Limit Laws for the Number of Groups formed by Social Animals under the Extra Clustering Model (Extended Abstract)

Michael DRMOTA
Institute for Discrete Mathematics and Geometry
Technical University of Vienna
1040 Vienna
Austria

Michael FUCHS* and Yi-Wen LEE
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, 300
Taiwan

January 14, 2014

Abstract

We provide a complete description of the limiting behaviour of the number X_n of groups that are formed by social animals when the number n of animals tends to infinity. The analysis is based on a random model by Durand, Blum and François, where it is assumed that groups are formed more likely by animals which are genetically related. The random variable X_n can be described by a stochastic recurrence equation that is very similar to equations that occur in the stochastic analysis of divide-and-conquer algorithms although it does not fall into already known cases. In particular, we obtain (in the most interesting) "neutral model" a curious central limit theorem, where the normalizing factor is $\sqrt{\mathrm{Var}(X_n)/2}$. In the non-neutral (or extra clustering) cases the results are completely different. We obtain either a mixture of a discrete and a continuous limit law or just a discrete limit law.

1 Introduction and Results

An important problem in biology is to understand animal group patterns of social animals like wolves, lions, springboks, deers, gazelles, elephants, etc. For this purpose, many models have been proposed such as fusion/fissions models, models based on kin-selection and game-theoretic models; see Durand, Blum and François [7] for a thorough discussion and references.

*Parts of this research was done while this author visited the Institut für Diskrete Mathematik und Geometrie, Technical University of Vienna. He thanks the department for hospitality and the NSC for financial support (NSC-102-2918-I-009-012).

In [7], the authors proposed a new model which is based on the reasonable assumption that groups are formed more likely by animals which are genetically related. They used a phylogenetic tree as decision criteria for genetic relatedness. The groups then are all the maximal *clades* of the tree, where the clade of a given external node (which corresponds to an animal) in the phylogenetic tree is the set of all external nodes (animals) belonging to the tree rooted at the parent of the given external node (what we call here *clade* was called *minimal clade* in Blum and François [3] and Chang and Fuchs [5]).

We describe the model of [7] now in more details. Let n be the number of animals. We consider a random phylogenetic tree of size n which is a rooted, ordered, binary tree with n external nodes (and consequently, $n-1$ internal nodes). As random model, we consider the Yule-Harding model which can be described as follows: start with the root of the tree which has two external nodes as children; choose one external node uniformly at random and replace it by a cherry (an internal node with two external nodes); recursively repeat this procedure until a tree with n external nodes is constructed. Note that alternatively, this random model can be also described by a coalescence process starting from the n external nodes; see [5] for more details. Finally, the binary search tree model from computer science is also equivalent to this model; see Blum, François and Janson [4].

As mentioned above, the decision of genetic relatedness will be based on the random phylogenetic tree of size n (this was called the *neutral model* in [7]). Let X_n denote the number of maximal clades of the tree, or in other words, the number of groups formed by the n animals. Then, X_n can be computed recursively as follows

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{if } I_n = 1 \text{ or } I_n = n-1; \\ X_{I_n} + X^*_{n-I_n}, & \text{otherwise,} \end{cases} \qquad (n \geq 3), \tag{1}$$

where the initial condition is given by $X_2 = 1$, X^*_n is an independent copy of X_n and I_n is the size of the left subtree of the random phylogenetic tree which has a uniform distribution on $\{1, \ldots, n-1\}$.

Distributional recurrences of the above type have been investigated in many recent studies, since similar recurrences hold for shape parameters in binary search trees and performance characteristics of the quicksort algorithm; see Hwang and Neininger [14] for a very general framework. Note, however, that the framework from [14] does not apply to the current situation as will become clear from the results below.

Mean and higher moments of X_n are easy to derive by either directly considering recurrences arising from the recursive definition (1) and similar ideas as in [14] or by using generating functions and singularity analysis; see Chapter VI in Flajolet and Sedgewick [13]. Durand and François in [8] used the latter tools in order to obtain the following result for the mean: they proved that, as $n \to \infty$,

$$\mathbb{E}(X_n) \sim \frac{1-e^{-2}}{4} n. \tag{2}$$

Using an extension of their argument, one can also derive higher moments. For the variance, one obtains that, as $n \to \infty$,

$$\operatorname{Var}(X_n) \sim \frac{(1-e^{-2})^2}{4} n \log n \tag{3}$$

and for all higher central moments that

$$\mathbb{E}\left(X_n - \mathbb{E}(X_n)\right)^k \sim (-1)^k \frac{2k}{k-2} \left(\frac{1-e^{-2}}{4}\right)^k n^{k-1}, \qquad (k \geq 3). \tag{4}$$

Note that from this, the limiting distribution of X_n (centralized and normalized) cannot be identified with the method of moments.

In order to find the limiting distribution of X_n, we will use bivariate generating functions and singularity perturbation analysis. Our first main result is the following (surprising) central limit theorem.

Theorem 1. *For the number of groups X_n under the neutral model, we have*

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)/2}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ denotes the standard normal distribution and $\xrightarrow{d}$ denotes convergence in distribution.

Remark 1. We point out the curious normalization (half the variance). The same phenomenon was observed by Janson and Kersting [15] for the total length of external branches in the Kingman's coalescent.

Remark 2. A similar situation as the one above was faced by Drmota, Iksanov, Möhle and Rösler in [9], where they analyzed the number of cuts needed to isolate the root of a random recursive tree, a sequence of random variables for which the limiting distribution also cannot be found from its asymptotic moments; see Panholzer [16]. In fact, we will prove Theorem 1 with a similar approach as in [9].

The above neutral model was extended in [7] to the *extra clustering model* in order to test whether genetic relatedness is really the main driving force behind the group formation process (as seems not to be the case for social animals with many predators such as deers, gazelles and springboks; see Figure 3 in [7]). We recall the definition from [7].

Let $0 \le p \le 1$ be a fixed probability. Then, the number of groups (which with a slight abuse of notation, we will again denote by X_n) is given as follows: $X_2 = 1$ and

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{with probability } p; \\ \text{neutral model (1)}, & \text{with probability } 1-p, \end{cases} \qquad (n \ge 3). \tag{5}$$

Note that $p = 0$ is the above neutral model. For the other values of p, we will extend our above analysis. (Note that the authors of [7] again did not provide any result beyond the mean.)

Theorem 2. *(a) Let $0 < p < 1/2$. Then, we have*

$$\frac{X_n}{n^{1-2p}} \xrightarrow{d} X,$$

where the law of X is the sum of a discrete distribution of measure $p/(1-p)$ that is concentrated at zero and a continuous distribution on $[0,\infty)$ with density

$$f(x) = \frac{4(1-2p)^3}{1-p}\frac{1}{2\pi i}\int_{\mathcal{H}} t^{-2p}e^{-\delta(p)t^{1-2p}x-t}\,dt = \frac{4(1-2p)^3}{1-p}\sum_{k\ge 0}\frac{(-\delta(p))^k}{k!\Gamma(2(k+1)p-k)}x^k, \tag{6}$$

where $\mathcal{H}$ is the Hankel contour (starting in the upper half plane and winding around 0 counterclockwise) and

$$\delta(p) = \frac{(1-2p)^2 W_{p,(1-2p)/2}(-2(1-p))}{4^{p-1}(1-p)^{2p}M_{p,(1-2p)/2}(-2(1-p))},$$

where $M_{\kappa,\mu}(z)$ and $W_{\kappa,\mu}(z)$ are the Whittaker M and Whittaker W function.

We also have convergence of all moments, where the moments of X are given by $\mathbb{E}(X^k) = d_k/\Gamma(k(1-2p)+1)$ and d_k is given by

$$d_1 =: c(p) = \frac{1}{e^{2(1-p)}}\int_0^1 (1-t)^{-2p}e^{2(1-p)t}(1-(1-p)t^2)\mathrm{d}t$$

and for $k \ge 2$,

$$d_k = \frac{2(1-p)}{(k-1)(1-2p)}\sum_{j=0}^{k-2}\binom{k-1}{j}d_{k-1-j}d_{j+1}.$$

(b) Let $1/2 \le p \le 1$ Then, we have

$$X_n \xrightarrow{d} X,$$

where X is a discrete random variable with probability generating function

$$\mathbb{E}\left(u^X\right) = \frac{1-\sqrt{1-4p(1-p)u}}{2(1-p)}.$$

In particular, for $1/2 < p \le 1$, we also have convergence of moments, where the moments are given by $\mathbb{E}(X^k) = e_k$ and e_k is given by $e_1 = p/(2p-1)$ and for $k \ge 2$,

$$e_k = \frac{2(1-p)}{2p-1}\sum_{j=0}^{k-2}\binom{k-1}{j} d_{k-1-j}d_{j+1} + \frac{p}{2p-1}.$$

In the case $p = 1/2$, we only have weak convergence. More precisely, the moments of X are infinite and the moments of X_n are given by

$$\mathbb{E}(X_n^k) \sim \frac{k!J_{2k-1}}{(2k-1)!2^{2k-1}}\log^{2k-1} n,$$

where J_{2k-1} are the tangent numbers (or Euler numbers of odd index)

Remark 3. In the cases $0 < p < 1/2$ and $1/2 < p \le 1$, we have convergence of moments and the result can be proved by the method of moments (see below).

For $p = 1/2$, we use a variant of the method of proof of Theorem 1. In fact, this method can also be applied to the cases $0 < p < 1/2$ and $1/2 < p \le 1$ to give an alternative proof of our result in these cases. For the first range such an approach gives in addition that the moment generating function of X has the integral representation

$$\mathbb{E}\left(e^{yX}\right) = \frac{1}{2\pi i}\int_{\mathcal{H}} \Phi(y,t)e^{-t}\mathrm{d}t,$$

where $\mathcal{H}$ is the Hankel contour and

$$\Phi(y,t) = \frac{4(1-2p)^2 - ypm(p)4^p(1-p)^{2p-1}t^{2p-1}}{4(1-2p)^2 t - ym(p)4^p(1-p)^{2p}t^{2p}}$$

with determination of the powers in t chosen such that the branch cut is at $[0,\infty)$ and the constant $m(p)$ is given by $m(p) = M_{p,(1-2p)/2}(-2(1-p))/W_{p,(1-2p)/2}(-2(1-p))$.

It is now an easy exercise to check that this is precisely the moment generating function of the probability measure that is the sum of a discrete distribution of measure $p/(1-p)$ concentrated at zero and a continuous distribution on $[0,\infty)$ with density (6).

We conclude the introduction with a short sketch of the paper. In the next section, we consider the neutral model and derive asymptotic expansions of moments of X_n. Moreover, a strong law of large numbers is proved as well. In Section 3, we will sketch the proof of Theorem 1. In Section 4, we will prove the moment convergency part of Theorem 2, part (a). More details and the proofs of the remaining claims of Theorem 2 are postponed to the journal version of this paper.

2 Moments and Strong Law of Large Numbers - Neutral Model

In this section, we will re-prove (2) and prove (3). Moment pumping can then be used to obtain (4); for the latter method which was frequently used in the analysis of algorithm see, e.g., Chern, Fuchs and Hwang [6] and Fill and Kapur [11] and references therein.

Now, in order to prove (2) and (3), first observe that (1) yields

$$\mathbb{E}(e^{yX_n}) = \frac{2}{n-1}e^y + \frac{1}{n-1}\sum_{j=2}^{n-2}\mathbb{E}(e^{yX_j})\mathbb{E}(e^{yX_{n-j}}), \qquad (n \geq 3)$$

with initial condition $\mathbb{E}(e^{yX_2}) = e^y$. Next, set

$$X(y,z) = \sum_{n\geq 2}\mathbb{E}(e^{yX_n})z^n.$$

Then, by a straightforward computation

$$z\frac{\partial}{\partial z}X(y,z) = X(y,z) + X^2(y,z) + e^y z^2 + \frac{2e^y z^3}{1-z} \tag{7}$$

with initial condition $X(y,0) = 0$.

From (7), we obtain differential equations for the generating functions of moments of X_n by differentiation. For instance, if we set

$$E(z) = \sum_{n\geq 2}\mathbb{E}(X_n)z^n \qquad \text{and} \qquad S(z) = \sum_{n\geq 2}\mathbb{E}(X_n^2)z^n,$$

we obtain that

$$E'(z) = \left(\frac{1}{z} + \frac{2z}{1-z}\right)E(z) + \frac{z(1+z)}{1-z}, \tag{8}$$

$$S'(z) = \left(\frac{1}{z} + \frac{2z}{1-z}\right)S(z) + \frac{2}{z}E(z)^2 + \frac{z(1+z)}{1-z} \tag{9}$$

with $E(0) = S(0) = 0$. Note that both of these differential equations (as well as the corresponding differential equations for higher moments) have all the same form.

In order to solve them, we need the following lemma.

Lemma 1. *Consider*

$$f'(z) = \left(\frac{1}{z} + \frac{2z}{1-z}\right)f(z) + g(z),$$

where $f(0) = 0$. *Then,*

$$f(z) = \frac{z}{(1-z)^2e^{2z}}\int_0^z \frac{(1-t)^2e^{2t}}{t}g(t)\mathrm{d}t.$$

Proof. Straightforward. ∎

Now, for the mean, applying the above lemma to (8), we obtain that

$$E(z) = \frac{(-1+e^{2z}+2ze^{2z}-2z^2e^{2z})z}{(1-z)^2 4e^{2z}}.$$

From this, by a standard application of singularity analysis, we obtain that

$$\mathbb{E}(X_n) = \frac{1-e^{-2}}{4}n + \mathcal{O}(1).$$

Next, for the second moment, by another application of Lemma 1, we have

$$S(z) = \frac{z}{(1-z)^2 e^{2z}} \int_0^z \left(\frac{2(1-t)^2 e^{2t}}{t^2} E(t)^2 + (1-t)^2 e^{2t} \right) \mathrm{d}t.$$

Plugging into this the above expression for $E(t)$ and using Maple yields the following singularity expansion for the integrand, as $t \to 1$,

$$\frac{2(1-t)^2 e^{2t}}{t^2} E(t)^2 + (1-t)^2 e^{2t} \sim \frac{(e^2-1)^2}{8e^2} \frac{1}{(1-t)^2} + \frac{(e^2-1)^2}{4e^2} \frac{1}{1-t}.$$

Then, by the closure properties of singularity analysis from Fill, Flajolet and Kapur [10], we obtain that, as $z \to 1$,

$$S(z) \sim \frac{(1-e^{-2})^2}{8} \frac{1}{(1-z)^3} + \frac{(1-e^{-2})^2}{4} \frac{1}{(1-z)^2} \log\left(\frac{1}{1-z}\right).$$

Hence, by singularity analysis

$$\mathbb{E}(X_n^2) \sim \frac{(1-e^{-2})^2}{16} n^2 + \frac{(1-e^{-2})^2}{4} n \log n.$$

This and the above expansion for the mean yields (3).

Using the above results, we can now prove that X_n satisfies a strong law of large numbers.

Theorem 3. *We have, as $n \to \infty$,*

$$P\left(\lim_{n\to\infty} \left| \frac{X_n}{\mathbb{E}(X_n)} - 1 \right| = 0 \right) = 1.$$

Proof. First, consider $n = k^2$. Then, by Chebyshev's inequality,

$$P\left(\left| \frac{X_{k^2}}{\mathbb{E}(X_{k^2})} - 1 \right| \geq \epsilon \right) = P(|X_{k^2} - \mathbb{E}(X_{k^2})| \geq \epsilon \mathbb{E}(X_{k^2})) = \mathcal{O}\left(\frac{\log k}{k^2} \right).$$

Thus,

$$\sum_{k\geq 1} P\left(\left| \frac{X_{k^2}}{\mathbb{E}(X_{k^2})} - 1 \right| \geq \epsilon \right) < \infty.$$

Consequently, by the Lemma of Borel-Cantelli, we have, a.s.,

$$\frac{X_{k^2}}{\mathbb{E}(X_{k^2})} \to 1. \tag{10}$$

Now, for general n, find k such that

$$k^2 \leq n < (k+1)^2.$$

Then, by the fact that X_n is non-decreasing, we have

$$\frac{X_{k^2}}{\mathbb{E}(X_n)} \leq \frac{X_n}{\mathbb{E}(X_n)} \leq \frac{X_{(k+1)^2}}{\mathbb{E}(X_n)}.$$

From this, (10), and the expansion of the mean, the claim follows. ∎

Finally, for the central moments, we set

$$\bar{X}(y,z) = X(y, ze^{-ya}),$$

where $a = (1-e^{-2})/4$. Then, (7) becomes

$$z\frac{\partial}{\partial z}\bar{X}(y,z) = \bar{X}(y,z) + \bar{X}^2(y,z) + e^{y(1-2a)}z^2 + \frac{2e^{y(1-3a)}z^3}{1-ze^{-ya}}.$$

From this, by using similar arguments as before and moment pumping, one obtains (4).

3 Sketch of Proof of Theorem 1

In the previous section, we saw that the limiting distribution of X_n cannot be obtained from the method of moments. To solve this problem, we will work directly with the bivariate generating function $X(y,z)$ which satisfies the Riccati differential equation (7). Note that Flajolet, Gourdon and Martinez in [12] proposed a theory for deriving limit laws for bivariate generating functions satisfying Riccati differential equations. However, their theory does not apply to the present situation due to singularities in differential equations (see (12) below). Therefore, we have to devise another approach.

We start by solving (7). First set

$$\tilde{X}(y,z) = \frac{X(y,z)}{z}.$$

Then,

$$\frac{\partial}{\partial z}\tilde{X}(y,z) = \tilde{X}^2(y,z) + e^y\frac{1+z}{1-z}.$$

Next, set

$$\tilde{X}(y,z) = -\frac{V'(y,z)}{V(y,z)}. \tag{11}$$

Then,

$$V''(y,z) + e^y\frac{1+z}{1-z}V(y,z) = 0. \tag{12}$$

This differential equation is a variant of Whittaker's differential equation and has the following solution

$$V(y,z) = \mathrm{M}_{-e^{y/2},1/2}\left(2e^{y/2}(z-1)\right) + c(y)\mathrm{W}_{-e^{y/2},1/2}\left(2e^{y/2}(z-1)\right),$$

where $\mathrm{M}_{\kappa,\mu}(z)$ and $\mathrm{W}_{\kappa,\mu}(z)$ are the Whittaker M and Whittaker W function and

$$c(y) = -\frac{\left(e^{y/2}-1\right)\mathrm{M}_{-e^{y/2}+1,1/2}\left(-2e^{y/2}\right)}{\mathrm{W}_{-e^{y/2}+1,1/2}\left(-2e^{y/2}\right)}.$$

Whittaker functions are well-studied objects and a lot of (asymptotic) properties are known; see Chapter 6 in Beals and Wong [1]. Using these properties, we obtain the following lemma.

Lemma 2. *Let* $|y| < \eta$ *and*

$$\Delta = \{z \in \mathbb{C} \,:\, |z| < 1+\delta, \arg(1-z) \neq \pi\},$$

where $\eta, \delta > 0$. *Then,* $V(y,z)$ *is analytic in* Δ *and satisfies*

$$V(y,z) = 2(z-1) + 2ay + 4ay(z-1)\log(z-1) + \mathcal{O}(\max\{y(z-1),(z-1)^2\}), \tag{13}$$
$$V'(y,z) = 2 + 4ay\log(z-1) + \mathcal{O}(\max\{y, z-1\}), \tag{14}$$

where $a = (1-e^{-2})/4$.

Next, we need information about the zeros of $V(y,z)$ for small y.

Lemma 3. *For* η, δ *sufficiently small,* $V(y,z)$ *has only one (simple) zero* $z_0(y)$ *in* z *which satisfies, as* $y \to 0$,

$$z_0(y) = 1 - ay + 2a^2y^2\log y + \mathcal{O}(y^2).$$

Sketch of Proof. The existence follows by a standard application of Roché's theorem.

As for the proof of the asymptotic expansion, note that for y small, the zero satisfies $z_0(y) = 1 + o(1)$. Using (13) and bootstrapping, we obtain that, as $y \to 0$,

$$z_0(y) = 1 - ay + o(y).$$

Using another bootstrapping step, this can be refined to

$$z_0(y) = 1 - ay + 2a^2y^2 \log y + o(y^2 \log y).$$

Yet another bootstrapping step gives the following refined error bound

$$z_0(y) = 1 - ay + 2a^2y^2 \log y + \mathcal{O}(y^2).$$

This is the claimed result. ∎

From the last two lemmas and (11), we see that $X(y, z)$ has a logarithmic singularity at $z = 1$ and a polar singularity at $z = z_0(y)$ for y small (note that both singularities coalesce as $y \to 0$). The main property for the proof of Theorem 1 is the following proposition.

Proposition 1. *Let* $y = it/(2a\sqrt{n \log n})$. *Then,*

$$\mathbb{E}(e^{yX_n}) = z_0(y)^{-n} + \mathcal{O}\left(\frac{(\log n)^3}{n}\right).$$

Before proving it, we show how to use it to complete the proof of Theorem 1. As in the proposition, we set $y = it/(2a\sqrt{n \log n})$. Then, as $n \to \infty$,

$$z_0(y) = 1 - \frac{it}{2\sqrt{n \log n}} + \frac{t^2}{4n} + \mathcal{O}\left(\frac{\log \log n}{n \log n}\right).$$

Plugging this into the proposition above, we obtain that

$$\mathbb{E}(e^{yX_n}) = \exp\left(\frac{it\sqrt{n}}{2\sqrt{\log n}} - \frac{t^2}{4}\right)\left(1 + \mathcal{O}\left(\frac{\log \log n}{\log n}\right)\right)$$

which gives the claimed result.

Hence it remains to prove Proposition 1.

Proof. We again assume that $y = it/(2a\sqrt{n \log n})$ for some real number t. Then, by Cauchy integration, we have

$$\begin{aligned}\mathbb{E}(e^{yX_n}) &= [z^n]\, X(y, z) = -[z^{n-1}]\frac{V'(y, z)}{V(y, z)} \\ &= -\frac{1}{2\pi i}\int_\gamma \frac{V'(y, z)}{V(y, z)}\frac{dz}{z^n} \\ &= z_0(y)^{-n} - \frac{1}{2\pi i}\int_{\gamma'} \frac{V'(y, z)}{V(y, z)}\frac{dz}{z^n},\end{aligned}$$

where γ is a small positively oriented cycle of radius < 1 and $\gamma' = \gamma'_1 \cup \gamma'_2$ with

$$\gamma'_1 = \{z = 1 + v/n : v \in \mathcal{H}_n\},$$

where $\mathcal{H}_n$ denotes the *major part* of a Hankel contour

$$\mathcal{H}_n = \{v \in \mathbb{C} : |v| = 1,\ \Re(v) \le 0\} \cup \{v \in \mathbb{C} : 0 \le \Re(v) \le (\log n)^2,\ \Im(v) = \pm 1\}$$

and γ_2' completes the contour with an almost-cycle of radius $R = |1 + ((\log n)^2 + i)/n|$, so that we have to add the residue

$$\operatorname{Res}\left(\frac{V'(y,z)}{V(y,z)} z^{-n}, z = z_0(y)\right) = z_0(y)^{-n}.$$

By (13) and (14), it follows that, for $z \in \gamma_1'$,

$$\frac{V'(y,z)}{V(y,z)} = \frac{2\sqrt{n \log n}}{it} + \mathcal{O}\left((\log n)^3\right).$$

Hence, we obtain that

$$\begin{aligned}
-\frac{1}{2\pi i}\int_{\gamma_1'} \frac{V'(y,z)}{V(y,z)} \frac{dz}{z^n} &= -\frac{1}{2\pi i}\int_{\mathcal{H}_n} \left(\frac{2\sqrt{n\log n}}{it} + \mathcal{O}\left((\log n)^3\right)\right) e^{-v} \left(1 + \mathcal{O}\left(\frac{(\log n)^4}{n}\right)\right) \frac{dv}{n} \\
&= \frac{1}{\pi}\frac{\sqrt{n\log n}}{nt}\int_{\mathcal{H}_n} e^{-v}\, dv + \mathcal{O}\left(\frac{(\log n)^3}{n}\right) \\
&= \mathcal{O}\left(\sqrt{\frac{\log n}{n}} e^{-(\log n)^2}\right) + \mathcal{O}\left(\frac{(\log n)^3}{n}\right) \\
&= \mathcal{O}\left(\frac{(\log n)^3}{n}\right).
\end{aligned}$$

Finally, suppose that $|z| = R = |1 + ((\log n)^2 + i)/n|$. Here, we can use (13) and (14) and the property $z_0(y) = 1 + \mathcal{O}(1/\sqrt{n \log n})$ to deduce that

$$\frac{V'(y,z)}{V(y,z)} = \mathcal{O}\left(\frac{n}{(\log n)^2}\right).$$

Consequently,

$$-\frac{1}{2\pi i}\int_{\gamma_2'} \frac{V'(y,z)}{V(y,z)} \frac{dz}{z^n} = \mathcal{O}\left(\frac{n}{(\log n)^2} R^{-n}\right) = \mathcal{O}\left(\frac{1}{n}\right).$$

This completes the proof of Proposition 1. ∎

4 Extra Clustering Model

Here, we consider the extra clustering model and give a proof of the moment convergency part of Theorem 2, part (a). We will use the method of moments. First, observe that by (5) the bivariate generating function

$$X(y,z) = \sum_{n \ge 2} \mathbb{E}(e^{yX_n}) z^n$$

satisfies the Riccati differential equation

$$z\frac{\partial}{\partial z} X(y,z) = X(y,z) + (1-p)X(y,z)^2 + e^y \frac{z^2(1-(1-p)z^2)}{(1-z)^2} \tag{15}$$

with $X(y,0)=0$.

Set

$$E^{[k]}(z)=\frac{\partial^k}{\partial z^k}X(y,z)\Big|_{y=0}.$$

Then, by differentiation, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}z}E^{[k]}(z)=\left(\frac{1}{z}+\frac{2(1-p)z}{1-z}\right)E^{[k]}(z)+\frac{2(1-p)}{z}\sum_{j=0}^{k-2}\binom{k-1}{j}E^{[k-1-j]}(z)E^{[j+1]}(z)$$
$$+\frac{z(1-(1-p)z^2)}{(1-z)^2}$$

with $E^{[k]}(0)=0$. Again, all these differential equations have the same shape and we have the following extension of Lemma 1.

Lemma 4. *Consider*

$$f'(z)=\left(\frac{1}{z}+\frac{2(1-p)z}{1-z}\right)f(z)+g(z),$$

where $f(0)=0$. *Then,*

$$f(z)=\frac{z}{(1-z)^{2(1-p)}e^{2(1-p)z}}\int_0^z\frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t}g(t)\mathrm{d}t.$$

Proof. Straightforward. ∎

Using this and moment pumping, we obtain the following result.

Proposition 2. *Let* $p<1/2$. *Then, as* $n\to\infty$,

$$\mathbb{E}(X_n^k)\sim\frac{d_k}{\Gamma(k(1-2p)+1)}n^{k(1-2p)},$$

where d_k *is the sequence from Theorem 2, part (a).*

Proof. We use induction to show that, as $z\to 1$,

$$E^{[k]}(z)\sim\frac{d_k}{(1-z)^{k(1-2p)+1}}.$$

The claimed result follows then by singularity analysis.

First, start with $k=1$ and set $E(z)=E^{[1]}(z)$. Then, from Lemma 4,

$$E(z)=\frac{z}{(1-z)^{2(1-p)}e^{2(1-p)z}}\int_0^z(1-t)^{-2p}e^{2(1-p)t}\left(1-(1-p)t^2\right)\mathrm{d}t.$$

Observe that the integrand satisfies, as $t\to 1$,

$$(1-t)^{-2p}e^{2(1-p)t}(1-(1-p)t^2)\sim pe^{2(1-p)}(1-t)^{-2p}.$$

Thus, from the closure properties of singularity analysis, as $z\to\infty$,

$$E(z)\sim\frac{c(p)}{(1-z)^{2(1-p)}}$$

which is the claimed result for $k = 1$.

Now, for the general case, assume that claim holds for all k' with $k' < k$. In order to prove it for k, we again use Lemma 4 which now gives an integral representation for $E^{[k]}(z)$ with integrand

$$\frac{2(1-p)(1-t)^{2(1-p)}e^{2(1-p)t}}{t^2}\sum_{j=0}^{k-2}\binom{k-1}{j}E^{[k-1-j]}(z)E^{[j+1]}(z) + (1-t)^{-2p}e^{2(1-p)t}(1-(1-p)t^2)$$
$$\sim (k-1)(1-2p)d_k e^{2(1-p)}\cdot\frac{1}{(1-t)^{(k-1)(1-2p)+1}},$$

as $t \to 1$, where the last asymptotics follows from the induction hypothesis. Now, another application of the closure properties of singularity analysis yields the claimed result. ∎

Proof of moment convergence in Theorem 2, part (a). By Theorem 30.1 in Billingsley [2], it suffices to show that

$$\sum_{k\geq 1}\frac{d_k z^k}{\Gamma(k(1-2p)+1)k!}$$

has a positive radius of convergence. This clearly follows from the estimate

$$d_k \leq A^k k! k^{k(1-2p)}$$

for a suitable large A which will be proved by induction. First, by suitable choosing A, it is clear that we can assume that the estimate holds for all small k. Assume now that the claim holds for $k' < k$. In order to prove it for k, we plug the induction hypothesis into the recurrence of d_k. This gives

$$\begin{aligned} d_k &\leq A^k k!\frac{2(1-p)}{k(k-1)(1-2p)}\sum_{j=0}^{k-2}(j+1)(k-1-j)^{(k-1-j)(1-2p)}(j+1)^{(j+1)(1-2p)} \\ &\leq A^k k!\frac{2(1-p)}{k(1-2p)}\sum_{j=1}^{k-1}\left((k-j)^{k-j}j^j\right)^{1-2p} \end{aligned}$$

Now, note that $(k-j)^{k-j}j^j$ is decreasing for $0 < j \leq k/2$. Choose j_0 such that $j_0 > 1/(1-2p)$. Then,

$$d_k \leq A^k k!\frac{2(1-p)}{k(1-2p)}\left(2j_0 k^{(k-1)(1-2p)} + k^{1+(k-j_0)(1-2p)}j_0^{j_0(1-2p)}\right) \leq A^k k! k^{k(1-2p)},$$

where the last inequality holds for k large enough. This concludes the induction step. ∎

Similarly, one can prove the moment convergency part of Theorem 2, part (b) for $1/2 < p \leq 1$. As for $p = 1/2$, by using moment pumping, one obtains the following result.

Proposition 3. *Let $p = 1/2$. Then, as $n \to \infty$,*

$$\mathbb{E}(X_n^k) \sim \frac{k!J_{2k-1}}{(2k-1)!2^{2k-1}}\log^{2k-1} n,$$

where J_{2k-1} are the tangent numbers (or Euler numbers of odd index).

Thus, as for $p = 0$, the limit law again cannot be characterized via the moments. However, working directly with the Riccati differential equation (15) whose solution can again be expressed in terms of the Whittaker functions one obtains a proof for the case $p = 1/2$ (all other parts of Theorem 2 can also be proved via such a bivariate approach). Details are postponed to the journal version of this paper.

References

[1] R. Beals and R. Wong. *Special Functions: A Graduate Text*, Cambridge Studies in Advanced Mathematics, 126, Cambridge University Press, Cambridge, 2010.

[2] P. Billingsley. *Probability and Measure*, third edition, Wiley Series in Probability and Mathematical Statistics, A Wiley-Interscience Publication, John Wiley & Sons, Inc., New York, 1995.

[3] M. G. B. Blum and O. François (2005). Minimal clade size and external branch length under the neutral coalescent, *Adv. in Appl. Probab.*, **37:3**, 647-662.

[4] M. G. B. Blum, O. François and S. Janson (2006). The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance, *Ann. Appl. Probab.*, **16:4**, 2195-2214.

[5] H. Chang and M. Fuchs (2010). Limit Theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

[6] H.-H. Chern, M. Fuchs and H.-K. Hwang (2007). Phase changes in random point quadtrees, *ACM Trans. Alg.*, **3:3**, 51 pages.

[7] E. Durand, M. G. B. Blum and O. François (2007). Prediction of group patterns in social mammals based on a coalescent model, *J. Theoret. Biol.*, **249:2**, 262–270.

[8] E. Durand and O. François (2010). Probabilistic analysis of a genealogical model of animal group patterns, *J. Math. Biol.*, **60:3**, 451–468.

[9] M. Drmota, A. Iksanov, M. Möhle, U. Rösler (2009). A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree, *Random Structures Algorithms*, **34:3**, 319–336.

[10] J. A. Fill, P. Flajolet, N. Kapur (2004). Singularity analysis, Hadamard products, and tree recurrences, *J. Comput. Appl. Math.*, **174:2**, 271–313.

[11] J. A. Fill and N. Kapur (2004). Limiting distributions for additive functionals on Catalan trees, *Theor. Comput. Sci.*, **326:1-3**, 69–102.

[12] P. Flajolet, X. Gourdon, C. Martinez (1997). Patterns in random binary search trees, *Random Structures Algorithms*, **11:3**, 223–244.

[13] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2009.

[14] H.-K. Hwang and R. Neininger (2002). Phase change of limit laws in the quicksort recurrences under varying toll functions, SIAM J. Comput., **31:6**, 1687-1722.

[15] S. Janson and G. Kersting (2011). On the total external length of the Kingman coalescent, *Electronic J. Probability*, **16**, 2203–2218.

[16] A. Panholzer (2004). Destruction of recursive trees, In: *Mathematics and Computer Science III*, Birkhäuser, Basel, 267–280.