac{(2k-1)C_{k-2}^2}{4^{2k-3}} \right) n + \mathcal{O}(1), \quad (n \to \infty).$$

(ii) *For* $k \to \infty$ *and* $n - k \to \infty$,

$$\mathbb{E}(X_{n,k}) \sim \mathrm{Var}(X_{n,k}) \sim \frac{n}{\sqrt{4\pi k^3}} \quad (n \to \infty).$$

(iii) *For constant* $n - k = l \geq 0$,

$$\mathbb{E}(X_{n,k}) = \frac{(l+1)C_l}{2^{2l+1}} + \mathcal{O}\left(\frac{1}{n}\right), \quad (n \to \infty),$$

and

$$\mathrm{Var}(X_{n,k}) = \frac{(l+1)C_l}{2^{2l+1}}\left(1 - \frac{(l+1)C_l}{2^{2l+1}}\right) + \mathcal{O}\left(\frac{1}{n}\right), \quad (n \to \infty).$$

Moreover, we again have a central limit theorem with Berry–Esseen bound that holds for small k.

Theorem 5 *Let* $3 \leq k = o(n^{2/3})$. *Then,*

$$\sup_{-\infty<x<\infty}\left|P\left(\frac{X_{n,k} - \mathbb{E}(X_{n,k})}{\sqrt{\mathrm{Var}(X_{n,k})}} \leq x\right) - \Phi(x)\right| = \mathcal{O}\left(\frac{k^{3/4}}{\sqrt{n}}\right).$$

Again, the central limit theorem does not hold beyond this range. However, as above, we have a Poisson approximation result.

Theorem 6 *Let* $k \to \infty$ *and* $n - k \to \infty$. *Then,*

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

More precisely,

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mathbb{E}(X_{n,k}))) = \begin{cases} \mathcal{O}\left(1/\sqrt{k}\right), & \text{if } n \geq k^{3/2}; \\ \mathcal{O}\left(n/k^2\right), & \text{if } k^{1+\epsilon} \leq n < k^{3/2}; \\ \mathcal{O}\left(\mathbb{E}(X_{n,k})\right), & \text{if } n < k^{1+\epsilon}, \end{cases}$$

where $\epsilon > 0$ *is an arbitrarily small constant.*

So, $X_{n,k}$ is again well approximated by a Poisson random variable unless k is either very small or very large. In the latter two cases, the Poisson random variable has to be replaced by a standard normal random variable and a Bernoulli random variable, respectively.

The proofs of the results in the uniform case will follow from a rather different method compared to the approach used in the Yule–Harding case. This is largely due to the fact that involved generating functions can be solved explicitly. Hence, the results are obtained more easily by using complex-analytic tools. For more details we refer the reader to Sect. 3.

In order to conclude the introduction, we give a brief sketch of the paper. First, since we intend to prove results for two different random models, we will split the

paper into two parts; the first part (Sect. 2) will discuss the Yule Harding model and the second part (Sect. 3) the uniform model. Every part will consist of three sections which will be concerned with deriving results for moments, discussing the validity of the central limit theorem, and finally proving our Poisson approximation results, respectively. We will end the paper with some concluding remarks.

Notations Subsequently, ϵ will always denote a small positive real number and c a large constant. Moreover, the values of both ϵ and c may be different from one occurrence to the next.

2 Yule–Harding model

In this section, we are going to investigate the statistical properties of $X_{n,k}$ under the Yule–Harding model. As already mentioned in the introduction, we will use the elementary method introduced in Fuchs (2008) which was based on studying the underlying recurrence satisfied by the moments and applying the method of moments and its refinements.

In fact, some of the results below will follow similarly as in Fuchs (2008) and in order to avoid repetition we will not give any details. We will, however, discuss in details a more simplified proof of the central limit theorem (without the Berry–Esseen bound) and refined bounds for the total variation distance; the former will constitute a simplification of the approach in Feng et al. (2008) for k-pronged nodes as well.

2.1 Moments

In this section, we will compute moments of $X_{n,k}$. Therefore, we will work out in details the approach briefly sketched in Fuchs (2008) for k-pronged nodes; see Feng et al. (2008) for a similar approach. This part should be compared with Rosenberg (2006) where the same results are proved for k-pronged nodes and k-caterpillars, but with a more complicated approach.

First, note that $X_{n,k}$ satisfies (1) with

$$P(I_n = j) = \frac{1}{n-1}, \quad 1 \le j \le n-1,$$

where the latter follows straightforwardly from the probabilistic description of the random model.

Next, we consider $P_{n,k}(z) = \mathbb{E}(\exp\{X_{n,k}z\})$. Then, (1) translates into

$$P_{n,k}(z) = \frac{1}{n-1}\sum_{j=1}^{n-1} P_{j,k}(z)P_{n-j,k}(z), \quad (n > k)$$

with $P_{n,k}(z) = 1$ for $n < k$ and $P_{k,k}(z) = p_k(e^z - 1) + 1$. Differentiating this recurrence m times and evaluating at $z = 0$ give a corresponding recurrence for the m-th

moment. The key observation is that all these recurrences are of the following type

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

where $a_{n,k} = 0$ for $n < k$, $b_{n,k} = 0$ for $n \le k$, $a_{k,k}$ is determined by the initial conditions and $b_{n,k}$ for $n > k$ is a function of moments of lower order. Moreover, a similar computation reveals that also all central moments satisfy a recurrence of the latter shape. So, we start by investigating this recurrence.

First, this recurrence can be easily solved. Therefore, consider $(n-1)a_{n,k} - (n-2)a_{n-1,k}$ and iterate the resulting recurrence. Then, for $k < l < n$,

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

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

In order to find the mean value, we set $b_{n,k} = 0$ and $a_{k,k} = p_k$ in the last formula above. Then,

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

Obviously, $\mu_{k,k} = p_k$ and $\mu_{n,k} = 0$ for $n < k$.

The computation of the variance $\sigma^2_{n,k} := \mathrm{Var}(X_{n,k})$ is slightly more involved. First, note that the variance satisfies the above recurrence with

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

A lengthy computation gives

$$b_{n,k} = \begin{cases} \frac{2(k-1)(3k-2)p_k^2}{3(n-1)k(k+1)}, & \text{if } n > 2k; \\ \frac{4(k-1)(3k^2-k-1)p_k^2}{3k(k+1)^2(2k-1)}, & \text{if } n = 2k. \end{cases}$$

Then, by plugging this into (2) with $l = 2k$,

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

where $n > 2k$. So, we first need to compute $\sigma^2_{2k,k}$.

Lemma 1 *We have*

$$\sigma^2_{2k,k} = \frac{2(4k^2 + 2k - 2 + (k^2 - 14k + 9)p_k)p_k}{(k+1)^2(2k-1)}.$$

Proof First, observe that $X_{2k,k}$ only takes on the values 0, 1, 2. A simple combinatorial argument shows that $P(X_{2k,k} = 2) = p_k^2/(2k-1)$. The other probabilities are easily computed from the latter by

$$\frac{4p_k}{k+1} = \mathbb{E}(X_{2k,k}) = 2P(X_{2k,k} = 2) + P(X_{2k,k} = 1)$$

and $P(X_{2k,k} = 2) + P(X_{2k,k} = 1) + P(X_{2k,k} = 0) = 1$. Overall,

$$X_{2k,k} = \begin{cases} 2, & \text{with probability } p_k^2/(2k-1); \\ 1, & \text{with probability } 4p_k/(k+1) - 2p_k^2/(2k-1); \\ 0, & \text{with probability } 1 - 4p_k/(k+1) + p_k^2/(2k-1). \end{cases}$$

The result follows now by a straightforward computation. □

Plugging the latter result into the formula above together with some simplifications yields

$$\sigma^2_{n,k} = \frac{2(4k^3 + 4k^2 - k - 1 - (11k^2 - 5)p_k)p_k n}{k(k+1)^2(2k-1)(2k+1)},$$

for $n > 2k$. Finally, for the range $n < 2k$, we deduce from the above result for the mean value

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

To sum up, we have proved the following result.

Proposition 1 *We have,*

$$\mu_{n,k} = \begin{cases} \frac{2p_k n}{k(k+1)}, & \text{if } n > k; \\ p_k, & \text{if } n = k; \\ 0, & \text{if } n < k \end{cases}$$

and

$$\sigma^2_{n,k} = \begin{cases} \frac{2(4k^3+4k^2-k-1-(11k^2-5)p_k)p_k n}{k(k+1)^2(2k-1)(2k+1)}, & \text{if } n > 2k; \\ \frac{2(4k^2+2k-2+(k^2-14k+9)p_k)p_k}{(k+1)^2(2k-1)}, & \text{if } n = 2k; \\ \frac{2(k^2+k-2np_k)p_k n}{k^2(k+1)^2}, & \text{if } k < n < 2k; \\ p_k(1-p_k), & \text{if } n = k; \\ 0, & \text{if } n < k. \end{cases}$$

The latter result immediately gives the following corollary.

Corollary 1 *As $k \to \infty$, we have*

$$\mu_{n,k} \sim \sigma_{n,k}^2 \sim \frac{2p_k n}{k^2}, \qquad (n \to \infty).$$

Moreover, higher moments could be computed by this approach as well. The computation, however, becomes more and more involved. We will see in the next section that this problem becomes easier when only the main order term in the asymptotic expansion is sought.

2.2 Central limit theorem

Now, we will turn to limiting distributions of $X_{n,k}$. First, it is well known that for fixed k, the following central limit theorem holds (see Hwang and Neininger 2002)

$$\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Moreover, the Berry–Esseen bound was derived by Hwang (2003) and is $\mathcal{O}(n^{-1/2})$.

Our first result extends the range of validity of the above central limit theorem.

Theorem 7 *Let $p_k n/k^2 \to \infty$. Then,*

$$\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Due to the result for constant k, we can focus on $k \to \infty$ as $n \to \infty$. First, consider $\bar{P}_{n,k}(z) = \mathbb{E}(\exp\{(X_{n,k} - \mu_{n,k})z\})$. Then, (1) translates into

$$\bar{P}_{n,k}(z) = \frac{1}{n-1} \sum_{j=1}^{n-1} \bar{P}_{j,k}(z) \bar{P}_{n-j,k}(z) e^{\Delta_{n,j,k} z} \qquad (n > k) \tag{4}$$

with $\bar{P}_{n,k}(z) = 1$ for $n < k$ and $\bar{P}_{k,k}(z) = e^{-p_k z}(p_k e^z - p_k + 1)$ and

$$\Delta_{n,j,k} = \mu_{j,k} + \mu_{n-j,k} - \mu_{n,k}.$$

Next, we introduce $A_{n,k}^{(m)} = \mathbb{E}(X_{n,k} - \mu_{n,k})^m$. Differentiating (4) m times ($m \geq 1$) and setting $z = 0$ reveal

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

where $A_{n,k}^{(m)} = 0$ for $n < k$, $A_{k,k}^{(m)} = p_k(1-p_k)^m + (1-p_k)(-p_k)^m$ and

$$B_{n,k}^{(m)} = \sum_{\substack{i_1+i_2+i_3=m \\ 0\le i_1,i_2<m}} \binom{m}{i_1, i_2, i_3} \frac{1}{n-1} \sum_{j=1}^{n-1} A_{j,k}^{(i_1)} A_{n-j,k}^{(i_2)} \Delta_{n,j,k}^{i_3}. \tag{5}$$

We first consider the case $m = 2$. Here, $A_{n,k}^{(2)} = \sigma_{n,k}^2$ and as already mentioned in the previous section

$$B_{n,k}^{(2)} = \frac{1}{n-1} \sum_{j=1}^{n-1} \Delta_{n,j,k}^2.$$

Even though we have obtained an asymptotic expansion of the variance as $k \to \infty$ in Corollary 1, we give here a second and more simplified proof of this result. Therefore, observe that for $n > k$

$$\Delta_{n,j,k} = \begin{cases} 0, & \text{if } k < j < n-k; \\ \mathcal{O}(p_k/k), & \text{if } j < k,\ j \neq n-k \text{ or } j > n-k,\ j \neq k; \\ p_k + \mathcal{O}(p_k/k), & \text{if } j = k,\ j \neq n-k \text{ or } j = n-k,\ j \neq k; \\ 2p_k + \mathcal{O}(p_k/k), & \text{if } j = k \text{ and } j = n-k \end{cases}$$

which yields

$$A_{k,k}^{(2)} = p_k(1-p_k), \quad B_{n,k}^{(2)} = \begin{cases} 2p_k^2/(n-1) + \mathcal{O}\left(p_k^2/k^2\right), & \text{if } n \neq 2k; \\ 4p_k^2/(2k-1) + \mathcal{O}\left(p_k^2/k^2\right), & \text{if } n = 2k, \end{cases}$$

where all implied constants are absolute. Plugging this into (3) then reveals

$$\begin{aligned} \sigma_{n,k}^2 &= \frac{2p_k(1-p_k)n}{k^2} + 4p_k^2 n \sum_{k<j<n} \frac{1}{(j-1)j(j+1)} + \mathcal{O}\left(\frac{p_k n}{k^3} + \frac{p_k^2}{n}\right), \\ &= \frac{2p_k n}{k^2} + \mathcal{O}\left(\frac{p_k n}{k^3} + \frac{p_k^2}{n}\right), \end{aligned}$$

where the implied constant is absolute. So, we obtain the bound

$$\sigma_{n,k}^2 = \mathcal{O}\left(\frac{p_k n}{k^2}\right) \tag{6}$$

which holds uniformly in n and k with $n > k$. Moreover, if $k \to \infty$ as $n \to \infty$, we have

$$\sigma_{n,k}^2 \sim \frac{2p_k n}{k^2}. \tag{7}$$

Next, we are going to show that both (6) and (7) can be extended to all moment as well. Therefore, we will use (3) together with induction. This is a standard method that is called "moment-pumping" and was applied to numerous problems; see Chern (2007) and references therein.

We first extend (6).

Proposition 2 *For $m \geq 1$,*

$$A_{n,k}^{(m)} = \mathcal{O}\left(\max\left\{\frac{p_k n}{k^2}, \left(\frac{p_k n}{k^2}\right)^{m/2}\right\}\right)$$

uniformly in $n > k$.

Proof First, note that the claim trivially holds for $m = 1$ and was proved for $m = 2$ above. Now, we assume that the claim holds for all m' with $m' < m$. We will establish that it holds for m as well.

Before starting with the proof, we need a notation. For fixed k denote by j_k the smallest integer such that $p_k j_k / k^2 \geq 1$.

Now, we can start with the proof. First consider (5) and break the involved sums into two parts

$$B_{n,k}^{(m)} = \sum_{\substack{i_1+i_2+i_3=m\\ 0\leq i_1,i_2<m}} \sum_{j=k \text{ or } j=n-k} + \sum_{\substack{i_1+i_2+i_3=m\\ 0\leq i_1,i_2<m}} \sum_{\substack{j=1\\ j\neq k,\ j\neq n-k}}^{n-1} =: \Sigma_1 + \Sigma_2.$$

We will bound the two parts separately. We start with the first one. Therefore, observe that

$$\begin{aligned}\Sigma_1 &= \mathcal{O}\left(\frac{1}{n} \sum_{\substack{i_1+i_2+i_3=m\\ 0\leq i_1,i_2<m}} \binom{m}{i_1, i_2, i_3} A_{k,k}^{(i_1)} A_{n-k,k}^{(i_2)} \Delta_{n,k,k}^{i_3}\right)\\ &= \mathcal{O}\left(\frac{p_k}{n} \sum_{i=0}^{m-1} A_{n-k,k}^{(i)}\right) = \mathcal{O}\left(\frac{p_k}{n}\left(\frac{p_k n}{k^2}\right)^{(m-1)/2} + \frac{p_k}{n}\right).\end{aligned}$$

Next, we will consider the second part which we again break into two parts

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

The second of the two sums can be bounded as follows

$$\Sigma_{2,2} = \mathcal{O}\left(\frac{1}{n}\sum_{\substack{i_1+i_2+i_3=m\\ 0\le i_1,i_2<m,\ 0<i_3}}\binom{m}{i_1,i_2,i_3}\sum_{j<k}A_{j,k}^{(i_1)}A_{n-j,k}^{(i_2)}\Delta_{n,j,k}^{i_3}\right)$$
$$= \mathcal{O}\left(\frac{p_k}{kn}\sum_{i=0}^{m-1}\sum_{j<k}A_{n-j,k}^{(i)}\right) = \mathcal{O}\left(\frac{p_k}{n}\left(\frac{p_k n}{k^2}\right)^{(m-1)/2} + \frac{p_k}{n}\right).$$

So, what is left is to bound $\Sigma_{2,1}$. Therefore, we break it into three parts

$$\Sigma_{2,1} \le \sum_{\substack{i_1+i_2=m\\ 0\le i_1,i_2<m}}\sum_{j\le j_k,\ j\ne k} + \sum_{\substack{i_1+i_2=m\\ 0\le i_1,i_2<m}}\sum_{j_k<j<n-j_k} + \sum_{\substack{i_1+i_2=m\\ 0\le i_1,i_2<m}}\sum_{j\ge n-j_k,\ j\ne n-k}$$
$$=: \Sigma_{2,1,1} + \Sigma_{2,1,2} + \Sigma_{2,1,3}.$$

Due to symmetry, the bound for the first and last of the three sums will be the same. Therefore, we will concentrate on the first one which can be treated as follows

$$\Sigma_{2,1,1} = \frac{1}{n-1}\sum_{i=1}^{m-1}\binom{m}{i}\sum_{j\le j_k,\ j\ne k}A_{j,k}^{(i)}A_{n-j,k}^{(m-i)}$$
$$= \begin{cases}\mathcal{O}\left((j_k/n)\left(p_k n/k^2\right)^{(m-1)/2}\right), & n > 2j_k;\\ \mathcal{O}\left((p_k n/k^2)^2\right), & n \le 2j_k.\end{cases}$$

Finally, we have

$$\Sigma_{2,1,2} = \frac{1}{n-1}\sum_{i=1}^{m-1}\binom{m}{i}\sum_{j_k<j<n-j_k}A_{j,k}^{(i)}A_{n-j,k}^{(m-i)}$$
$$= \mathcal{O}\left(\left(\frac{p_k n}{k^2}\right)^{m/2}\sum_{i=1}^{m-1}\binom{m}{i}\int_0^1 x^{i/2}(1-x)^{(m-i)/2}\mathrm{d}x\right) = \mathcal{O}\left(\left(\frac{p_k n}{k^2}\right)^{m/2}\right).$$

Collecting all terms above yields

$$B_{n,k}^{(m)} = \mathcal{O}\left(\left(\frac{p_k n}{k^2}\right)^{m/2} + \left(\frac{p_k n}{k^2}\right)^2 + \frac{p_k}{k}\right) \tag{8}$$

for $m \geq 5$. Now we plug this together with $A_{k,k}^{(m)} = \mathcal{O}(p_k)$ into (3) and obtain

$$\begin{aligned} & A_{n,k}^{(m)} \\ & = \mathcal{O}\left(\frac{p_k n}{k^2}\right) + \mathcal{O}\left(\frac{p_k^{m/2} n}{k^m} \sum_{k<j<n} j^{m/2-2} + \frac{p_k^2 n}{k^4} \sum_{k<j<n} 1 + \frac{p_k n}{k} \sum_{k<j<n} j^{-2}\right) + B_{n,k}^{(m)} \\ & = \mathcal{O}\left(\frac{p_k n}{k^2} + \left(\frac{p_k n}{k^2}\right)^{m/2}\right). \end{aligned}$$

For $m = 4$, we have to replace the second term in (8) by $(p_k n/k^2)^{3/2}$. The claim then follows as above.

For $m = 3$, we have to be slightly more careful. Here the second term in (8) has to be replaced by the above bound for $\Sigma_{2,1,1}$. Since the above arguments still work for the first and third term in (8), we just have to concentrate on the contribution of the new second term. Therefore, observe that

$$n \sum_{k<j<n} \frac{\Sigma_{2,1,1}}{j^2} = \mathcal{O}\left(\frac{p_k^2 n}{k^4} \sum_{k<j\leq 2j_k} 1 + \frac{p_k j_k n}{k^2} \sum_{2j_k<j<n} j^{-2}\right) = \mathcal{O}\left(\frac{p_k n}{k^2}\right).$$

Hence, also in this case, we obtain the claimed bound. This concludes the induction step and hence our claim is established. □

Next, we will refine our previous result for the range where the claimed central limit theorem holds.

Proposition 3 *For $p_k n/k^2 \to \infty$ and $k \to \infty$ as $n \to \infty$, we have*

$$A_{n,k}^{(2m-1)} = o\left(\left(\frac{p_k n}{k^2}\right)^{m-1/2}\right);$$
$$A_{n,k}^{(2m)} \sim g_m \left(\frac{2p_k n}{k^2}\right)^m,$$

for $m \geq 1$, where $g_m = (2m)!/(2^m m!)$.

Proof We again use induction on m. Note that for $m = 1$ the first assertion is trivial and the second assertion follows from (6). Now, assume the assertions hold for all m' with $m' < m$. We will show that they hold for m as well.

Therefore, we again first consider (5). Note that the proof of the last proposition yields

$$B_{n,k}^{(l)} = \sum_{i=1}^{l-1} \binom{l}{i} \frac{1}{n-1} \sum_{j_k<j<n-j_k} A_{j,k}^{(i)} A_{n-j,k}^{(l-i)} + o\left(\left(\frac{p_k n}{k^2}\right)^{l/2}\right),$$

where j_k is defined as in the proof of the last proposition. We fix an $\epsilon > 0$ and split the sum into three parts

$$\sum_{i=1}^{l-1} \sum_{j_k < j \le \epsilon n} + \sum_{i=1}^{l-1} \sum_{\epsilon n < j < (1-\epsilon)n} + \sum_{i=1}^{l-1} \sum_{(1-\epsilon)n \le j < n - j_k} =: \Sigma_1 + \Sigma_2 + \Sigma_3.$$

We first concentrate on the second of the three parts. Therefore, we consider two cases. First, if $l = 2m - 1$ is odd, then either i or $2m - 1 - i$ is odd. Hence,

$$\begin{aligned}\Sigma_2 &= o\left(\left(\frac{p_k n}{k^2}\right)^{m-1/2} \sum_{i=1}^{2m-2} \binom{2m-1}{i} \int_{\epsilon}^{1-\epsilon} x^{i/2}(1-x)^{(2m-1-i)/2} \mathrm{d}x\right) \\ &= o\left(\left(\frac{p_k n}{k^2}\right)^{m-1/2}\right).\end{aligned}$$

Second, if $l = 2m$ is even, then the above reasoning shows that the sum over the odd indices i has the same bound as above. As for the sum over the even indices, we have

$$\begin{aligned}&\sum_{i=1}^{m-1} \binom{2m}{2i} \frac{1}{n} \sum_{\epsilon n < j < (1-\epsilon)n} A_{j,k}^{(2i)} A_{n-j,k}^{(2m-2i)} \\ &\quad \sim \left(\frac{2p_k n}{k^2}\right)^m \sum_{i=1}^{m-1} \binom{2m}{2i} g_i g_{m-i} \int_{\epsilon}^{1-\epsilon} x^i (1-x)^{m-i} \mathrm{d}x.\end{aligned}$$

So, overall

$$\Sigma_2 \sim \left(\frac{2p_k n}{k^2}\right)^m \sum_{i=1}^{m-1} \binom{2m}{2i} g_i g_{m-i} \int_{\epsilon}^{1-\epsilon} x^i (1-x)^{m-i} \mathrm{d}x.$$

As for the first and third sum above, using the uniform bound from our last proposition shows that

$$\Sigma_1 = \Sigma_3 = \mathcal{O}(\epsilon \Sigma_2).$$

So, by letting $\epsilon \to 0$, we see that the main contribution comes from the second sum. Overall, we have

$$\begin{aligned}B_{n,k}^{(2m-1)} &= o\left(\left(\frac{p_k n}{k^2}\right)^{m-1/2}\right); \\ B_{n,k}^{(2m)} &\sim \bar{g}_m \left(\frac{2p_k n}{k^2}\right)^m,\end{aligned}$$

where

$$\bar{g}_m = \sum_{i=1}^{m-1} \binom{2m}{2i} g_i g_{m-i} \frac{\Gamma(i+1)\Gamma(m-i+1)}{\Gamma(m+2)} = \frac{m-1}{m+1} g_m.$$

Now, we plug this together with $A_{k,k}^{(l)} = \mathcal{O}(p_k)$ into (3). This gives

$$A_{n,k}^{(l)} = \mathcal{O}\left(\frac{p_k n}{k^2}\right) + 2n \sum_{k<j<\epsilon n} \frac{B_{j,k}^{(l)}}{j(j+1)} + 2n \sum_{\epsilon n<j<n} \frac{B_{j,k}^{(l)}}{j(j+1)} + B_{n,k}^{(l)},$$

where $\epsilon > 0$ is again fixed. Using our uniform bound from the last proposition again shows that the main contribution comes from the third and fourth term. First, for $l = 2m-1$, we have

$$A_{n,k}^{(2m-1)} = o\left(\frac{p_k^{m-1/2} n}{k^{2m-1}} \sum_{j<n} j^{m-5/2} + \left(\frac{n}{k^3}\right)^{m-1/2}\right) = o\left(\left(\frac{p_k n}{k^2}\right)^{m-1/2}\right).$$

Finally, for $l = 2m$, we have

$$A_{n,k}^{(2m)} \sim 2\bar{g}_m \left(\frac{2p_k n}{k^2}\right)^m \int_{\epsilon}^{1} x^{m-2} \mathrm{d}x + \bar{g}_m \left(\frac{2p_k n}{k^2}\right)^m.$$

Letting $\epsilon \to 0$ and simplifying the right hand side yield the claimed result also for even moments. This concludes the induction step and hence the proof is finished as well. □

Theorem 7 now follows from the previous proposition by the theorem of Fréchet–Shohat; see Loève (1977).

As for the Berry–Esseen bound, we can use the method from Fuchs (2008) which constitutes a refinement of the previous approach. Since there are only minor technical differences compared to the situation discussed in Fuchs (2008), we only state the result and omit the proof details.

Theorem 8 *Let $p_k n/k^2 \to \infty$. Then,*

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

2.3 Poisson approximation

With the proof method introduced in the last section, it can be shown that the limit distribution of $X_{n,k}$ is Poisson for $p_k n/k^2 \to c \ge 0$; see Feng et al. (2008) for similar results. Hence, Theorem 7 gives the maximal range for which the central limit theorem holds.

Instead of proving such a result, we will prove the stronger Poisson approximation result stated in the introduction for the special case of nodes with given minimal clade size. Before, we can do so, we need local limit theorems for $X_{n,k}$. The following two results also follow from the method in Fuchs (2008).

Proposition 4 (i) *Let $p_k n/k^2 \to \infty$. Then,*

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

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

(ii) *Let $k < n$. Then,*

$$P(X_{n,k} = l) = e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} + \mathcal{O}\left(\frac{p_k^2 n}{k^3}\right)$$

uniformly in l.

From the last proposition together with the bounds from Proposition 2 of the last section, we will obtain quite sharp bounds for the total variation distance between $X_{n,k}$ and a Poisson random variable with the same mean. Note that these bounds improve upon the bounds given in Fuchs (2008).

Theorem 9 *Let $k < n$ and $k \to \infty$. Then,*

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \begin{cases} \mathcal{O}\left((p_k/k)^{\alpha/(3\alpha+1)}\right), & \text{if } \mu_{n,k} \geq 1; \\ \mathcal{O}\left((p_k/k)^{\alpha} \cdot \mu_{n,k}\right), & \text{if } \mu_{n,k} < 1, \end{cases}$$

where $\alpha = 2m/(2m+1)$ with a fixed (but arbitrary) $m \geq 1$.

Proof We start by considering the case where $\mu_{n,k} \geq 1$. Here, we will split the sum in the formula of the total variation distance into two parts

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \frac{1}{2} \sum_{|l-\mu_{n,k}| < \eta\sqrt{\mu_{n,k}}} |\cdots| + \frac{1}{2} \sum_{|l-\mu_{n,k}| \geq \eta\sqrt{\mu_{n,k}}} |\cdots| =: \Sigma_1 + \Sigma_2. \tag{9}$$

where η will be chosen below. In order to bound the second part, observe that from Proposition 2,

$$P(|X_{n,k} - \mu_{n,k}| \geq \eta\sqrt{\mu_{n,k}}) = \mathcal{O}\left(\eta^{-2m}\right) \tag{10}$$

for all $m \geq 1$. Moreover, the same bound holds as well when $X_{n,k}$ is replaced by $\mathrm{Po}(\mu_{n,k})$. Consequently,

$$\Sigma_2 = \mathcal{O}\left(\eta^{-2m}\right)$$

for all $m \geq 1$.

Now, we consider three cases. First assume that $\mu_{n,k} \geq k^2/p_k^2$ and choose $\eta = (k/p_k)^\epsilon$ with $\epsilon > 0$ sufficiently small. By Proposition 4, part (i), we have

$$P(X_{n,k} = l) = \frac{1}{\sqrt{2\pi\,\mu_{n,k}}} \exp\left(-\frac{(l-\mu_{n,k})^2}{2\mu_{n,k}}\right)\left(1 + \mathcal{O}\left((1+x^2+|x|^3)\frac{p_k}{k}\right)\right) \tag{11}$$

uniformly for x with $|x| < \eta$, where x is such that $l = \mu_{n,k} + x\sqrt{\mu_{n,k}}$. Here, we used the following expansion for the variance

$$\sigma_{n,k}^2 = \mu_{n,k}\left(1 + \mathcal{O}\left(\frac{p_k}{k}\right)\right)$$

which follows from Proposition 1. Next, by the well-known local limit theorem for the Poisson distribution

$$e^{-\mu_{n,k}}\frac{(\mu_{n,k})^l}{l!} = \frac{1}{\sqrt{2\pi\,\mu_{n,k}}} \exp\left(-\frac{(l-\mu_{n,k})^2}{2\mu_{n,k}}\right)\left(1 + \mathcal{O}\left(\left(1+|x|^3\right)\frac{1}{\sqrt{\mu_{n,k}}}\right)\right)$$

again uniformly in x with $|x| < \eta$. Plugging this into Σ_1, we obtain

$$\Sigma_1 = \mathcal{O}\left(\frac{p_k}{k}\right).$$

The same bound holds for Σ_2 as well. Hence, for the first range, we obtain the estimate p_k/k which is even better than the claimed one.

As second range, we consider $(k/p_k)^{2\alpha/(3\alpha+1)} \leq \mu_{n,k} < k^2/p_k^2$ and again choose $\eta = k^\epsilon/p_k^\epsilon$. Then, the above reasoning works as well with the only difference that p_k/k in (11) has to be replaced by $1/\sqrt{\mu_{n,k}}$. So, the bound for Σ_1 becomes

$$\Sigma_1 = \mathcal{O}\left(\frac{1}{\sqrt{\mu_{n,k}}}\right) = \mathcal{O}\left(\left(\frac{p_k}{k}\right)^{\alpha/(3\alpha+1)}\right).$$

Again the same bound holds for Σ_2 as well. Hence, we are done in this range.

For the third range, we consider $1 \leq \mu_{n,k} < (k/p_k)^{2\alpha/(3\alpha+1)}$ and choose $\eta = (k/(p_k\mu_{n,k}^{3/2}))^{1/(2m+1)}$. Moreover, we use the expansion of Proposition 4, part (ii) instead of (11) above. This yields the following bound

$$\Sigma_1 = \mathcal{O}\left(\frac{p_k^2 n}{k^3}\eta\sqrt{\mu_{n,k}}\right) = \mathcal{O}\left(\left(\frac{p_k}{k}\right)^{\alpha/(3\alpha+1)}\right).$$

Again the same bound holds for Σ_2. Consequently, the claim is proved for this range as well.

For the final range where $\mu_{n,k} < 1$, we split the sum in the formula for the total variation distance slightly different

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \frac{1}{2} \sum_{|l-\mu_{n,k}|<\eta} |\cdots| + \frac{1}{2} \sum_{|l-\mu_{n,k}|\geq\eta} |\cdots| =: \Sigma_1 + \Sigma_2, \quad (12)$$

where $\eta = (k/p_k)^{1/(2m+1)}$. Then, as in the third case above, we obtain for Σ_1 the bound

$$\Sigma_1 = \mathcal{O}\left(\frac{p_k^2 n}{k^3}\eta\right) = \mathcal{O}\left(\left(\frac{p_k}{k}\right)^{\alpha} \mu_{n,k}\right).$$

As for Σ_2, we use Proposition 2 and obtain

$$\Sigma_2 = \mathcal{O}\left(\frac{\mu_{n,k}}{\eta^{2m}}\right) = \mathcal{O}\left(\left(\frac{p_k}{k}\right)^{\alpha} \mu_{n,k}\right).$$

Hence, the claimed result follows in the present case as well. This concluded the proof. □

Remark 1 The bounds in the previous theorem are still not optimal. In order to get better bounds, one needs to improve upon the second local limit theorem of Proposition 4. An improvement in the same style as in the (easier) uniform case below will lead to the following sharp bound

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \mathcal{O}\left(\frac{p_k}{k} \cdot \min\{1, \mu_{n,k}\}\right).$$

3 Uniform model

Now, we will turn to the uniform model which assigns the same probability to every phylogenetic tree of size n. Here, we will use a completely different approach based on complex-analytic tools from the analysis of algorithms. The latter area is concerned with analyzing algorithms on random inputs. One of the standard approaches to do this is to use generating functions. If the generating functions are explicit (which will be the case here), then asymptotic properties of the encoded sequences are most easiest obtained from complex-analytic properties of the functions. Many sophisticated tools have been developed along this line. Most of the tools can be considered classic by now and are found in the standard textbooks of the area; see Flajolet and Sedgewick (2009), Knuth (1997, 1998a,b) and Szpankowski (2001).

In particular, the case of fixed k is quickly derived by these standard tools (see below for more detailed references). Hence, we will mainly focus on the case where k is allowed to grow with n which will turn out to be more involved. Here, we will use an approach introduced in Baron et al. (1996) for studying the number of predecessors in random mappings which itself was based on singularity analysis, a standard method from the analysis of algorithms. However, we will make some technical improvements to obtain the optimal Berry–Esseen bound as well as sharp bounds for the total variation distance.

3.1 Moments

We will start by investigating mean value and variance. As already mentioned in the introduction, $X_{n,k}$ satisfies (1) as well. The crucial difference is the distribution of I_n which is given as

$$P(I_n = j) = \frac{C_{j-1}C_{n-1-j}}{C_{n-1}}, \qquad (1 \le j < n).$$

First, we introduce the probability generating function $Q_{n,k}(u) = \mathbb{E}(u^{X_{n,k}})$. Then, the above recurrence becomes

$$Q_{n,k}(u) = \sum_{j=1}^{n-1} \frac{C_{j-1}C_{n-1-j}}{C_{n-1}} Q_{j,k}(u) Q_{n-j,k}(u), \quad (n > k)$$

with initial conditions $Q_{n,k}(u) = 1$ for $n < k$ and $Q_{k,k}(u) = p_k(u-1) + 1$. Next, we introduce the bivariate generating function

$$G_k(u, z) = \sum_{n \ge 1} C_{n-1} Q_{n,k}(u) z^n.$$

Then, the above recurrence translates into the following quadratic equation

$$G_k(u, z) = G_k(u, z)^2 + C_{k-1} p_k (u-1) z^k + z$$

with solution

$$G_k(u, z) = \frac{1 - \sqrt{1 - 4C_{k-1}p_k(u-1)z^k - 4z}}{2}. \tag{13}$$

So, compared with the Yule–Harding case, the generating function is here explicitly computable. This will make things much more easier. All results below will be deduced with the help of (13).

First, we can quickly compute moments from the last expression by differentiation. For instance,

$$\begin{aligned}\mu_{n,k} := \mathbb{E}(X_{n,k}) &= \frac{1}{C_{n-1}}[z^n]\frac{\partial}{\partial u} G_k(u, z)\Big|_{u=1} = \frac{C_{k-1}p_k}{C_{n-1}}[z^{n-k}]\frac{1}{\sqrt{1-4z}} \\ &= \frac{(n-k+1)C_{k-1}C_{n-k}p_k}{C_{n-1}}\end{aligned}$$

for $n \ge k$. Obviously, $\mu_{n,k} = 0$ for $n < k$.

As for the variance, a similar computation reveals

$$\begin{aligned}\mathbb{E}(X_{n,k}(X_{n,k}-1)) &= \frac{1}{C_{n-1}}[z^n]\frac{\partial^2}{\partial u^2}G_k(u,z)\Big|_{u=1}\\ &= \frac{(n-2k+2)(n-2k+1)C_{k-1}^2 C_{n-2k+1}p_k^2}{C_{n-1}}\end{aligned}$$

for $n \geq 2k$. Consequently, $\sigma_{n,k}^2 := \mathrm{Var}(X_{n,k})$ equals

$$\begin{aligned}&\frac{(n-k+1)C_{k-1}C_{n-k}p_k}{C_{n-1}} + \frac{(n-2k+2)(n-2k+1)C_{k-1}^2 C_{n-2k+1}p_k^2}{C_{n-1}}\\ &-\frac{(n-k+1)^2 C_{k-1}^2 C_{n-k}^2 p_k^2}{C_{n-1}^2}\end{aligned} \tag{14}$$

for $n \geq 2k$. The corresponding formula for the range $k \leq n < 2k$ follows from the above expression for the mean value. The remaining range $n < k$ is trivial.

Overall, we have the following expression for mean and variance.

Proposition 5 *We have,*

$$\mu_{n,k} = \begin{cases} \frac{(n-k+1)C_{k-1}C_{n-k}p_k}{C_{n-1}}, & \text{if } n \geq k;\\ 0, & \text{if } n < k \end{cases}$$

and

$$\sigma_{n,k}^2 = \begin{cases} (14), & \text{if } n \geq 2k;\\ \frac{(n-k+1)C_{k-1}C_{n-k}p_k(C_{n-1}-(n-k+1)C_{k-1}C_{n-k}p_k)}{C_{n-1}^2}, & \text{if } k \leq n < 2k;\\ 0, & \text{if } n < k. \end{cases}$$

This proposition gives the following corollary.

Corollary 2 (a) *For constant k,*

$$\mu_{n,k} = \frac{C_{k-1}p_k}{4^{k-1}}n + \frac{(k-1)C_{k-1}p_k}{2\cdot 4^{k-1}} + \mathcal{O}\left(\frac{1}{n}\right), \quad (n\to\infty)$$

and, as $n \to \infty$,

$$\begin{aligned}\sigma_{n,k}^2 = &\left(\frac{C_{k-1}p_k}{4^{k-1}} - \frac{(2k-1)C_{k-1}^2 p_k^2}{4^{2k-2}}\right)n + \frac{(k-1)C_{k-1}p_k}{2\cdot 4^{k-1}}\\ &-\frac{(3k^2-4k+1)C_{k-1}^2 p_k^2}{2\cdot 4^{2k-2}} + \mathcal{O}\left(\frac{1}{n}\right).\end{aligned}$$

(b) *As* $k \to \infty$ *and* $n - k \to \infty$,

$$\mu_{n,k} \sim \sigma_{n,k}^2 \sim \frac{p_k}{\sqrt{\pi} k^{3/2}} n, \quad (n \to \infty).$$

(c) *For constant* $n - k = l \geq 0$,

$$\mu_{n,k} = \frac{(l+1) C_l p_{n-l}}{4^l} + \mathcal{O}\left(\frac{1}{n}\right), \quad (n \to \infty)$$

and

$$\sigma_{n,k}^2 = \frac{(l+1) C_l p_{n-l}}{4^l} \left(1 - \frac{(l+1) C_l p_{n-l}}{4^l}\right) + \mathcal{O}\left(\frac{1}{n}\right), \quad (n \to \infty).$$

Proof All results can easily be derived with Maple from the following well-known expansion for the Catalan numbers (see page 186 in Flajolet and Sedgewick 1995)

$$C_n = \frac{4^n}{\sqrt{\pi} n^{3/2}} \left(1 + \frac{9}{8n} + \mathcal{O}\left(\frac{1}{n^2}\right)\right), \qquad (n \to \infty).$$

We just indicate how to show part (b). Therefore, note that for $k \leq \epsilon n$ with $\epsilon < 1$, we have

$$\mu_{n,k} = \frac{C_{k-1} p_k}{4^{k-1}} n + \mathcal{O}\left(\frac{p_k}{\sqrt{k}}\right), \quad (n \to \infty) \tag{15}$$

and

$$\sigma_{n,k}^2 = \left(\frac{C_{k-1} p_k}{4^{k-1}} - \frac{(2k-1) C_{k-1}^2 p_k^2}{4^{2k-2}}\right) n + \mathcal{O}\left(\frac{p_k}{\sqrt{k}}\right), \quad (n \to \infty). \tag{16}$$

By expanding C_{k-1} as well, the claim is easily proved. So, what is left is to show that both $\mu_{n,k}$ and $\sigma_{n,k}^2$ tend to 0 as $k \geq \epsilon n$ and $n - k \to \infty$. Therefore, we use the following expansion for the mean

$$\mu_{n,k} = \frac{p_k}{\sqrt{\pi}} \left(\frac{n}{k}\right)^{3/2} \frac{1}{\sqrt{n-k}} \left(1 + \mathcal{O}\left(\frac{1}{k}\right) + \mathcal{O}\left(\frac{1}{n-k}\right)\right), \quad (n \to \infty).$$

From this the claim follows. The variance is slightly more involved, but handled similarly. □

3.2 Central limit theorem

Now, we turn to limit laws. As for the Yule–Harding Model, we start by briefly discussing the case of fixed k. Here, a result from the treatise of Flajolet and Sedgewick

(see Theorem IX.12 in Flajolet and Sedgewick 2009) immediately gives the following central limit theorem

$$\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Moreover, the above result also yields the Berry–Esseen bound which is of order $\mathcal{O}(n^{-1/2})$.

So, we can again concentrate on the case $k \to \infty$. Here, we will use a variant of the proof of the above result to show that the central limit theorem remains valid in the (maximal) range where $\mu_{n,k} \to \infty$; a similar approach was used in Baron et al. (1996).

Theorem 10 *Let $p_k n/k^{3/2} \to \infty$. Then,*

$$\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} \mathcal{N}(0, 1).$$

For the proof of the above theorem, we have to revisit the proof of Theorem IX.12 in Flajolet and Sedgewick (2009) which roughly consisted of two parts: first using a uniform version of singularity analysis (see Lemma IX.2 in Flajolet and Sedgewick 2009) and then applying the quasi-power theorem (see Theorem IX.9 in Flajolet and Sedgewick 2009). In the current situation, the largest differences will occur in the first step since also uniformity in k is needed.

We start by collecting a couple of properties of (13).

Lemma 2 *The three properties below hold for k suitable large.*

(i) *For $|u| \le 1 + \epsilon$, the polynomial*

$$F(u, z) := 1 - 4C_{k-1}p_k(u - 1)z^k - 4z = 0$$

has a unique, analytic solution $\rho_k(u)$ inside the circle $|z| \le 1/4 + c/k$. Moreover,

$$\rho_k(u) = \frac{1}{4} - \frac{C_{k-1}p_k}{4^k}(u - 1) + \frac{kC_{k-1}^2 p_k^2}{4^{2k-1}}(u - 1)^2 + \mathcal{O}\left(\frac{p_k^3}{k^{5/2}}|u - 1|^3\right) \tag{17}$$

uniformly in u with $|u| \le 1 + \epsilon$.

(ii) *We have,*

$$G_k(u, z) = \frac{1}{2} - \frac{\sqrt{k - 4(k-1)\rho_k(u)}}{2}\sqrt{1 - \frac{z}{\rho_k(u)}} + \mathcal{O}\left(p_k\sqrt{k}|1 - z/\rho_k(u)|^{3/2}\right)$$

uniformly for $|u| \le 1 + \epsilon$, $|z - \rho_k(u)| \le 1/k$, and $\arg(1 - z/\rho_k(u)) \ne \pi$.

(iii) *$G_k(u, z)$ is uniformly bounded for $|u| \le 1 + \epsilon$, $|z| \le 1/4 + c/k$, and $\arg(1 - z/\rho_k(u)) \ne \pi$.*

Proof Let $|u| \leq 1+\epsilon$. In order to prove the existence and uniqueness of a solution of $F(z,u) = 0$, we use Rouché's theorem (see page 270 in Flajolet and Sedgewick 2009 or any standard textbook on complex analysis). Therefore, we choose $f(z) = 1-4z$, $g(z) = -4C_{k-1}p_k(u-1)z^k$. Then, for a suitable constant c_1,

$$|g(z)| \leq \frac{c_1}{k^{3/2}} < \frac{4c}{k} \leq |f(z)|$$

for all z with $|z| = 1/4 + c/k$ and k sufficiently large. Hence, the existence and uniqueness of $\rho_k(u)$ is established. Moreover, since $\rho_k(u)$ is a simple root, we have

$$\frac{dF}{dz}(u, \rho_k(u)) \neq 0.$$

Consequently, the implicit function theorem implies that $\rho_k(u)$ is analytic for u with $|u| \leq 1+\epsilon$. Finally, (17) follows from $F(u, \rho_k(u)) = 0$ by implicit differentiation. This concludes the proof of part (i).

In order to prove part (ii), we expand $F(u,z)$ around $z = \rho_k(u)$. This yields

$$F(u,z) = (k-4(k-1)\rho_k(u))\left(1-\frac{z}{\rho_k(u)}\right) + \mathcal{O}\left(\sqrt{k}p_k\left(1-\frac{z}{\rho_k(u)}\right)^2\right),$$

where $|u| \leq 1+\epsilon$ and $|z| \leq 1/4 + c/k$. Plugging this into (13) together with another Taylor series expansion gives the claimed result.

Finally, part (iii) is trivial. □

From the latter result, we can deduce the following proposition.

Proposition 6 *For $c \leq k \leq Cn/(\ln n)^2$ with c and C large enough,*

$$Q_{n,k}(u) = \frac{\sqrt{k-4(k-1)\rho_k(u)}}{4^n}\rho_k(u)^{-n}\left(1+\mathcal{O}\left(\frac{p_k\sqrt{k}}{n}\right)\right)$$

uniformly in u with $|u| \leq 1+\epsilon$.

Proof This follows from the properties of the above lemma together with the traditional proof method used in singularity analysis; see Flajolet and Sedgewick (2009). We only have to be careful that the contour of integration is inside the domain, where we have a unique singularity. The latter is ensured for k with $k \leq Cn/(\ln n)^2$. □

Now, we can prove Theorem 10.

Proof of Theorem 10 First by the above proposition,

$$Q_{n,k}(e^{it/\sigma_{n,k}}) = \exp\left\{-n\ln\left(4\rho_k\left(e^{it/\sigma_{n,k}}\right)\right) + 1/2\ln\left(k-4(k-1)\rho_k\left(e^{it/\sigma_{n,k}}\right)\right)\right\} \times\left(1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right).$$

Then, by using (17) and a lengthy computation,

$$Q_{n,k}(e^{it/\sigma_{n,k}}) = \exp\left\{ita_{n,k} - \frac{t^2}{2}b_{n,k}\right\}\left(1 + \mathcal{O}\left(\frac{|t| + |t|^3}{\sigma_{n,k}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \quad (18)$$

uniformly in $|t| \le C\sigma_{n,k}$, where

$$a_{n,k} = \frac{n}{\sigma_{n,k}} \cdot \frac{C_{k-1}p_k}{4^{k-1}}, \quad b_{n,k} = \frac{n}{\sigma_{n,k}^2}\left(\frac{C_{k-1}p_k}{4^{k-1}} - \frac{(2k-1)C_{k-1}^2 p_k^2}{4^{2k-2}}\right).$$

Next, by (15) and (16),

$$a_{n,k} - \frac{\mu_{n,k}}{\sigma_{n,k}} = \mathcal{O}\left(\frac{1}{\sigma_{n,k}}\right), \quad b_{n,k} = \mathcal{O}\left(\frac{1}{\sigma_{n,k}^2}\right).$$

Consequently, the characteristic function satisfies

$$\varphi_{n,k}(t) := e^{-it\mu_{n,k}/\sigma_{n,k}} Q_{n,k}\left(e^{it/\sigma_{n,k}}\right) = e^{-t^2/2}\left(1 + \mathcal{O}\left(\frac{|t| + |t|^3}{\sigma_{n,k}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)$$

uniformly in $|t| \le C\sigma_{n,k}$. The result follows from this by Lévy's continuity theorem; see Petrov (1975). □

Finally, also the (optimal) Berry–Esseen bound can be derived.

Theorem 11 *Let $p_k n/k^{3/2} \to \infty$. Then,*

$$\sup_{-\infty<x<\infty}\left|P\left(\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \le x\right) - \Phi(x)\right| = \mathcal{O}\left(\frac{k^{3/4}}{\sqrt{p_k n}}\right).$$

Proof This follows from the expansion for the characteristic function in the above proof together with the Berry–Esseen inequality; see Petrov (1975). □

3.3 Poisson approximation

As in the Yule–Harding case, the central limit theorem just holds in the range of $\mu_{n,k} \to \infty$. We will again show that a Poisson random variable approximates $X_{n,k}$ well in a much larger range of k.

Before we can make the last statement precise, we again have to prove local limit theorems.

Proposition 7 (i) *Let $p_k n/k^{3/2} \to \infty$. Then,*

$$P(X_{n,k} = \lfloor \mu_{n,k} + x\sigma_{n,k} \rfloor) = \frac{e^{-x^2/2}}{\sqrt{2\pi\sigma_{n,k}^2}}\left(1 + \mathcal{O}\left(\left(1 + |x|^3\right)\frac{k^{3/4}}{\sqrt{p_k n}}\right)\right)$$

uniformly in $x = o((p_k n)^{1/6}/k^{1/4})$.

(ii) *Let $k \leq cn/(\ln n)^2$. Then,*

$$P(X_{n,k} = l) = e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} + \mathcal{O}\left(\frac{p_k^2 n}{k^2}\right) + \mathcal{O}\left(\frac{p_k\sqrt{k}}{n}\right)$$

uniformly in l.

Proof Part (i) follows from expansion (18) and Cauchy's formula; see for instance Hwang (2003) where a similar local limit theorem is derived.

As for part (ii), we will actually prove a more refined result than the one claimed above.

We first consider the range where $k^\epsilon/p_k^{2\epsilon} \leq \mu_{n,k} \leq \epsilon k/p_k^2$ with $\epsilon > 0$ suitable small. Then from Proposition 6, (17) and Taylor series expansion

$$Q_{n,k}(u) = \exp\left\{a_{n,k}(u-1) + b_{n,k}(u-1)^2 + \mathcal{O}\left(\frac{p_k^2\mu_{n,k}}{k}|u-1|^3\right)\right\} \times \left(1 + \mathcal{O}\left(\frac{p_k\sqrt{k}}{n}\right)\right), \tag{19}$$

where

$$a_{n,k} = (n + (k-1)/2)\frac{C_{k-1}p_k}{4^{k-1}} = \mu_{n,k} + \mathcal{O}\left(\frac{p_k\sqrt{k}}{n}\right)$$

and

$$b_{n,k} = \mathcal{O}\left(\frac{p_k\mu_{n,k}}{\sqrt{k}}\right).$$

From Cauchy's formula, we obtain

$$P(X_{n,k} = l) = \frac{1}{2\pi i}\int_{|u|=1} Q_{n,k}(u)\frac{\mathrm{d}u}{u^{l+1}} = \int_{|u-1|\leq\eta_1,|u|=1} + \int_{\eta_2\geq|u-1|>\eta_1,|u|=1} + \int_{|u-1|>\eta_2,|u|=1} =: I_1 + I_2 + I_3,$$

where $\eta_1 = (\mu_{n,k})^{-1/2+\epsilon}$ and $\eta_2 = (\mu_{n,k})^{-1/4}$. We first bound the third integral

$$I_3 \ll \exp\left\{-c\sqrt{\mu_{n,k}} + \mathcal{O}\left(\frac{p_k\mu_{n,k}}{\sqrt{k}}\right)\right\} \ll \exp\left\{-c_0\sqrt{\mu_{n,k}}\right\},$$

where c_0 is a suitable, positive constant. Next, for the second integral, observe that

$$I_2 = \frac{1}{2\pi i}\int_{\eta_2\geq|u-1|>\eta_1,|u|=1} e^{a_{n,k}(u-1)}\left(1 + \mathcal{O}\left(\frac{p_k\mu_{n,k}}{\sqrt{k}}(u-1)^2\right)\right)\frac{\mathrm{d}u}{u^{l+1}} \ll \exp\left\{-c(\mu_{n,k})^{2\epsilon}\right\}.$$

Finally, for the first integral, we use the above expansion and obtain

$$I_1 = e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \left(1 + \mathcal{O}\left(\frac{p_k \mu_{n,k}}{\sqrt{k}} \Delta^{(1)}_{n,k,l} + \frac{p_k^2 \mu_{n,k}^2}{k} \Delta^{(2)}_{n,k,l}\right)\right) + \mathcal{O}\left(\left(\frac{p_k}{\sqrt{k}}\right)^{1+\epsilon} \frac{1}{\sqrt{\mu_{n,k}}}\right), \tag{20}$$

where

$$\Delta^{(1)}_{n,k,l} = \left| \frac{l(l-1)}{\mu_{n,k}^2} - \frac{2l}{\mu_{n,k}} + 1 \right|$$

and

$$\Delta^{(2)}_{n,k,l} = \left| \frac{l(l-1)(l-2)(l-3)}{\mu_{n,k}^4} - \frac{4l(l-1)(l-2)}{\mu_{n,k}^3} + \frac{6l(l-1)}{\mu_{n,k}^2} - \frac{4l}{\mu_{n,k}} + 1 \right|.$$

Overall, we obtain the claimed result of the proposition as special case.

For the remaining range of $\mu_{n,k} < k^\epsilon / p_k^{2\epsilon}$ the above line of reasoning does not work since the estimates of I_2 and I_3 are not necessarily small. However, here we do not need to break the integral into three parts since higher order terms in the above expansion are small anyway. More precisely, from the above expansion and Cauchy's formula

$$\begin{aligned} P(X_{n,k} = l) &= \frac{1}{2\pi i} \int\limits_{|u|=1} Q_{n,k}(u) \frac{\mathrm{d}u}{u^{l+1}} \\ &= e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \left(1 + \mathcal{O}\left(\frac{p_k \mu_{n,k}}{\sqrt{k}} \Delta^{(1)}_{n,k,l}\right)\right) + \mathcal{O}\left(\frac{p_k^2 \mu_{n,k}}{k} + \frac{p_k \sqrt{k}}{n}\right), \end{aligned} \tag{21}$$

where $\Delta^{(1)}_{n,k,l}$ is as above. This concludes the proof of part (ii) of the proposition. □

From the last proposition, we can deduce our claimed result.

Theorem 12 *Let $k \to \infty$ and $n - k \to \infty$. Then,*

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \begin{cases} \mathcal{O}\left(p_k/\sqrt{k} \cdot \min\{1, \mu_{n,k}\}\right), & \text{if } \mu_{n,k} \geq (p_k/\sqrt{k})^{1-\epsilon}; \\ \mathcal{O}\left(\mu_{n,k}\right), & \text{if } \mu_{n,k} < (p_k/\sqrt{k})^{1-\epsilon}, \end{cases}$$

where $\epsilon > 0$ is an arbitrarily small constant.

Proof First, note that the proof of this result is trivial for the range where $\mu_{n,k} \to 0$ (this is the range where $p_k n/k^{3/2} \to 0$ and $n-k \to \infty$). This follows from the following estimate

$$\begin{aligned} d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) &\le \sum_{l\ge 1} \left| P(X_{n,k}=l) - e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \right| \\ &= P(X_{n,k} \ge 1) + P(\mathrm{Po}(\mu_{n,k}) \ge 1) \le 2\mu_{n,k}. \end{aligned}$$

Hence, we can focus on the other ranges. First assume that $\mu_{n,k} \ge 1$. Here, we proceed as in the proof of the corresponding result for the Yule–Harding model. Consequently, we first split the sum in the formula for the total variation distance as in (9). In order to bound the second sum, observe that

$$P(|X_{n,k} - \mu_{n,k}| \ge \eta\sqrt{\mu_{n,k}}) \le e^{-s\mu_{n,k} - s\eta\sqrt{\mu_{n,k}}} \mathbb{E}\left(e^{sX_{n,k}}\right),$$

where s will be chosen below. From (19), we obtain

$$\mathbb{E}\left(e^{sX_{n,k}}\right) = \mathcal{O}\left(e^{\mu_{n,k}(e^s-1)}\right)$$

uniformly for s with $|s| \le 1/\sqrt{\mu_{n,k}}$. Plugging this into the bound above and choosing $s = 1/\sqrt{\mu_{n,k}}$ yields

$$P(|X_{n,k} - \mu_{n,k}| \ge \eta\sqrt{\mu_{n,k}}) = \mathcal{O}\left(e^{-\eta}\right).$$

A similar bound holds when $X_{n,k}$ is replaced by $\mathrm{Po}(\mu_{n,k})$. Hence,

$$\Sigma_2 = \mathcal{O}\left(e^{-\eta}\right). \tag{22}$$

In order to bound the first sum in (9), we consider three cases. The first case, where $\mu_{n,k} \ge \epsilon k/p_k^2$ is treated as in the proof of Theorem 9.

For the second case, we assume that $k^\epsilon/p_k^{2\epsilon} \le \mu_{n,k} \le \epsilon k/p_k^2$, where ϵ is a suitable small constant. Then, we choose $\eta = k^\epsilon/p_k^{2\epsilon}$. We can use (20) in order to get the bound

$$\begin{aligned} \Sigma_1 = \mathcal{O}\left(\frac{p_k\mu_{n,k}}{\sqrt{k}} \sum_{l\ge 0} e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \Delta^{(1)}_{n,k,l} + \frac{p_k^2\mu_{n,k}^2}{k} \sum_{l\ge 0} e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \Delta^{(2)}_{n,k,l} \right) \\ + \mathcal{O}\left(\frac{p_k}{\sqrt{k}}\right). \end{aligned}$$

Next, observe that

$$\sum_{l\ge 0} e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \Delta^{(1)}_{n,k,l} = \frac{1}{\mu_{n,k}^2} \sum_{l\ge 0} e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} |(l-\mu_{n,k})^2 - l| = \mathcal{O}\left(\frac{1}{\mu_{n,k}}\right).$$

Similarly,

$$\sum_{l\geq 0} e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \Delta^{(2)}_{n,k,l} = \mathcal{O}\left(\frac{1}{\mu^2_{n,k}}\right).$$

Plugging the latter two estimates into the above bound yields

$$\Sigma_1 = \mathcal{O}\left(\frac{p_k}{\sqrt{k}}\right).$$

Due to (22) and our choice of η the same bound holds for Σ_2 as well. This proves the claim in this case.

As for the third and final case, we consider the range $1 \leq \mu_{n,k} \leq k^\epsilon / p_k^{2\epsilon}$ and again choose $\eta = k^\epsilon / p_k^{2\epsilon}$. Then, we use (21) to get the bound

$$\Sigma_1 = \mathcal{O}\left(\frac{p_k \mu_{n,k}}{\sqrt{k}} \sum_{l\geq 0} e^{-\mu_{n,k}} \frac{(\mu_{n,k})^l}{l!} \Delta^{(1)}_{n,k,l}\right) + \mathcal{O}\left(\frac{p_k^2 \eta \mu_{n,k}^{3/2}}{k} + \frac{p_k\sqrt{k}}{n}\eta\sqrt{\mu_{n,k}}\right),$$

The first term is treated as above. The second term can be further bounded as

$$\frac{p_k^2 \eta \mu_{n,k}^{3/2}}{k} + \frac{p_k\sqrt{k}}{n}\eta\sqrt{\mu_{n,k}} \ll \left(\frac{p_k}{\sqrt{k}}\right)^{2-5\epsilon} + \left(\frac{p_k}{\sqrt{k}}\right)^{2-2\epsilon} \cdot \frac{1}{\sqrt{\mu_{n,k}}} \ll \frac{p_k}{\sqrt{k}}.$$

Hence, we get the same bound for Σ_1 as above. Moreover, again due to (22) the same bound holds for Σ_2 as well. Hence, the result is for this case established as well.

Next, assume that $\mu_{n,k} \leq 1$. Here, we use (12). In order to bound Σ_2 observe that

$$P(|X_{n,k} - \mu_{n,k}| \geq \eta) \leq e^{-s\mu_{n,k} - s\eta} \mathbb{E}\left(e^{sX_{n,k}}\right).$$

From (19), we obtain

$$\mathbb{E}\left(e^{sX_{n,k}}\right) = \mathcal{O}\left(e^{\mu_{n,k}(e^s-1)}\right)$$

uniformly for s with $|s| \leq c$ where c is an arbitrary constant. Consequently,

$$P(|X_{n,k} - \mu_{n,k}| \geq \eta) = \mathcal{O}\left(e^{-c\eta}\right).$$

The same bound holds for $\mathrm{Po}(\mu_{n,k})$ as well. Hence,

$$\Sigma_1 = \mathcal{O}\left(e^{-c\eta}\right).$$

Now, we again choose $\eta = k^{\epsilon}/p_k^{2\epsilon}$. For Σ_1, we obtain

$$\Sigma_1 = \mathcal{O}\left(\frac{p_k\mu_{n,k}}{\sqrt{k}}\right) + \mathcal{O}\left(\frac{p_k^2\eta\mu_{n,k}}{k} + \frac{p_k\sqrt{k}\eta}{n}\right).$$

For the second term, we obtain

$$\frac{p_k^2\eta\mu_{n,k}}{k} + \frac{p_k\sqrt{k}\eta}{n} \ll \left(\frac{p_k}{\sqrt{k}}\right)^{2-2\epsilon}\mu_{n,k} + \left(\frac{p_k}{\sqrt{k}}\right)^{2-2\epsilon}\frac{1}{\mu_{n,k}} \ll \frac{p_k\mu_{n,k}}{\sqrt{k}}.$$

The same bound holds for Σ_2 as well. Hence, the Theorem is proved. □

4 Conclusion

In this paper, we proposed a general framework for deriving statistical properties of the occurrences of patterns in phylogenetic trees under the Yule–Harding model and the uniform model. An important feature of the current study is that our results are useful for the whole range of possible sizes of the pattern. Apart from exact and asymptotic expansions for mean value and variance, we were mainly concerned with limit laws. We demonstrated that for both models the Poisson distribution provides a good approximation for almost the whole range of the size of the pattern. When the pattern size is small, however, the normal distribution should be used. For the uniform model, we have in addition a small range with large pattern size, where the Bernoulli distribution yields a better approximation. Moreover, we also obtained sharp bounds for the error of approximation.

In recent years, phenomena of the above type have been observed for shape parameters of many discrete structures and the name "phase change" has been ascribed to them. Hence, our results show that the limit law of the number of occurrence of a given pattern in a random phylogenetic tree provides yet another example of a phase change, namely, it changes from normal to Poisson for pattern sizes that are fixed to pattern sizes that grow to infinity as the size of the tree tends to infinity. Moreover, for the uniform model, there is a second phase change to Bernoulli for pattern sizes that are close to the size of tree.

Acknowledgments The authors are indebted to the two anonymous referees for many helpful suggestions and comments. The second author acknowledges partial support by National Science Council under the grant NSC-96-2628-M-009-012.

References

Aldous DJ (1991) The continuum random tree II: an overview. In: Barlow NT, Bingham NH (eds) Stochastic analysis. Cambridge University Press, Cambridge, pp 23–70

Baron G, Drmota M, Mutafchiev L (1996) Predecessors in random mappings. Comb Prob Comput 5:317–335

Blum M, François O (2005) External branch length and minimal clade size under the neutral coalescent. Adv Appl Prob 37:647–662

Blum M, Bortolussi N, Durand E, François O (2006a) APTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics 22:363–364
Blum M, François O, Janson S (2006b) The mean, variance and limiting distributions of two statistics sensitive to phylogenetic tree balance. Ann Appl Prob 16:2195–2214
Chern H-H, Fuchs M, Hwang H-K (2007) Phase changes in random point quadtrees. ACM Trans Alg 3:51
Darwin C (1859) The origin of species, reprinted by Penguin Books. London
Drmota M, Hwang H-K (2005) Profile of random trees: correlation and width of random recursive trees and binary search trees. Adv Appl Prob 37:321–341
Drmota M, Janson S, Neininger R (2008a) A functional limit theorem for the profile of search trees. Ann Appl Prob 18:288–333
Drmota M, Gittenberger B, Panholzer A, Prodinger H, Ward MD (2008b) On the shape of the fringe of various types of random trees. Math Math Appl Sci (in press)
Feng Q, Mahmoud H, Panholzer A (2008) Phase changes in subtree varieties in random recursive trees and binary search trees. SIAM J Discrete Math 22:160–184
Fill JA, Kapur N (2004) Limiting distributions for additive functionals on Catalan trees. Theor Comp Sci 326:69–102
Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge
Flajolet P, Gourdon X, Martinez C (1997) Patterns in random binary search trees. Random Struct Alg 11:223–224
Flajolet P, Sedgewick R (1995) An introduction to the analysis of algorithms. Addison-Wesley Professional, Reading
Fuchs M (2008) Subtree sizes in recursive trees and binary search trees: Berry–Esseen bound and Poisson approximation. Comb Prob Comput 17:661–680
Fuchs M, Hwang H-K, Neininger R (2007) Profiles of random trees: limit theorems for random recursive trees and binary search trees. Algorithmica 46:367–407
Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 3:44–77
Hwang H-K (2003) Second phase changes in random m-ary search trees and generalized quicksort: convergence rates. Ann Prob 31:609–629
Hwang H-K (2007) Profiles of random trees: plane-oriented recursive trees. Random Struct Alg 30:380–413
Hwang H-K, Neininger R (2002) Phase change of limit laws in the quicksort recurrences under varying toll functions. SIAM J Comput 31:1687–1722
Hwang H-K, Nicodème P, Park G, Szpankowski W (2008) Profiles of tries. SIAM J Comput 38:1821–1880
Knuth DE (1997) The art of computer programming, vol 1, 3rd edn. In: Fundamental algorithms. Addison-Wesley, Reading
Knuth DE (1998) The art of computer programming, vol 2, 3rd edn. In: Seminumerical algorithms. Addison-Wesley, Reading
Knuth DE (1998) The art of computer programming, vol. 3, 2nd edn. In: Sorting and searching. Addison-Wesley, Reading
Loève M (1977) Probability theory. I, 4th edn. Springer, New York
McKenzie A, Steel M (2000) Distribution of cherries for two models of trees. Math Biosci 164:81–92
McKenzie A, Steel M (2001) Properties of phylogenetic trees generated by Yule-type specification models. Math Biosci 170:91–112
Mooers AO, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 72:31–54
Mooers AO, Heard SB (2002) Using tree shape. Syst Biol 51:833–834
Pemantle R (2000) Generating functions with high-order poles are nearly polynomial. In: Mathematics and computer science (Versailles, 2000). Birkhäuser, Baesl, pp 305–321
Pemantle R, Wilson MC (2002) Asymptotics of multivariate sequences I: smooth points of the singular variety. J Comb Theory Ser A 97:129–161
Pemantle R, Wilson MC (2004) Asymptotics of multivariate sequences, Part II: multiple points of the singular variety. Comb Prob Comp 13:735–761
Pemantle R, Wilson MC (2008) Twenty combinatorial examples of asymptotics derived from multivariate generating functions. SIAM Rev 20:199–272
Petrov VV (1975) Sums of indepedent random variables, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82. Springer, New York
Rosenberg NA (2006) The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Ann Comb 10:129–146

Semple C, Steel M (2003) Phylogenetics, Oxford University Press, Oxford
Stanley RP (1997) Enumerative combinatorics, vol 1. Cambridge University Press, Cambridge
Stanley RP (1999) Enumerative combinatorics, vol 2. Cambridge University Press, Cambridge
Szpankowski W (2001) Average-case analysis of algorithms on sequences. Wiley, New York
Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willies. Philos Trans R Soc London B 213:21–87