J. Math. Biol. (2016) 73:123–159
DOI 10.1007/s00285-015-0941-9

Mathematical Biology

Stochastic analysis of the extra clustering model for animal grouping

Michael Drmota[1] · **Michael Fuchs**[2] · **Yi-Wen Lee**[2]

Received: 20 November 2014 / Revised: 23 July 2015 / Published online: 31 October 2015

Abstract We consider the extra clustering model which was introduced by Durand et al. (J Theor Biol 249(2):262–270, 2007) in order to describe the grouping of social animals and to test whether genetic relatedness is the main driving force behind the group formation process. Durand and François (J Math Biol 60(3):451–468, 2010) provided a first stochastic analysis of this model by deriving (amongst other things) asymptotic expansions for the mean value of the number of groups. In this paper, we will give a much finer analysis of the number of groups. More precisely, we will derive asymptotic expansions for all higher moments and give a complete characterization of the possible limit laws. In the most interesting case (neutral model), we will prove a central limit theorem with a surprising normalization. In the remaining cases, the limit law will be either a mixture of a discrete and continuous law or a discrete law. Our results show that, except of in degenerate cases, strong concentration around the mean value takes place only for the neutral model, whereas in the remaining cases there is also mass concentration away from the mean.

Keywords Social animals · Number of groups · Moments · Limit laws · Singularity perturbation analysis

Mathematics Subject Classification 05A16 · 60F05 · 92B05

✉ Michael Fuchs
mfuchs@math.nctu.edu.tw

[1] Institute for Discrete Mathematics and Geometry, Technical University of Vienna,
1040 Vienna, Austria

[2] Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan

1 Introduction and model

A basic and important problem in biology is to gain an understanding of the dynamics of the group formation process of social animals, which are animals who spend their lives in groups, for instance, wolves, gazelles, elephants, lions, etc. In order to solve this problem, biologists have proposed many models for animal grouping, e.g., fusion/fission models, kinship models and models based on game theory; see the introduction of Durand et al. (2007) for a detailed discussion.

Some of these models use genetic relatedness as one of the driving forces behind the group formation process. Moreover, in many real-world studies, it has been observed that animals within a group are indeed often genetically related. Thus, in Durand et al. (2007), the authors proposed a simplification of previous models where only genetic relatedness is used to decide which animals belong to the same group.

The advantage of such a simplified model is that one can use the coalescent process in order to define group patterns. Moreover, the model is simple enough to devise statistical tests with which one can test whether genetic relatedness really is the major driving force behind the group formation process. For this, the authors of Durand et al. (2007) defined the *extra clustering model* which depends on a parameter $0 \leq p \leq 1$. The parameter gives the probability of additional group formation which does not correspond to genetic relatedness. Hence, for $p = 0$, no other factors than genetic relatedness are present and in this case, the authors of Durand et al. (2007) called their model the *neutral model*. From a statistical point of view, one is now interested in testing the hypothesis $p = 0$ against $p > 0$. For this purpose, the authors of Durand et al. (2007) used the maximum-likelihood test and applied it to real-world data. The outcome was a good fit of the neutral model for many classes of social animals except classes which have many predators, likely, because in this case, security is another important reason why animals huddle together.

A first probabilistic analysis of the extra clustering model was carried out by Durand and Françis (2010) who derived asymptotic expansions for the mean number of groups. However, the knowledge of only the mean value might give little information about the distribution of the number of groups. Thus, in this paper, we will look at more refined properties. Firstly, we will derive asymptotic expansions of variances and higher moments of the number of groups. These results will show that there is a strong concentration around the mean value for the neutral model ($p = 0$), whereas for the extra clustering model with $0 < p < 1$ such a concentration does not take place. Secondly, we will prove limiting distribution results for all values of p which further highlight the above mentioned concentration phenomena and also give precise insight into the behaviour of these limiting distributions. In the neutral model, a surprising central limit theorem holds which from a theoretical angle adds a further layer of richness to the model proposed by Durand et al. In the case $0 < p < 1/2$ there is not only no concentration but a mass concentration at 0 with probability $p/(1 - p)$, which means that there is positive probability that the number of groups is significantly smaller than the expected number. Finally, a phase change will be observed at $p = 1/2$, where the transition from many small groups to one big group takes place.

Before going into more details, we will give a precise definition of the extra clustering model. We start with the case $p = 0$ (neutral model). Here, the model is defined via

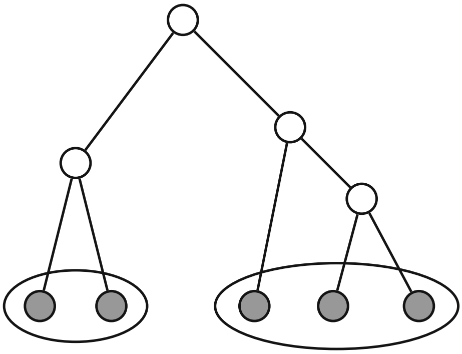

Fig. 1 The tree arising from the coalescent process applied to five animals (*grey nodes*). The number of groups in this example is two

the Kingman coalescent (Kingman 1982): start with n animals which are considered to be singleton units; at every time point pick uniformly at random two units and let them coalesce; continue this until only one unit is left. This random process can be depicted by a rooted, binary tree, where the animals are the leaves and every coalescent event corresponds to the creation of an internal node. If the leaves are drawn at the bottom and the root at the top, then the Kingman coalescent corresponds to a random process building the tree bottom-up; see Fig. 1. Alternatively, one can build a random tree top-down as follows: start with the root and two leaves; choose a leave uniformly at random and replace it by an internal node with two leaves; do this until n leaves are created. It is well-known that these two random processes yield the same random model on the set of all rooted, binary trees; see Blum et al. (2006). Moreover, this random model is also equivalent to the Yule-Harding model on phylogenetic trees; see Chang and Fuchs (2010) for details.

We recall some properties of the above random tree. First, if the two subtrees of the root have size j and $n-j$, respectively, then given the size, the two subtrees are again random trees generated by the same model. Moreover, the (random) size of the subtrees is j and $n-j$ with $1 \leq j \leq n-1$ with equal probabilities, i.e., probability $1/(n-1)$; for these properties see, e.g., Chang and Fuchs (2010).

The above random tree was used in Durand et al. (2007) to define the random number of groups. More precisely, consider n animals and construct the above random tree. This random tree describes genetic relatedness of the animals. In particular, for a given leaf of the tree, all the animals belonging to the subtree rooted at the father are genetically closely related to the leaf and this set of animals is called a *clade*; see Blum and François (2005) and Chang and Fuchs (2010). The number of groups of the n animals is now given by the number of *maximal clades*; see Fig. 1. In the sequel, we will denote this number by X_n. From the top-down construction of the random tree and the above stochastic properties, we immediately see that X_n satisfies the following distributional recurrence

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

where $X_2 = 1$, I_n has a uniform distribution on $\{1, \ldots, n-1\}$, and X_n^* denotes an independent copy of X_n. This recurrence is explained as follows: the number of groups is computed as the sum of the number of groups of the subtrees of the root unless there is only one maximal clade which is the case if and only if one of the subtrees has size one.

Recurrences of the above type have been extensively studied over the last few decades because they also arise in the analysis of certain algorithms and data structures from computer science. In particular, in Hwang and Neininger (2002), the authors proposed a very general framework to limit laws of sequences of random variable satisfying distributional recurrence similar to (1). Our above recurrence, although closely related, however, does not fall into the framework of Hwang and Neininger (2002). In particular, new phenomena not observed before for these recurrences will appear and this makes a detailed analysis of (1) highly interesting.

We next explain the extra clustering model from Durand et al. (2007). As mentioned before, this model depends on a probability p which describes the probability of extra clustering in the group formation process. More precisely, the recurrence (1) for the number of groups is replaced by the following distributional recurrence

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{with probability } p; \\ 1, & \text{with probability } 1-p \text{ and } I_n \in \{1, n-1\}; \qquad (n \geq 3), \\ X_{I_n} + X^*_{n-I_n}, & \text{with probability } 1-p \text{ and } I_n \notin \{1, n-1\}, \end{cases} \tag{2}$$

where $X_2 = 1$ and notation is as above. Note that $p = 0$ corresponds to the neutral model. For this model, the authors in Durand and Françqis (2010) computed the following asymptotic expansion of the mean

$$\mathbb{E}(X_n) \sim \begin{cases} \frac{c(p)}{\Gamma(2(1-p))} n^{1-2p}, & \text{if } 0 \leq p < 1/2; \\ \frac{\log n}{2}, & \text{if } p = 1/2; \\ \frac{p}{2p-1}, & \text{if } 1/2 \leq p \leq 1, \end{cases} \tag{3}$$

where

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

We will refine this result by proving asymptotic expansions for the variance and all higher moments and by investigating the limiting distribution of X_n for all p. Our results together with discussions and comparisons with the results from Durand and Françqis (2010) will be given in the next section.

We conclude the introduction by pointing out that a preliminary version of this paper already appeared as an extended abstract (Drmota et al. 2014). The present version contains full proofs of our results [in Drmota et al. (2014) the proof of the neutral model was only sketched and only some special cases of the extra clustering model were treated]. Moreover, we correct the expression for the density of the continuous part of the limiting distribution in the case $0 < p < 1/2$ which was stated wrongly in Drmota et al. (2014).

2 Results

In this section, we will state our results and discuss them. We start with the neutral model. Note that in this case, we have from (3) that $\mathbb{E}(X_n) = (1 - e^{-2})n/4 + \mathcal{O}(1)$.

Theorem 1 *Suppose that* $p = 0$. *Then, we have*

$$\operatorname{Var}(X_n) = \frac{(1 - e^{-2})^2}{4} n \log n + cn + \mathcal{O}(\log n) \tag{4}$$

with

$$\begin{aligned} c = {} & \frac{(4\gamma - 3)(1 - e^{-2})^2}{16} \\ & + e^{-2} \int_0^1 \left(\frac{(1 - (1 + 2t - 2t^2)e^{2t})^2}{8(1 - t)^2 e^{2t}} + (1 - t)^2 e^{2t} - \frac{e^2(1 - e^{-2})^2(3 - 2t)}{8(1 - t)^2} \right) \mathrm{d}t \\ = {} & -0.45679\cdots, \end{aligned}$$

where γ *denotes Euler's constant, and for all* $k \geq 3$,

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

Furthermore,

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

where $N(0, 1)$ *denotes the standard normal distribution, and*

$$X_n \sim \mathbb{E}(X_n) \quad a.s.$$

with the coupling arising from the top-down construction of the random tree.

Remark 1 Note that according to the above result, the standard deviation has order $\sqrt{n \log n}$ which shows strong concentration of the number of groups around the mean (which is of order n). For instance, for 100 animals ($n = 100$), we obtain a standard deviation of $6.35582\cdots$ Comparing this with the real value $6.82125\cdots$ which can be computed from (1), we see that this is a quite good approximation.

However, the convergence to the normal limit law is slow; see Fig. 2 for a plot of the limiting distribution functions and the exact distribution function for $n = 100, 200, 400, 800$ which were computed from (1) (computations of the exact

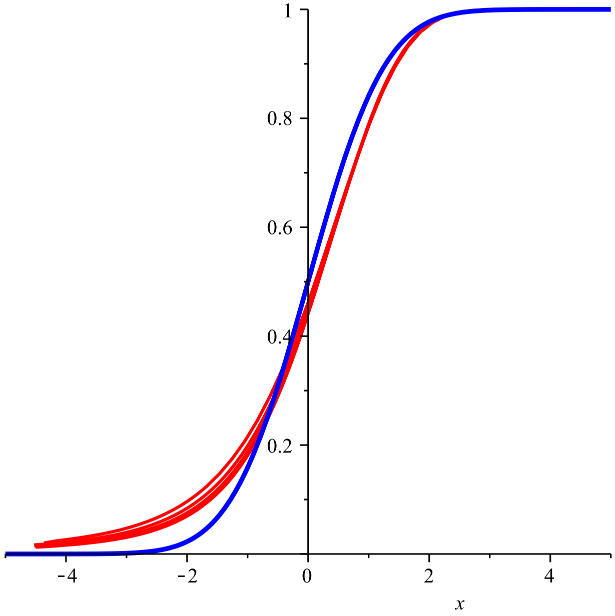

Fig. 2 The distribution functions for $n = 100, 200, 400, 800$ and limiting distribution function of the number of groups under the neutral model

distribution function for n beyond 1000 are getting rather time-consuming). Looking at the data gathered in Durand et al. (2007), sample sizes are too small to make our limit result applicable (the only larger class considered in Durand et al. (2007) are browsing springboks from the Etosha National Park in Namibia where 1064 animals have been observed).

Remark 2 Mathematically there are two surprising facts. First, there is the curious normalization by half of the variance. Note that a similar (but seemingly unrelated) phenomenon was also observed by Janson and Kersting (2011) in their analysis of the total external path length of the Kingman coalescent. A probabilistic proof of the above central limit theorem shedding further light on the curious normalization was given recently by Janson; see Janson (2014). Second, the asymptotics of centralized moments do not correspond to the moments of the normal distribution. Thus, the central limit theorem cannot be found from the method of moments; for the latter method see Section 30 in Billingsley (1995). This is in sharp contrast to Hwang and Neininger (2002), where the method of moments was applied to many examples of X_n satisfying a recurrence similar to (1).

We next turn to the extra clustering model with $p > 0$. From the data presented in Durand et al. (2007), we see that p usually does not exceed $1/2$. Thus, the case $0 < p < 1/2$ (together with $p = 0$) is of particular relevance for real-world applications and this is the range we will treat next.

Theorem 2 *Suppose that* $0 < p < 1/2$. *Then, for all* $k \geq 1$,

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

where d_k *is recursively given by* $d_1 = c(p)$ *and for* $k \geq 2$

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

Moreover,

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

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

$$f(x) = -\delta(p) \frac{1-2p}{1-p} \sum_{k \geq 0} \frac{\delta(p)^k}{k!\Gamma(2(k+1)p-k)} x^k, \tag{5}$$

where

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

and where $M_{\kappa,\mu}(z)$ *and* $W_{\kappa,\mu}(z)$ *are the Whittaker M and W functions* (*see Section* 6.7 *in* Beals and Wong 2010).

Remark 3 The most remarkable fact is that the limiting distribution has a discontinuous part at 0 (with probability $p/(1-p)$) which shows that X_n is significantly smaller than $\mathbb{E}(X_n)$ with positive probability $p/(1-p)$. It is also of interest to look at the densities of the continuous part. Figure 3 shows a plot of the density functions for several values of p. Note that the density is getting less peaked for p closer to $1/2$ which reflects the fact that the distribution becomes less and less concentrated. In particular, for $p = 1/4$, one obtains

$$f(x) = 0.3780064347 \cdots e^{-0.2525054668 \cdots x^2}.$$

For other values of p the resulting expressions are usually less explicit.

Note also that in contrast to the asymptotic expansion of the mean ($k = 1$), for the variance

$$\mathrm{Var}(X_n) \sim \left(\frac{2(1-p)}{(1-2p)\Gamma(3-4p)} - \frac{1}{\Gamma(2(1-p))^2} \right) c(p)^2 n^{2(1-2p)}, \tag{6}$$

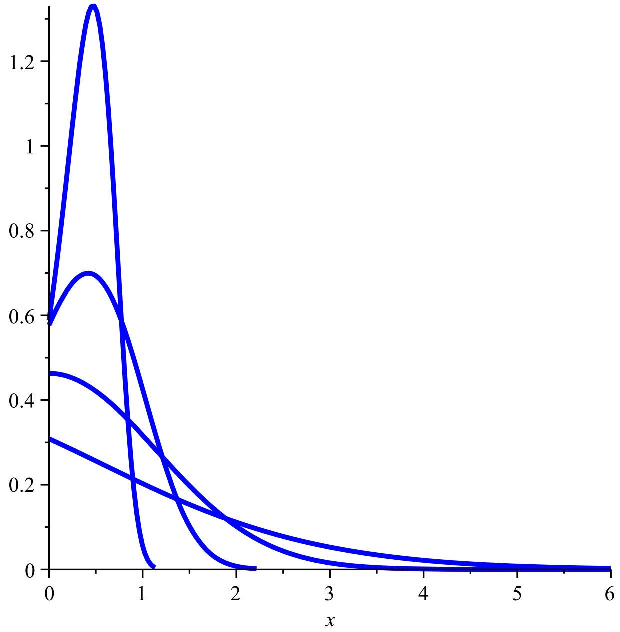

Fig. 3 The density $f(x)$ of the continuous part of the limiting distribution of the number of groups X for $p = 1/8, 3/16, 1/4, 5/16$ (*top* to *bottom*)

if we let $p \to 0$, we do not recover the result of the neutral model. Thus one expects a less accurate approximation in (6) for p close to 0; see Fig. 4 which shows a plot of the relative error of the standard deviation for $n = 100$ when using the approximation of (6) and different values of p. We see that indeed the error is large for small values of p. Moreover, the error is also large for values of p approaching $1/2$. The latter was also observed for the mean in Durand et al. (2007), however, the approximation of the mean is also very accurate for small values of p. The minimum of the relative error in Fig. 4 is attained at a value of p close to 0.09 (that is why we plotted the relative error of the standard deviation only for values of p in the vicinity of this minimum).

Remark 4 We have again plotted the limiting distribution functions and the exact distribution functions for $n = 100, 200, 400, 800$ and two values of p, namely, $p = 0.02$ and $p = 0.24$; see Fig. 5. The reason for these two choices of p comes from Durand et al. (2007). The former is the maximal-likelihood estimate of p for Alaska wolves and the second is the maximum-likelihood estimate for browsing springboks from the Etosha National Park in Namibia (they have one of the smallest values of p and the largest value of p, respectively, from the real-world data presented in Durand et al. 2007).

The remaining range of $1/2 \leq p \leq 1$ is less important from a practical point of view. Nevertheless, we will give results for this range as well for the sake of completeness. Our results show again that no strong concentration around the mean takes place (except in the trivial case $p = 1$). We start with the case $p = 1/2$.

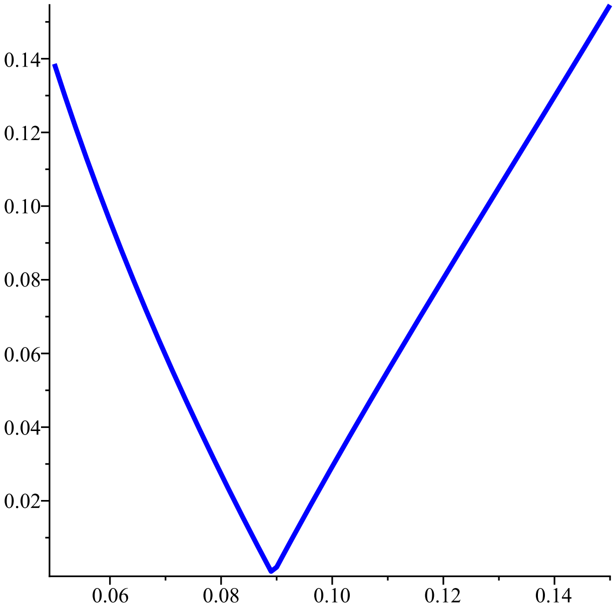

Fig. 4 Relative error of standard deviation of the number of groups for $n = 100$ as a function of p

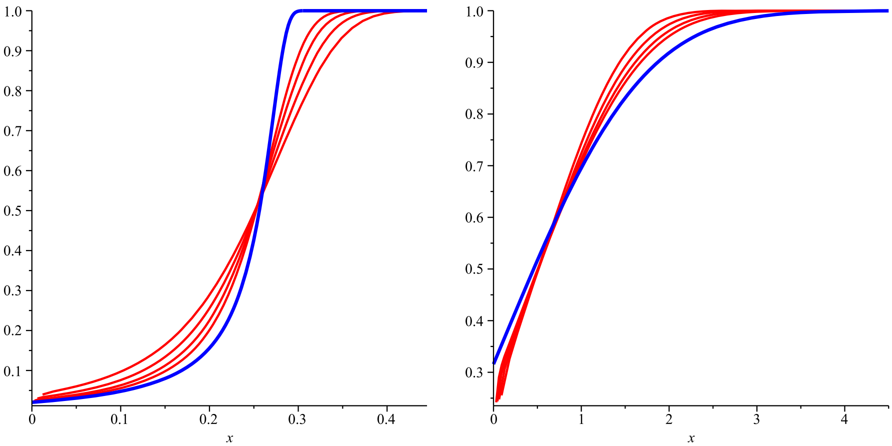

Fig. 5 The distribution functions for $n = 100, 200, 400, 800$ of $p = 0.02$ (*left* Alaska wolves) and $p = 0.24$ (*right* browsing springboks) and their limiting distribution functions

Theorem 3 *Suppose that* $p = 1/2$. *Then, we have*

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

where J_{2k-1} *are the Euler numbers of odd index* (*see, e.g., page* 144 *in* Flajolet and Sedgewick 2009). *Furthermore,*

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

where X *has a discrete law on* $\{1, 2, \ldots\}$ *which is given by*

$$P(X = k) = \frac{2^{-2k}}{2k-1}\binom{2k}{k}.$$

Note that in this case the moments of X_n do not converge.

Finally, we turn to the case $1/2 < p \leq 1$.

Theorem 4 *Suppose that* $1/2 < p \leq 1$. *Then*, *for all* $k \geq 1$,

$$\mathbb{E}(X_n^k) \sim e_k,$$

where e_k *is recursively given by* $e_1 = p/(2p-1)$ *and for* $k \geq 2$

$$e_k = \frac{1-p}{2p-1}\sum_{j=1}^{k-1}\binom{k}{j}e_j e_{k-j} + \frac{p}{2p-1}.$$

Moreover,

$$X_n \xrightarrow{d} X$$

with convergence of all moments, *where* X *has a discrete law on* $\{1, 2, \ldots\}$ *which is given by*

$$P(X = k) = \frac{p^k(1-p)^{k-1}}{2(2k-1)}\binom{2k}{k}.$$

Remark 5 Note that Theorems 3 and 4 can be merged. However, there is an significant difference between these results: in Theorem 3 we only have convergence in distribution, whereas in Theorem 4 also all moments do converge.

Overall, our above results combined give a full picture of the limiting behavior of the number of groups under the extra clustering model. In particular, we see that the limit law is continuous for $p = 0$, is a mixture of a discrete and continuous distribution for $0 < p < 1/2$, and finally becomes discrete as $1/2 \leq p \leq 1$ (and degenerates at $p = 1$). The most important range for real-world applications is $0 \leq p < 1/2$. Here, a notable phenomenon is the strong concentration around the mean for the neutral model which, however, does not hold for the extra clustering model with $p > 0$. This shows a quantitative difference between the neutral model and the extra clustering model with $p > 0$, something which was not visible from the previous analysis of the mean value (Durand and Françis 2010). Moreover, from a mathematical point of view, an interesting aspect is that the limit law can be obtained via the method of moments if and only if $0 < p < 1/2$ and $1/2 < p \leq 1$, but not in the two cases $p = 0$ and $p = 1/2$.

We conclude the introduction with a short sketch of the paper. Since the derivation of the moments of X_n is quite technical but standard, we have put this analysis to Appendix 1 (as main tool we will use singularity analysis which will briefly reviewed at the beginning of this appendix). In Sect. 3, we will introduce the mathematical tools needed for the proofs of our limit laws. More precisely, this section will contain a short discussion of Whittaker functions and some of their properties which are needed in the proofs. Moreover, we will explain our approach to limit laws via singularity perturbation analysis. The proofs of the limiting distribution results will then be contained in Sect. 4 for $p = 0$ and Appendix 2 for $0 < p \leq 1$. We will end the paper with a conclusion.

3 Whittaker functions and singularity perturbation analysis

In this section, we will explain our analytic method used for proving our limiting distribution results of the number of groups under the extra clustering model (Theorems 1–4). The method will rely on the explicit solution of (7) which will involve Whittaker functions. Thus, properties of Whittaker functions will play a crucial role and we will recall them below. The method itself then uses singularity perturbation analysis and will also be explained in details below.

We consider the moment-generating function of X_n which by (2) satisfies the recurrence, for $n \geq 3$,

$$\mathbb{E}\big(e^{yX_n}\big) = pe^y + (1-p)\frac{2}{n-1}e^y + \frac{1-p}{n-1}\sum_{j=2}^{n-2}\mathbb{E}\big(e^{yX_j}\big)\mathbb{E}\big(e^{yX_{n-j}}\big)$$

with initial condition $\mathbb{E}\left(e^{yX_2}\right) = e^y$ (the above sum is equal to 0 for $n = 3$). Next, set

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

Then, by a straightforward computation

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

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

Solution of (7). Note that (7) is a Riccati differential equation for which a standard solution procedure exists. Therefore, set

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

Then, (7) becomes

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

with $\tilde{X}(y, 0) = 0$. Next, set

$$\tilde{X}(y, z) = -\frac{1}{1-p} \cdot \frac{V'(y, z)}{V(y, z)},$$

where $V(y, 0) = 1$ and differentiation is with respect to z. Then, we obtain the second-order differential equation

$$V''(y, z) + (1-p)e^{y}\frac{1-(1-p)z^2}{(1-z)^2}V(y, z) = 0$$

with $V(y, 0) = 1$ and $V'(y, 0) = 0$. This differential equation is a variant of Whittaker's differential equation. Thus, its solution can be expressed in terms of the Whittaker functions as follows

$$\begin{aligned} V(y, z) = {} & M_{-(1-p)e^{y/2}, \sqrt{1-4p(1-p)e^{y}}/2}(2(1-p)e^{y/2}(z-1)) \\ & + c(y) W_{-(1-p)e^{y/2}, \sqrt{1-4p(1-p)e^{y}}/2}(2(1-p)e^{y/2}(z-1)) \end{aligned}$$

with

$$c(y) = \frac{(1 + \sqrt{1-4p(1-p)e^{y}} - 2(1-p)e^{y/2}) M_{-(1-p)e^{y/2}+1, \sqrt{1-4p(1-p)e^{y}}/2}(-2(1-p)e^{y/2})}{2 W_{-(1-p)e^{y/2}+1, \sqrt{1-4p(1-p)e^{y}}/2}(-2(1-p)e^{y/2})}.$$

We will work in the next section with this explicit solution. Consequently, we will need some background knowledge on Whittaker functions which we will recall next.

Whittaker functions. Here, we gather some properties of the Whittaker functions. The exposition will follow Section 6 in Beals and Wong (2010).

We start with the definition of the Whittaker functions which are independent solutions of Whittaker's differential equation

$$v''(z) + \left(-\frac{1}{4} + \frac{\kappa}{z} + \frac{1-4\mu^2}{4z^2}\right) v(z) = 0.$$

They can be expressed as follows

$$\begin{aligned} M_{\kappa,\mu}(z) &= e^{-z/2} z^{\mu+1/2} M\left(\mu - \kappa + \frac{1}{2}, 1 + 2\mu, z\right), \\ W_{\kappa,\mu}(z) &= e^{-z/2} z^{\mu+1/2} U\left(\mu - \kappa + \frac{1}{2}, 1 + 2\mu, z\right). \end{aligned}$$

Here, $M(a, c; z)$ and $U(a, c; z)$ are the Kummer functions. The former is defined for all $a, c, z \in \mathbb{C}$ with $c \neq 0, -1, -2, \ldots$ by the following series

$$M(a, c; z) = \sum_{\ell=0}^{\infty} \frac{(a)_\ell}{(c)_\ell \ell!} z^\ell,$$

where $(a)_\ell$ is the Pochhammer symbol

$$(a)_\ell := a(a+1)\cdots(a+\ell-1).$$

Note that the above expression shows that $M(a,c;z)$ is analytic in all three variables. The definition of the Kummer function of second kind, namely $U(a,c;z)$, is slightly more involved. More precisely, $U(a,c;z)$ is defined as

$$U(a,c;z) = \frac{\Gamma(1-c)}{\Gamma(a+1-c)} M(a,c;z) + \frac{\Gamma(c-1)}{\Gamma(a)} z^{1-c} M(a+1-c, 2-c; z)$$

for all a, c, z with $c \notin \mathbb{Z}$. The definition can be extended to $c = m \in \mathbb{N}$ (where the limit exists) as follows

$$\begin{aligned} U(a,m;z) = \frac{(-1)^m}{\Gamma(a+1-m)(m-1)!} \Bigg(& M(a,m;z)\log z \\ & + \sum_{\ell=0}^{\infty} \frac{(a)_\ell}{(m)_\ell \ell!} \left(\psi(a+\ell) - \psi(\ell+1) - \psi(m+\ell)\right) z^\ell \Bigg) \\ & + \frac{(m-2)!}{\Gamma(a)} z^{1-m} \sum_{\ell=0}^{m-2} \frac{(a+1-m)_\ell}{(2-m)_\ell \ell!} z^\ell \end{aligned}$$

with $\psi(z) = \Gamma'(z)/\Gamma(z)$. We will in the sequel choose the determination of log and powers such that we have a branch cut at $[0,\infty)$. Then, from the above definitions we obtain the following lemma.

Lemma 1 *Assume that $\mu \neq -1/2, -1, -3/2, \ldots$ Then, both Whittaker functions are analytic on $\mathbb{C} \backslash [0,\infty)$.*

Finally, note that the above expressions also give singularity expansions as $z \to 0$. For instance, if $\mu \neq -1/2, -1, -3/2, \ldots$, then

$$M_{\kappa,\mu}(z) \sim z^{\mu+1/2}, \quad (z \to 0)$$

and if in addition $\mu \neq 0, 1/2, 1, \ldots$, then

$$W_{\kappa,\mu}(z) \sim \begin{cases} \frac{\Gamma(2\mu)}{\Gamma(\mu-\kappa+1/2)} z^{-\mu+1/2}, & \text{if } \mu > 0; \\ \frac{\Gamma(-2\mu)}{\Gamma(-\mu-\kappa+1/2)} z^{\mu+1/2}, & \text{if } \mu < 0. \end{cases}$$

Similar expansions can be found for $W_{\kappa,\mu}(z)$ when $\mu = 0, 1/2, 1, \ldots$ as well.

Singularity perturbation analysis. We will now explain our method of proof of our limit laws from Sect. 1. The method is based on singularity perturbation analysis, a term coined by Flajolet and Lafforgue (1994). The idea is to directly work with the moment-generating function of X_n which by Cauchy's integral formula is obtained from $X(y,z)$ by

$$\mathbb{E}\big(e^{yX_n}\big) = \frac{1}{2\pi i}\int_\gamma \frac{X(y,z)}{z^{n+1}}\,\mathrm{d}z. \tag{8}$$

Here, y is considered to be a parameter for which we assume that $|y| < \eta$ with $\eta > 0$ suitably small.

In order to use (8), one has to choose a suitable contour γ and to study the singularity structure of $X(y,z)$. Due to the above explicit expression for $X(y,z)$, we see that the singularities are either the branch point singularities (with moving branch-cut) of the Whittaker functions or are poles arising from the zeros of $V(y,z)$. By doing a change of variable in (8) (replacing $e^{y/2}(z-1)$ by $z-1$), we can consider

$$\begin{aligned}\tilde{V}(y,z) = V(z, 1+e^{-y/2}(z-1)) &= M_{-(1-p)e^{y/2},\sqrt{1-4p(1-p)e^y}/2}(2(1-p)(z-1))\\ &\quad + c(y)W_{-(1-p)e^{y/2},\sqrt{1-4p(1-p)e^y}/2}(2(1-p)(z-1)).\end{aligned} \tag{9}$$

Now, the branch cut is fixed at $[1,\infty)$. As for the zeros of this function, we will prove that there are two cases:

- Case I: $p = 0$. Here, we will show that for $|z| < 1+\delta$ with a suitable δ, we have exactly one zero $z_0(y)$ of $\tilde{V}(y,z)$. Moreover, this zero has the property that it converges to the branch point singularity as y tends to 0.
- Case II: $p > 0$. Here, we will show that for $|z| < 1+\delta$ with a suitable δ, we have no zeros of $\tilde{V}(y,z)$.

The second case is more in line with other instances to which singularity perturbation analysis was applied; see Flajolet and Lafforgue (1994) and Chapter X of Flajolet and Sedgewick (2009). In this case, γ will be deformed into a Hankel-type contour; see the right contour in Fig. 6. The asymptotic evaluation of (8) is then immediate and the main term comes from the part of the contour close to the branch point singularity.

The first case is more involved, in particular, due to the fact that the polar singularity (arising from the zero of $\tilde{V}(y,z)$) coalesces with the branch point singularity as y tends to 0. Note that a somewhat similar situation was encountered in a recent study of Drmota et al. (2009). In fact, our approach in case I will resemble the one of Drmota et al. (2009). More precisely, we will again deform the contour into the same type of contour as in case II; see the left contour in Fig. 6. This will lead to a contribution coming from the polar singularity by a straightforward application of the residue theorem. Then, in contrast to case II, we will show that the contribution of the branch point singularity is negligible. From this, the unusual central limit theorem of the neutral model will follow.

Remark 6 Analytically, the unusual normalization in Theorem 1 arises from the two coalescing singularities. If, for instance, one would only have a polar singularity, then a central limit theorem with the usual normalization would hold; see Flajolet et al. (1997).

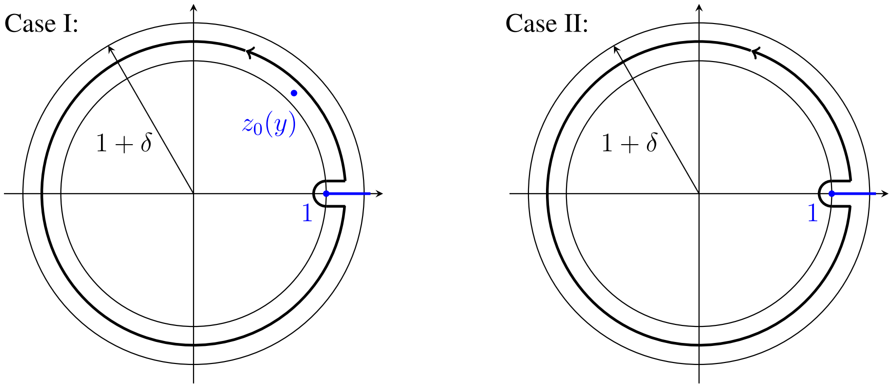

Fig. 6 The integration contour and singularities in the two cases. The only (but crucial) difference is the additional polar singularity in case I

4 Limit laws

In this section, we will prove the limiting distribution result for the neutral model (Theorem 1). The corresponding proofs of the limiting distribution results for the extra clustering model with $p > 0$ (Theorems 2–4) will be given in Appendix 2. For the proof, we will use singularity perturbation analysis and the properties of the Whittaker functions from the previous section.

We first collect some properties of $\tilde{V}(y,z)$ and $\tilde{V}'(y,z) = \frac{\partial}{\partial z}\tilde{V}(y,z)$ (for the definition see (9)).

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

$$\tilde{\Delta} = \{z \in \mathbb{C} \;:\; |z| < 1+\delta,\ \arg(z-1) \neq 0\},$$

where $\eta, \delta > 0$. Then, $\tilde{V}(y,z)$ and $\tilde{V}'(y,z)$ are analytic in $\tilde{\Delta}$ and satisfy

$$\begin{aligned}\tilde{V}(y,z) &= 2(z-1) + 2ay + 4ay(z-1)\log(z-1)\\ &\quad + \mathcal{O}(\max\{|y|^2, |y||z-1|, |z-1|^2\}),\end{aligned} \tag{10}$$

$$\tilde{V}'(y,z) = 2 + 4ay\log(z-1) + \mathcal{O}\left(\max\{|y|, |z-1|\}\right), \tag{11}$$

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

Proof This follows from the properties of the Whittaker function from Sect. 3. □

Next, we need the following lemma which was already announced in the previous section.

Lemma 3 *For η, δ sufficiently small, $\tilde{V}(y,z)$ as a function of z has only one (simple) zero $z_0(y)$ in $\tilde{\Delta}$. Moreover, we have, as $y \to 0$,*

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

Proof First note that $\tilde{V}(0, z) = 2(z-1)e^{z-1}$ which is an entire function with only one (simple) zero at $z = 1$. Next, $\tilde{V}(y, z)$ is analytic in both y and z in $\tilde{\Delta}$ and thus its zeros vary continuously with y. Also, note that because of (10) of the above lemma, there is no zero in a sufficiently small neighborhood of $z = 1$ for y sufficiently small (the limits as z tends to the branch-cut in the neighborhood are never equal to zero as well). Thus, for η, δ sufficiently small, we exactly have one zero in $\tilde{\Delta}$ which in addition must move to 1 as y tends to 0. This proves the first claim.

As for the proof of the second claim, we use bootstrapping. We already know that

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

Plugging this into (10), 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 second claim. □

Now, we state the key lemma for the proof of the central limit theorem.

Lemma 4 *Let* $y = it/(2a\sqrt{n \log n})$. *Then,*

$$\mathbb{E}\left(e^{yX_n}\right) = z_0(y)^{-n} + \mathcal{O}\left(\frac{1}{\log n}\right). \tag{12}$$

Proof For the proof, we use Cauchy's integral formula

$$\begin{aligned}\mathbb{E}\left(e^{yX_n}\right) &= [z^n]X(y, z) = -[z^{n-1}]\frac{V'(y, z)}{V(y, z)} \\ &= -\frac{1}{2\pi i}\int_{\tilde{\gamma}} \frac{V'(y, z)}{V(y, z)}\frac{\mathrm{d}z}{z^n} \\ &= -\frac{1}{2\pi i}\int_{\gamma} \frac{\tilde{V}'(y, w)}{\tilde{V}(y, w)}\frac{\mathrm{d}\omega}{(e^{-y/2}(w-1)+1)^n},\end{aligned}$$

where $\tilde{\gamma}$ is a small positively oriented circle centered at the origin and the last step follows from the change of variables $e^{y/2}(z-1) = w-1$. We now deform the contour γ into a contour γ' which is given by $\gamma' = \gamma_1' \cup \gamma_2'$ with

$$\gamma_1' = \{w = 1 + v/n \ : \ v \in \mathcal{H}_n\},$$

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

$$\mathcal{H}_n = \{v \in \mathbb{C} \,:\, |v| = 1, \Re(v) \le 0\} \\ \cup \left\{v \in \mathbb{C} \,:\, 0 \le \Re(v) \le \sqrt{(1+\delta')^2 n^2 - 1} - n, \Im(v) = \pm 1\right\}$$

(here, as usual $\Re$ and $\Im$ denote the real part and imaginary part of a complex number) and γ_2' completes the contour with an almost circle of radius $1 + \delta'$ with $0 < \delta' < \delta$; see Fig. 6 where γ_1' is the small noose around the branch cut and the almost circle γ_2' is the remainder of the contour. Note that the above integral then becomes

$$\mathbb{E}\big(e^{yX_n}\big) = (e^{-y/2}(z_0(y)-1)+1)^{-n} - \frac{1}{2\pi i}\int_{\gamma'} \frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)} \frac{\mathrm{d}\omega}{(e^{-y/2}(w-1)+1)^n}$$

since by the residue theorem, we have to add the residue

$$\operatorname{Res}\left(\frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)}(e^{-y/2}(w-1)+1)^{-n}, w = z_0(y)\right) = (e^{-y/2}(z_0(y)-1)+1)^{-n}.$$

In order to derive (12) from this, we first note that from $z_0(y) = 1+\mathcal{O}(1/\sqrt{n\log n})$, we obtain that

$$(e^{-y/2}(z_0(y)-1)+1)^{-n} = z_0(y)^{-n}\left(1+\mathcal{O}\left(\frac{1}{n\log n}\right)\right)^{-n} = z_0(y)^{-n} + \mathcal{O}\left(\frac{1}{\log n}\right).$$

Next for the integral, note that by (10) and (11) and again $z_0(y) = 1+\mathcal{O}(1/\sqrt{n\log n})$, we have for $w \in \gamma_1'$,

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

Moreover, for $w \in \gamma_1'$

$$|e^{-y/2}(w-1)+1|^{-n} \le \left(1+\frac{\Re(e^{-y/2}v)}{n}\right)^{-n} \le e^{-\Re(e^{-y/2}v)} = \mathcal{O}\big(e^{-\epsilon\Re(v)}\big) \quad (13)$$

for a suitable $\epsilon > 0$. Hence, we obtain that for $w \in \gamma_1'$,

$$\frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)} \cdot \frac{1}{(e^{-y/2}(w-1)+1)^n} = \mathcal{O}\left(\frac{n}{\log^2 n} e^{-\epsilon\Re(v)}\right).$$

Consequently,

$$-\frac{1}{2\pi i}\int_{\gamma_1'}\frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)}\frac{\mathrm{d}\omega}{(e^{-y/2}(w-1)+1)^n}$$
$$=-\frac{1}{2\pi i}\int_{\mathcal{H}_n}\mathcal{O}\left(\frac{n}{\log^2 n}e^{-\epsilon\Re(v)}\right)\frac{\mathrm{d}v}{n}=\mathcal{O}\left(\frac{1}{\log^2 n}\right).$$

Finally, suppose that $|w|=1+\delta'$. First, from the analyticity of $\tilde{V}(y,z)$ and $\tilde{V}'(y,z)$, we obtain that

$$\frac{\tilde{V}'(y,z)}{\tilde{V}(y,z)}=\mathcal{O}(1).$$

Moreover,

$$|e^{-y/2}(w-1)+1|\geq|w|+\mathcal{O}\left(\frac{1}{\sqrt{n\log n}}\right)\geq 1+\delta''$$

for n large enough with $0<\delta''<\delta'$. Thus,

$$-\frac{1}{2\pi i}\int_{\gamma_2'}\frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)}\frac{\mathrm{d}\omega}{(e^{-y/2}(w-1)+1)^n}=\mathcal{O}((1+\delta'')^{-n}).$$

Putting everything gives

$$\mathbb{E}\big(e^{yX_n}\big)=z_0(y)^{-n}+\mathcal{O}\left(\frac{1}{\log n}\right)+\mathcal{O}\left(\frac{1}{\log^2 n}\right)+\mathcal{O}((1+\delta'')^{-n})$$
$$=z_0(y)^{-n}+\mathcal{O}\left(\frac{1}{\log n}\right)$$

which is the claimed result. □

The proof of the central limit theorem (from Theorem 1) follows now from the last lemma.

Proof of the central limit theorem from Theorem 1 As in Lemma 4 set $y=it/(2a\sqrt{n\log n})$. Then, by the expansion of $z_0(y)$ from Lemma 3, we obtain

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

Inserting this into the result from Lemma 4 yields

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

and by rearranging

$$\mathbb{E}\left(e^{it(X_n-an)/(2a\sqrt{n\log n})}\right)=\exp\left(-\frac{t^2}{4}\right)\left(1+\frac{\log\log n}{\log n}\right).$$

Since $\exp(-t^2/4)$ is the characteristic function of a normal distribution with mean 0 and variance $1/2$, the claimed central limit theorem follows from this by Lévy's continuity theorem. □

The proofs of the remaining limiting distribution results can be found in Appendix 2. Note that in the cases $0 < p < 1/2$ and $1/2 < p \leq 1$ our results can be also derived from the moment asymptotics (given in Appendix 1) together with the property that the limiting distribution is characterized by its moments.

5 Conclusion

In this paper, we gave a detailed analysis of the extra clustering model which was recently introduced by Durand et al. (2007) because of two reasons: (i) to model the group formation process of social animals and (ii) to test whether genetic relatedness is the main driving force behind the group formation process. Our analysis extends the previous analysis of Durand and Françis (2010) which was concerned with asymptotic expansions of the mean of the number of groups formed by the animals. We derived all higher moments and completely classified the limiting distribution of the number of groups for all values of p. Our results are most relevant for the range $0 \leq p < 1/2$ which were the p values observed in real-word data. They show that the distribution of the number of groups is strongly concentrated around the mean for the neutral model, but not for the extra clustering model with $p > 0$. Thus, there is a phase change in the behaviour from $p = 0$ to $p > 0$, something which was not visible from previous results for the mean.

As for limiting distributions, our results show that the limit law is a continuous law for the neutral model ($p = 0$), a mixture of discrete and continuous law for the extra clustering model with $0 < p < 1/2$ and a discrete law for $1/2 \leq p \leq 1$. This transition from continuous to discrete is in fact expected since the extra clustering model is getting less random as p increases (because animals are more likely to form one huge group).

From a mathematical point of view, our results contain two surprises. First, not in all cases, the limit law can be obtained by the method of moments. In fact, we have seen two cases, namely, $p = 0$ and $p = 1/2$, where we have weak convergence but moments do not converge. The case of the neutral model is in particular surprising because the underlying sequence of random variables satisfies a divide-and-conquer recurrence of a type which often appears in computer science and for which in many previous studies an application of the method of moments led to the limit law. The second surprise is the curious limit law for the neutral model. In fact, our proof does not give a lot of insight of as to why this surprising result holds. A better explanation was given in a recent paper of Janson (2014). However, many things about this result are still shrouded in mystery, in particular, whether such a surprising result also holds for other classes of random trees such as random m-ary search trees (in this work, we considered trees which are equivalent to random binary search trees; for a definition of this family of trees as well as random m-ary search trees see Mahmoud 1992). We hope to come back to this question in a future work.

Acknowledgments An extended abstract (Drmota et al. 2014) and talk with preliminary materials of this paper were presented at the 25th International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms which took place in Paris from June 16 to June 20, 2014. The authors thank the reviewers of this abstract and the participants of the meeting for valuable input. Moreover, the authors also thank the referees of the present paper for helpful comments leading to an improvement of the presentation. M. Drmota acknowledges partial support by the Austrian Science Foundation, SFB F50 “Algorithmic and Enumerative Combinatorics”. M. Fuchs was partially supported by the Ministry of Science and Technology, Taiwan under the Grant MOST-103-2115-M-009-007-MY2.

Appendix 1: Moments

In this appendix we will investigate the moments of X_n. The method we use for this has already been used in many other studies and was nicknamed “moment pumping”; see, e.g., Chern et al. (2007) or Fill and Kapur (2004) and references therein. It is based on induction and singularity analysis. The latter is a standard tool of analytic combinatorics, see Chapter VI of Flajolet and Sedgewick (2009), and says—in a nutshell—that the leading asymptotic behavior of the coefficients a_n of a power series $f(z) = \sum_{n\geq 0} a_n z^n$ is mainly governed by the kind of the dominating singularity z_0 of $f(z)$ on the radius of convergence $|z| = |z_0| = R$. For example, if $z_0 = 1$ and we have

$$f(z) = A(1-z)^{\alpha} + \mathcal{O}((1-z)^{\beta})$$

for $z \to 1$, $z \in \Delta$, where α, β are real numbers and Δ is a so-called Δ-domain of the form

$$\Delta = \{z \in \mathbb{C} : |z| < 1+\delta,\ |\arg(z-1)| > \varphi\}, \quad (\delta > 0, 0 < \varphi < \pi/2),$$

then we have

$$a_n = A\frac{n^{-\alpha-1}}{\Gamma(-\alpha)} + \mathcal{O}\big(n^{\max\{-\beta-1,-\alpha-2\}}\big).$$

Actually, we can also work without an error term. For example, $f(z) \sim A(1-z)^{\alpha}$ as $z \to 1$, $z \in \Delta$, implies $a_n \sim An^{-\alpha-1}/\Gamma(-\alpha)$. This tool will be extensively used below.

We next explain in more details the above mentioned method of moment-pumping. Due to singularity analysis, it suffices to find singularity expansions for the generating functions of the moments of X_n. By differentiating (7) with respect to y and setting $y = 0$, we see that these generating functions satisfy differential equations. In particular, the resulting differential equations are all of the following general form

$$f'(z) = \left(\frac{1}{z} + \frac{2(1-p)z}{1-z}\right) f(z) + g(z), \tag{14}$$

where $g(z)$ is a function of generating functions of moments of smaller order. Thus, we have a recursive scheme with which generating functions of moments can be computed inductively once a general solution of the above differential equation is known. Such a solution is provided in the next lemma.

Lemma 5 *Let $f(z)$ and $g(z)$ be functions which are analytic at zero and satisfy*

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

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 This is proved by applying the standard approach for solving first-order differential equations. □

We will use this general solution and induction to obtain the singularity expansion (in a Δ-domain) of generating functions of moments of all order. Moreover, in the same way, generating functions of moments are also proved to be analytic in a suitable domain. Both these properties will follow from closure properties of singularity analysis; see Fill et al. (2004) or Section VI.10 in Flajolet and Sedgewick (2009).

We demonstrate first how this works for the neutral model and then apply a similar approach to the extra clustering model with $p > 0$.

Moments for $p = 0$. We start with mean and variance. Differentiating (7) with respect to y once and twice and setting $y = 0$ gives

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

and

$$S'(z) = \left(\frac{1}{z} + \frac{2z}{1-z}\right) S(z) + \frac{2}{z}M(z)^2 + \frac{z(1+z)}{1-z},$$

with the notation

$$M(z) = \sum_{n\geq 2} \mathbb{E}(X_n)z^n, \quad S(z) = \sum_{n\geq 2} \mathbb{E}(X_n^2)z^n.$$

Now, for the mean, an application of Lemma 5 gives

$$M(z) = \frac{(-1 + e^{2z} + 2ze^{2z} - 2z^2e^{2z})z}{(1-z)^2 4e^{2z}}. \tag{15}$$

Thus, as $z \to 1$, $z \in \Delta$,

$$M(z) = \frac{1-e^{-2}}{4} \cdot \frac{1}{(1-z)^2} + \mathcal{O}\left(\frac{1}{1-z}\right).$$

Consequently, by applying singularity analysis,

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

Next, for the second moment, again by Lemma 5

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

Using (15) together with a proper use of a computer algebra system (we used Maple), we obtain that for the integrand, as $t \to 1$, $t \in \Delta$,

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

This leads to,

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

as $z \to 1$, $z \in \Delta$. Hence, again by singularity analysis,

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

From this and the above expansion of the mean, we obtain that

$$\mathrm{Var}(X_n) \sim \frac{(1-e^{-2})^2}{4} n \log n.$$

Remark 7 More terms in the asymptotic expansions of the mean and variance can be obtained in a straightforward manner by computing more terms in the singularity expansion of the functions above and again applying singularity analysis. Such a refined computation in particular gives the claimed expansion for the variance from Theorem 1.

From the last two results, we also obtain a strong law of large numbers for X_n (with the coupling arising from the top-down construction of the random tree as explained in Sect. 1). This is the last statement of Theorem 1.

Lemma 6 *Suppose that* $p = 0$. *Then, we have*

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

In other words,

$$X_n \sim \mathbb{E}(X_n) \quad a.s.$$

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

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

for all $\epsilon > 0$, where in the last step, we used the above results for the mean and variance of X_n. A standard application of the lemma of Borel–Cantelli now gives

$$\lim_{k\to\infty}\frac{X_{k^2}}{\mathbb{E}(X_{k^2})} = 1 \quad \text{a.s.} \tag{16}$$

Next, for general n, find k such that

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

Note that by the above asymptotics for the mean, we have that

$$\mathbb{E}(X_{(k+1)^2}) \sim \mathbb{E}(X_{k^2}) \quad (k \to \infty). \tag{17}$$

Moreover, the fact that X_n is non-decreasing (from the coupling) gives

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

From this, the claimed results follows by using (16) and (17). □

This result suggests looking at central moments. Hence, we set

$$\bar{X}(y,z) := X(y, ze^{-ya}) = \sum_{n\geq 2}\mathbb{E}\left(e^{y(X_n - an)}\right)z^n$$

with $a := (1 - e^{-2})/4$. Then, (7) becomes (recall that $p = 0$)

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

Now, taking the k-th derivative with respect to y and setting $y = 0$ yields for

$$\bar{M}^{[k]}(z) := \frac{\partial^k}{\partial y^k} \bar{X}(y, z)\bigg|_{y=0}$$

the differential equation

$$\bar{M}^{[k]'}(z) = \left(\frac{1}{z} + \frac{2z}{1-z}\right) \bar{M}^{[k]}(z) + \frac{1}{z}\sum_{j=1}^{k-1}\binom{k}{j}\bar{M}^{[j]}(z)\bar{M}^{[k-j]}(z) + \bar{h}^{[k]}(z),$$

where $\bar{M}^{[k]}(0) = 0$ and

$$\bar{h}^{[k]}(z) = (1-2a)^k z + \frac{\mathrm{d}^k}{\mathrm{d}y^k}\frac{2e^{y(1-3a)}z^2}{1-ze^{-ya}}\bigg|_{y=0}.$$

This differential equation is of the type (14). Thus, we can apply Lemma 5 and induction to obtain the following lemma.

Lemma 7 *For* $k \geq 3$, *as* $z \to 1$, $z \in \Delta$,

$$\bar{M}^{[k]}(z) \sim \frac{2(-1)^k k! a^k}{(k-2)(1-z)^k}.$$

Proof First note that from the computations above for the mean and the variance, we have the following bounds, as $z \to 1$, $z \in \Delta$,

$$\bar{M}^{[1]}(z) = \mathcal{O}\left(\frac{1}{1-z}\right) \quad \text{and} \quad \bar{M}^{[2]}(z) = \mathcal{O}\left(\frac{1}{(1-z)^2}\log\frac{1}{1-z}\right). \tag{18}$$

We prove our claim by induction. Since the proofs for the base step and the induction step are the same, we merge them into one. So, assume that the claim holds for all $k' < k$. In order to show that it holds for k, we use Lemma 5 which yields

$$\bar{M}^{[k]}(z) = \frac{z}{(1-z)^2e^{2z}}\int_0^z \frac{(1-t)^2e^{2t}}{t}\left(\frac{1}{t}\sum_{j=1}^{k-1}\binom{k}{j}\bar{M}^{[j]}(t)\bar{M}^{[k-j]}(t) + \bar{h}^{[k]}(t)\right)\mathrm{d}t. \tag{19}$$

We first consider the two terms inside the bracket. For the first, by (18) and induction hypothesis, we obtain that, as $t \to 1$, $t \in \Delta$,

$$\frac{1}{t}\sum_{j=1}^{k-1}\binom{k}{j}\bar{M}^{[j]}(t)\bar{M}^{[k-j]}(t) = \mathcal{O}\left(\frac{1}{(1-t)^{k+\epsilon}}\right),$$

where $\epsilon > 0$ is an arbitrarily small constant (this constant comes from the additional log term of $\bar{M}^{[2]}(z)$). For the second term, it is not complicated to see that, as $t \to 1$, $t \in \Delta$,

$$\bar{h}^{[k]}(t) \sim \frac{2(-1)^k a^k}{(1-t)^{k+1}}.$$

Thus, for the integrand of (19), as $t \to 1$, $t \in \Delta$,

$$\frac{(1-t)^2 e^{2t}}{t}\left(\frac{1}{t}\sum_{j=1}^{k-1}\binom{k}{j}\bar{M}^{[j]}(t)\bar{M}^{[k-j]}(t)+\bar{h}^{[k]}(t)\right) \sim \frac{2e^2(-1)^k a^k}{(1-t)^{k-1}}.$$

Hence, by the closure properties of singularity analysis, as $z \to 1$, $z \in \Delta$,

$$\int_0^z \frac{(1-t)^2 e^{2t}}{t}\left(\frac{1}{t}\sum_{j=1}^{k-1}\binom{k}{j}\bar{M}^{[j]}(t)\bar{M}^{[k-j]}(t)+\bar{h}^{[k]}(t)\right)\mathrm{d}t \sim \frac{2e^2(-1)^k a^k}{(k-2)(1-z)^{k-2}}.$$

Inserting this into (19) gives the claimed result. □

The proposed expansion for all central moments of order higher than two (as stated in Theorem 1) now follows from Lemma 7 and singularity analysis. In particular, note that by

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k = \mathbb{E}(X_n - an)^k + \mathcal{O}(|\mathbb{E}(X_n - an)^{k-1}|),$$

we can easily transfer asymptotic results for $\mathbb{E}(X_n - an)^k$ to corresponding ones for $\mathbb{E}(X_n - \mathbb{E}(X_n))^k$.

Next we prove the results for the moments of the number of groups under the extra clustering model with $p > 0$. We will follow the same proof strategy as in the above proof for the case $p = 0$ with the only difference that we now directly work with moments instead of central moments.

Moments for $0 < p < 1/2$. Set

$$M^{[k]}(z) := \frac{\partial^k}{\partial y^k} X(y,z)\bigg|_{y=0} = \sum_{n\geq 2}\mathbb{E}X_n^k z^n.$$

Then, (7) implies that

$$M^{[k]'}(z) = \left(\frac{1}{z}+\frac{2(1-p)z}{1-z}\right)M^{[k]}(z) + \frac{1-p}{z}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(z)M^{[k-j]}(z) + h(z).$$

where $M^{[k]}(0) = 0$ and

$$h(z) = \frac{z(1-(1-p)z^2)}{(1-z)^2}.$$

By Lemma 5, the solution of this differential equation is given by

$$\begin{aligned} M^{[k]}(z) = {} & \frac{z}{(1-z)^{2(1-p)}e^{2(1-p)z}} \int_0^z \frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t} \\ & \times \left(\frac{1-p}{t} \sum_{j=1}^{k-1} \binom{k}{j} M^{[j]}(t) M^{[k-j]}(t) + h(t) \right) \mathrm{d}t. \end{aligned} \qquad (20)$$

Now, the asymptotic of the moments for the case $0 < p < 1/2$ (as stated in Theorem 2) follows from the next lemma by singularity analysis.

Lemma 8 *For $k \geq 1$, as $z \to 1$, $z \in \Delta$,*

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

Proof We start with $k = 1$. Here, according to (20), we have

$$M^{[1]}(z) = \frac{z}{(1-z)^{2(1-p)}e^{2(1-p)z}} \int_0^z \frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t} h(t)\,\mathrm{d}t.$$

Note that the integrand has the singularity expansion, as $t \to 1$, $t \in \Delta$,

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

Since $2p < 1$, applying the closure properties of singularity analysis yields, as $z \to 1$, $z \in \Delta$,

$$\int_0^z \frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t} h(t)\,\mathrm{d}t \sim e^{2(1-p)} d_1.$$

Inserting this into the above expression for $M^{[1]}(z)$ gives the claimed asymptotics for $k = 1$.

Now, assume that the claim is true for all $k' < k$. We want to show that it also holds for k. First, observe that by the induction hypothesis, the integrand of (20) satisfies, as $t \to 1$, $t \in \Delta$,

$$\frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t}\left(\frac{1-p}{t}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(t)M^{[k-j]}(t)+h(t)\right)$$
$$\sim(1-p)e^{2(1-p)}\left(\sum_{j=1}^{k-1}\binom{k}{j}d_jd_{k-j}\right)\cdot\frac{1}{(1-t)^{(k-1)(1-2p)-1}}.$$

Thus, by the closure properties of singularity analysis, as $z\to 1$, $z\in\Delta$,

$$\int_0^z\frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t}\left(\frac{1-p}{t}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(t)M^{[k-j]}(t)+h(t)\right)\mathrm{d}t$$
$$\sim\frac{e^{2(1-p)}d_k}{(1-z)^{(k-1)(1-2p)}}.$$

Inserting this into (20) gives, as $z\to 1$, $z\in\Delta$,

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

which is the claimed result. □

Moments for $p=1/2$. Again the proof is similar as in the previous case. More precisely, we show the following lemma.

Lemma 9 *For* $k\geq 1$, *as* $z\to 1$, $z\in\Delta$,

$$M^{[k]}(z)\sim\frac{b_k}{1-z}\log^{2k-1}\frac{1}{1-z},$$

where b_k *is recursively given by* $b_1=1/2$ *and for* $k\geq 2$

$$b_k=\frac{1}{2(2k-1)}\sum_{j=1}^{k-1}\binom{k}{j}b_jb_{k-j}. \tag{21}$$

Proof The proof is again by induction. We start with $k=1$. In this case, the integrand of (20) satisfies, as $t\to 1$, $t\in\Delta$,

$$\frac{(1-t)e^t}{t}h(t)\sim\frac{e}{2(1-t)}.$$

Thus, as $z\to 1$, $z\in\Delta$,

$$\int_0^z\frac{(1-t)e^t}{t}h(t)\,\mathrm{d}t\sim\frac{e}{2}\cdot\log\frac{1}{1-z}.$$

Inserting this into (20) gives the claimed result.

For the induction step, assume that the claim holds for all $k' < k$. Then, for the proof that it also holds for k, by the induction hypothesis, the integrand of (20) satisfies, as $t \to 1, t \in \Delta$,

$$\frac{(1-t)e^t}{t}\left(\frac{1}{2t}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(t)M^{[k-j]}(t)+h(t)\right)$$
$$\sim \frac{e}{2}\left(\sum_{j=1}^{k-1}\binom{k}{j}b_j b_{k-j}\right)\cdot\frac{1}{1-t}\log^{2k-2}\frac{1}{1-t}.$$

Hence, by the closure properties of singularity analysis, we obtain that, as $z \to 1$, $z \in \Delta$,

$$\int_0^z \frac{(1-t)e^t}{t}\left(\frac{1}{2t}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(t)M^{[k-j]}(t)+h(t)\right)\mathrm{d}t \sim eb_k\log^{2k-1}\frac{1}{1-z}.$$

Inserting this into (20) concludes the induction step. □

The asymptotic expansion of the moments in the case $p = 1/2$ follows from this by singularity analysis and the following lemma which solves the recurrence for b_k in terms of Euler numbers.

Lemma 10 *The solution of the recurrence* (21) *from Lemma* 9 *is given by*

$$b_k = \frac{k!J_{2k-1}}{(2k-1)!2^{2k-1}},$$

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

Proof We use generating functions. Set

$$B(z) = \sum_{k\ge 1} b_k \frac{z^k}{k!}.$$

Then, the recurrence (21) becomes

$$B(z)B'(z) = 2zB''(z) + B'(z) - 1/2$$

with $B(0) = 0$. Integrating yields

$$4zB'(z) = B(z)^2 + 2B(z) + z$$

which has the solution

$$B(z) = \sqrt{z}\tan\left(\frac{\sqrt{z}}{2}\right).$$

Expanding and using that $\tan(z)$ is the generating function of the Euler numbers gives the claim. □

Moments for $1/2 < p \le 1$. This final case is again treated similar to the two previous cases. More precisely, the result follows from the following lemma and singularity analysis.

Lemma 11 *For* $k \ge 1$, *as* $z \to 1$, $z \in \Delta$,

$$M^{[k]}(z) \sim \frac{e_k}{1-z}.$$

Proof Again, we use induction and (20). First, for $k = 1$, the integrand of (20) satisfies, as $t \to 1$, $t \in \Delta$,

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

Then, from the closure properties of singularity analysis, as $z \to 1$, $z \in \Delta$,

$$\int_0^z \frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t}h(t)\,\mathrm{d}t \sim \frac{pe^{2(1-p)}}{(2p-1)(1-z)^{2p-1}}.$$

Inserting this into (20) gives the claimed result.

Next, assume by induction that the claim holds for all $k' < k$. In order to show it for k note that the integrand of (20) satisfies, as $t \to 1$, $t \in \Delta$,

$$\frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t}\left(\frac{1-p}{t}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(t)M^{[k-j]}(t) + h(t)\right)$$
$$\sim e^{2(1-p)}\left((1-p)\sum_{j=1}^{k-1}\binom{k}{j}e_j e_{k-j} + p\right)\cdot\frac{1}{(1-t)^{2p}}.$$

Thus, by the closure properties of singularity analysis, we obtain that, as $z \to 1$, $z \in \Delta$,

$$\int_0^z \frac{(1-t)^{2(1-p)}e^{2(1-p)t}}{t}\left(\frac{1-p}{t}\sum_{j=1}^{k-1}\binom{k}{j}M^{[j]}(t)M^{[k-j]}(t) + h(t)\right) \sim \frac{e^{2(1-p)}e_k}{(1-z)^{2p-1}}.$$

Inserting this into (20) gives the desired result. □

Appendix 2: Limit laws for $0 < p \le 1$

Here we present the proofs of the limiting distribution results for the number of groups under the extra clustering model with $p > 0$ (Theorems 2–4). We will use a similar approach as for the proof of the central limit theorem of Theorem 1 in Sect. 4. Moreover, for $0 < p < 1/2$ and $1/2 < p \leq 1$ the limiting distributions can be derived, too, by the method of moments applied to the moment asymptotics given in Appendix 1.

Limiting distribution for $0 < p < 1/2$. As in the proof of the central limit theorem (in the case $p = 0$) we start by giving expansions for $\tilde{V}(z, y)$ and $\tilde{V}'(z, y)$.

Lemma 12 *Let* $|y| < \eta$ *and*

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

where $\eta, \delta > 0$. *Then,* $\tilde{V}(y, z)$ *and* $\tilde{V}'(y, z)$ *are analytic in* $\tilde{\Delta}$ *and satisfy*

$$\begin{aligned}\tilde{V}(y, z) &= 2^{1-p}(1-p)^{1-p}(z-1)^{1-p} - \frac{2^{p-1}(1-p)^{p+1}}{(1-2p)^2} m(p) y (z-1)^p \\ &\quad + \mathcal{O}\big(\max\{|y|^2|z-1|^p, |y||z-1|^{1-p}, |z-1|^{2-p}\}\big), \qquad (22)\\ \tilde{V}'(y, z) &= 2^{1-p}(1-p)^{2-p}(z-1)^{-p} - \frac{p2^{p-1}(1-p)^{p+1}}{(1-2p)^2} m(p) y (z-1)^{p-1} \\ &\quad + \mathcal{O}\big(\max\{|y|^2|z-1|^{p-1}, |y||z-1|^{-p}, |z-1|^{1-p}\}\big). \qquad (23)\end{aligned}$$

Proof This follows from the properties of Whittaker functions from Sect. 3. □

Next, we need to study zeros of $\tilde{V}(y, z)$. In contrast to Lemma 3, in the current case, we have no zeros.

Lemma 13 *For* η, δ *sufficiently small,* $\tilde{V}(y, z)$ *as a function of* z *has no zeros in* $\tilde{\Delta}$.

Proof First note that

$$\tilde{V}(0, z) = 2^{1-p}(1-p)^{1-p}(z-1)^{1-p} e^{-(1-p)(z-1)}$$

which has no zero in $\tilde{\Delta}$ and only tends to 0 on the branch-cut when z tends to 1. The latter property holds for $\tilde{V}(y, z)$ for η and δ small enough as well as can be easily seen from (22). Thus, due to the analyticity of $\tilde{V}(y, z)$, all its zeros have to escape to infinity as y tends to zero. Consequently, for η sufficiently small, $\tilde{V}(y, z)$ has no zero in $\tilde{\Delta}$. □

The main lemma in this context is the following one.

Lemma 14 *Let* $y = it/n^{1-2p}$. *Then,*

$$\mathbb{E}\left(e^{yX_n}\right) = \frac{1}{2\pi i} \int_{\mathcal{H}} \Phi(y, v) e^{-v} \mathrm{d}v + \mathcal{O}\left(\frac{\log^2 n}{n^{1-2p}}\right),$$

where $\mathcal{H}$ *is the Hankel contour starting in the upper half plane at* $+\infty$ *and winding around* 0 *counterclockwise before tending to* $+\infty$ *in the lower half plane and*

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

with determination of the powers in t chosen such that the branch cut is at $[0,\infty)$ *and*

$$m(p)=\frac{M_{p,(1-2p)/2}(-2(1-p))}{W_{p,(1-2p)/2}(-2(1-p))}.$$

Proof The proof is similar to the proof of Lemma 4 with the crucial difference that now the main contribution will come from the branch-point singularity (since there is no polar singularity). The starting point is again Cauchy's integral formula which as in Lemma 4 can be rewritten to

$$\mathbb{E}\big(e^{yX_n}\big)=-\frac{1}{2(1-p)\pi i}\int_{\gamma}\frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)}\frac{\mathrm{d}w}{(e^{-y/2}(w-1)+1)^n}.$$

We again deform the contour γ into a contour γ' which this time is $\gamma'=\gamma_1'\cup\gamma_2'\cup\gamma_3'$, where

$$\gamma_i'=\{w=1+v/n\ :\ v\in\mathcal{H}_n^{(i)}\},\quad i=1,2$$

with $\mathcal{H}_n^{(i)}$, $i=1,2$, given by

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

and γ_3' completes the contour with an almost circle of radius $1+\delta'$ with $\delta'<\delta$. The difference to the contour from Lemma 4 is that $\mathcal{H}_n$ there is split into the two parts $\mathcal{H}_n^{(1)}$ and $\mathcal{H}_n^{(2)}$. Moreover, note that now, deforming the contour in the integral above will leave the value of the integral unchanged.

We proceed by treating the three integrals corresponding to the above three parts of the contour. First, for γ_1' note that from (22) and (23), we obtain that

$$\frac{\tilde{V}'(y,w)}{\tilde{V}(y,w)}=n(1-p)\Phi(it,v)\left(1+\mathcal{O}\left(\frac{\log^{2-2p}n}{n^{1-2p}}\right)\right).$$

Moreover, we have that

$$(e^{-y/2}(w-1)+1)^{-n}=e^{-v}\left(1+\mathcal{O}\left(\frac{\log^2 n}{n^{1-2p}}\right)\right).$$

Thus,

$$\begin{aligned}&-\frac{1}{2(1-p)\pi i}\int_{\gamma_1'}\frac{\tilde{V}'(y,z)}{\tilde{V}(y,z)}\frac{\mathrm{d}w}{(e^{-y/2}(w-1)+1)^n}\\&\quad=-\frac{1}{2\pi i}\int_{\mathcal{H}_n^{(1)}}n\Phi(it,v)e^{-v}\left(1+\mathcal{O}\left(\frac{\log^{2-2p}n}{n^{1-2p}}\right)\right)e^{-v}\left(1+\mathcal{O}\left(\frac{\log^2 n}{n^{1-2p}}\right)\right)\frac{\mathrm{d}v}{n}\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2\pi i}\int_{\mathcal{H}_n^{(1)}} \Phi(it, v)e^{-v}\,\mathrm{d}v + \mathcal{O}\left(\frac{\log^2 n}{n^{1-2p}}\right)\\
&= \frac{1}{2\pi i}\int_{\mathcal{H}} \Phi(it, v)e^{-v}\,\mathrm{d}v + \mathcal{O}\left(\frac{\log^2 n}{n^{1-2p}}\right),
\end{aligned}
$$

where the last step follows by attaching the tails of the Hankel contour (which introduces a negligible error) and changing the orientation of the contour.

Next, we consider the contribution of the integral over γ_2'. Here, we have

$$\frac{\tilde{V}'(y, w)}{\tilde{V}(y, w)} = \mathcal{O}(1).$$

Moreover, by using (13) which holds in the current situation as well, we obtain that

$$
\begin{aligned}
&-\frac{1}{2(1-p)\pi i}\int_{\gamma_2'} \frac{\tilde{V}'(y, z)}{\tilde{V}(y, z)}\frac{\mathrm{d}w}{(e^{-y/2}(w-1)+1)^n}\\
&= -\frac{1}{2\pi i}\int_{\mathcal{H}_n^{(2)}} \mathcal{O}\big(e^{-\epsilon\Re(v)}\big)\frac{\mathrm{d}v}{n} = \mathcal{O}\left(\frac{1}{n}\right).
\end{aligned}
$$

Finally, the integral over γ_3' is exactly treated as in Lemma 4 and it contributes only an exponential decreasing error term. Collecting these three parts yields the claimed result. □

Now, we can complete the proof of the limiting distribution of Theorem 2.

Proof of the limiting distribution result for $0 < p < 1/2$. Weak convergence follows from the previous lemma.

In order to show that also moments converge, we work with the moments (stated in Theorem 2 and proved in Appendix 1) and use the method of moments. Accordingly, the only thing which one has to verify is that there is unique random variable whose moment sequence is given by $d_k/\Gamma(k(1-2p)+1)$. For this purpose, it suffices to show that

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

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

$$d_k \le A^k k! k^{k(1-2p)} \tag{24}$$

for a sufficiently large A. We will prove this by induction, where in the proof we will show how large one has to choose A.

First, by choosing A suitably, it is clear that we can assume that the estimate holds for all small k. Now, assume that it holds for all $k' < k$. In order to prove it for k, we insert the induction hypothesis into the recurrence for d_k (see Theorem 2). This gives

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

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

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

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

Finally, since we know already that X_n/n^{1-2p} weakly converges to X, we must have that $\mathbb{E}(X^k) = d_k/\Gamma(k(1-2p)+1)$.

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

This concludes the proof of the limiting distribution.

We complete the proof by showing that the limiting distribution has the shape stated in Theorem 2. It is sufficient to show that

$$
\int_0^{\infty} f(x) e^{yx}\, \mathrm{d}x = \frac{1}{2\pi i} \int_{\mathcal{H}} \left(\Phi(y, v) - \frac{p}{v(1-p)} \right) e^{-v}\, \mathrm{d}v
$$

for every fixed $y \in \mathbb{C}$. For this purpose, we use the series representation (5) and Hankel's representation of the reciprocal of the Gamma function

$$
\frac{1}{\Gamma(2(k+1)p-k)} = -\frac{1}{2\pi i} \int_{\mathcal{H}} (-v)^{-2(k+1)p+k} e^{-v}\, \mathrm{d}v.
$$

Next, we replace the Hankel contour $\mathcal{H}$ by the contour $\mathcal{H}'$ that starts in the upper half plane at $+e^{i\varphi}\infty$, winds around 0 counterclockwise before it tends to $+e^{-i\varphi}\infty$ in the lower half plane, where $0 < \varphi < \pi/2$ is chosen such that $(\pi - \varphi)(1-2p) < \pi/2$. In particular, we can choose $\mathcal{H}'$ in a way that $\Re(\delta(-v)^{1-2p} + y) < 0$ for all $v \in \mathcal{H}'$; note that $\delta = \delta(p) < 0$.

Hence, after interchanging the integral and the series and by evaluating the exponential series

$$
\sum_{k \geq 0} \frac{(\delta(-v)^{1-2p} x)^k}{k!} = e^{\delta(-v)^{1-2p} x}
$$

we can compute the (inner) integral

$$\int_0^\infty e^{(\delta(-v)^{1-2p}+y)x} = \frac{-1}{\delta(-v)^{1-2p}+y}$$

and finally get

$$\int_0^\infty f(x)e^{yx}\,\mathrm{d}x = \frac{1}{2\pi i}\int_{\mathcal{H}'}\left(\Phi(y,v) - \frac{p}{v(1-p)}\right)e^{-v}\mathrm{d}v.$$

It is clear that $\mathcal{H}'$ can be (again) replaced by $\mathcal{H}$ and so the result follows. □

Limiting distribution for $1/2 \le p < 1$. We will consider here the proofs of the limiting distribution of Theorems 3 and 4 which can be proved together. For the weak convergence, we will proceed as in the previous paragraph. The following two lemmas can be proved with the same method as before.

Lemma 15 *Let* $|t| < \eta$ *and*

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

where $\eta, \delta > 0$. *Then,* $\tilde{V}(it,z)$ *and* $\tilde{V}(it,z)$ *are analytic in* $\tilde{\Delta}$ *and satisfy*

$$\tilde{V}(it,z) = d(t)(z-1)^{\left(1-\sqrt{1-4p(1-p)e^{it}}\right)/2} + \mathcal{O}\left((z-1)^{\left(1+\sqrt{1-4p(1-p)e^{it}}\right)/2}\right), \tag{25}$$

$$\begin{aligned}\tilde{V}'(it,z) = d(t)&\left(\frac{1-\sqrt{1-4p(1-p)e^{it}}}{2}\right)(z-1)^{\left(-1-\sqrt{1-4p(1-p)e^{it}}\right)/2}\\ &+\mathcal{O}\left((z-1)^{\left(-1+\sqrt{1-4p(1-p)e^{it}}\right)/2}\right),\end{aligned} \tag{26}$$

where

$$\begin{aligned}d(t) = &\frac{\Gamma(\sqrt{1-4p(1-p)e^{it}}/2)}{\Gamma(1/2+(1-p)e^{it/2}+\sqrt{1-4p(1-p)e^{it}}/2)}\\ &\times(2(1-p))^{\left(1+\sqrt{1-4p(1-p)e^{it}}\right)/2}c(it).\end{aligned}$$

Lemma 16 *For* η, δ *sufficiently small,* $\tilde{V}(it,z)$ *as a function of* z *has no zeros in* $\tilde{\Delta}$.

From these two lemmas, we prove the following result.

Lemma 17 *Let* $|t| < \eta$ *with* η *sufficiently small. Then,*

$$\mathbb{E}\left(e^{itX_n}\right) \longrightarrow \frac{1-\sqrt{1-4p(1-p)e^{it}}}{2(1-p)}.$$

Proof Obviously, we can assume that $|t| > 0$. The proof is then similar to the one of Lemma 14. We only highlight differences. First, as in the proof of Lemma 14, we obtain that

$$\mathbb{E}\big(e^{itX_n}\big) = -\frac{1}{2(p-1)\pi i}\int_{\gamma}\frac{\tilde{V}'(it,w)}{\tilde{V}(it,w)}\frac{\mathrm{d}w}{(e^{-it/2}(w-1)+1)^n}.$$

Then, we again deform the contour to $\gamma' = \gamma_1' \cup \gamma_2' \cup \gamma_3'$.

The treatment of the integral over γ_2' and γ_3' is completely the same as in the proof of Lemma 14. Thus, we only have to concentrate on γ_1'. Here, we have from (25) and (26),

$$\frac{\tilde{V}'(it,w)}{\tilde{V}(it,w)} = \frac{1-\sqrt{1-4p(1-p)e^{it}}}{2}nv^{-1}\left(1+\mathcal{O}\left(\left(\frac{\log^2 n}{n}\right)^{\Re(\sqrt{1-4p(1-p)e^{it}})}\right)\right).$$

Note that the above real part is positive for $|t|$ small (even in the boundary case $p = 1/2$). Moreover, we have

$$(e^{-it/2}(w-1)+1)^{-n} = e^{-e^{-it/2}v}\left(1+\mathcal{O}\left(\frac{\log^2 n}{n^2}\right)\right).$$

Plugging into the above integral yields

$$\begin{aligned}
&-\frac{1}{2(p-1)\pi i}\int_{\gamma_1}\frac{\tilde{V}'(it,w)}{\tilde{V}(it,w)}\frac{\mathrm{d}w}{(e^{-it/2}(w-1)+1)^n}\\
&\quad= \frac{1-\sqrt{1-4p(1-p)e^{it}}}{2(p-1)}\left(-\frac{1}{2\pi i}\int_{\mathcal{H}_n^{(1)}}v^{-1}e^{-e^{-it/2}v}\mathrm{d}v\right)+o(1).
\end{aligned}$$

The last integral can be reduced to

$$-\frac{1}{2\pi i}\int_{\mathcal{H}_n^{(1)}}v^{-1}e^{-e^{-it/2}v}\,\mathrm{d}v = \frac{1}{2\pi i}\int_{\mathcal{H}}v^{-1}e^{-v}\,\mathrm{d}v + o(1) = 1 + o(1),$$

where the last step follows from Hankel's integral representation of $1/\Gamma(z)$. Thus, the part of the integral over γ_1' gives the main contribution and the result follows. □

We can now finish the proof of Theorems 3 and 4.

Proof of the limiting distribution result for $1/2 < p \le 1$. The weak convergence part follows from the last lemma. We just add that

$$\frac{1-\sqrt{1-4p(1-p)x}}{2(1-p)} = \sum_{k\ge 1}\frac{p^k(1-p)^{k-1}}{2(2k-1)}\binom{2k}{k}x^k.$$

Hence the limiting distribution of X is discrete with

$$P(X = k) = \frac{p^k(1-p)^{k-1}}{2(2k-1)}\binom{2k}{k}.$$

Next, we prove that when $1/2 < p < 1$, then also all moments converge. To this end, as in the case $0 < p < 1/2$, we only need to show that with the e_k's from Theorem 4, the following series

$$E(z) = \sum_{k\geq 1} e_k \frac{z^k}{k!}$$

has a positive radius of convergence. In fact, using the recurrence for e_k, we can directly show that

$$E(z) = \frac{\sqrt{1-4p(1-p)e^z}}{2} - 1$$

as must be the case. To prove this, note that the recurrence for e_k implies that

$$(2p-1)E'(z) - p = 2(1-p)E(z)E'(z) + p(e^z - 1)$$

with $E(0) = 0$. Integrating gives

$$(2p-1)E(z) - pz = (1-p)(E(z))^2 + p(e^z - z - 1).$$

Thus,

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

This proves the claimed result. □

References

Beals R, Wong R (2010) Special functions: a graduate text. Cambridge studies in advanced mathematics, vol 126. Cambridge University Press, Cambridge

Billingsley P (1995) Probability and measure, 3rd edn. Wiley series in probability and mathematical statistics. A Wiley-Interscience Publication/Wiley, New York

Blum MGB, François O (2005) Minimal clade size and external branch length under the neutral coalescent. Adv Appl Probab 37(3):647–662

Blum MGB, François O, Janson S (2006) The mean, variance and limiting distributions of two statistics sensitive to phylogenetic tree balance. Ann Appl Probab 16(4):2195–2214

Chang H, Fuchs M (2010) Limit theorems for patterns in phylogenetic trees. J Math Biol 60(4):481–512

Chern H-H, Fuchs M, Hwang H-K (2007) Phase changes in random point quadtrees. ACM Trans Algorithms 3(2):12

Durand E, Blum MGB, Francois O (2007) Prediction of group patterns in social mammals based on a coalescent model. J Theor Biol 249(2):262–270

Durand E, Francǫis O (2010) Probabilistic analysis of a genealogical model of animal group patterns. J Math Biol 60(3):451–468

Drmota M, Fuchs M, Lee Y-W (2014) Limit laws for the number of groups formed by social animals under the extra clustering model. In: Proceedings of the 25th international meeting on probabilistic, combinatorial and asymptotic methods for the analysis of algorithms. Discrete mathematics and theoretical computer science proceedings, pp 73–85

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

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

Fill JA, Kapur N (2004) Limiting distributions for additive functionals on Catalan trees. Theor Comput Sci 326(1–3):69–102

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

Flajolet P, Lafforgue T (1994) Search costs in quadtrees and singularity perturbation asymptotics. Discrete Comput Geom 12(2):151–175

Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge

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

Janson S (2014) Maximal clades in random binary search trees. Electron J Combin 22(1):31

Janson S, Kersting G (2011) On the total external length of the Kingman coalescent. Electron J Probab 16:2203–2218

Kingman JFC (1982) The coalescent. Stoch Process Appl 13(3):235–248

Mahmoud HM (1992) Evolution of random search trees. Wiley-Interscience, New York