Distributional Analysis of the Extra-Clustering Model with Uniformly Generated Phylogenetic Trees

Michael Fuchs, Chih-Hong Lee, Ariel R. Paningbatan*
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, 300
Taiwan

July 24, 2019

Abstract

We study the extra-clustering model for the group formation process of social animals when the underlying phylogenetic tree is generated by the uniform model. Moments and limit laws for the number of groups, the number of groups of fixed size and the largest group size are derived. Our results show that on average, there is only a finite number of groups one of which is very large whereas all others are small. This behavior is compared with the results of Durand and Françqis (2010) and Drmota, Fuchs and Lee (2014, 2016) who studied the extra-clustering model with phylogenetic trees generated by the Yule-Harding model.

1 Introduction

The *extra-clustering model*, proposed by Durand et al. in [4], is a model for the group formation process of social animals. Under this model, the number of groups N_n formed by n animals satisfies the distributional recurrence: for $n \geq 2$,

$$N_n \stackrel{d}{=} \begin{cases} N_{I_n} + N^*_{n-I_n}, & \text{if } K_n = 0 \text{ and } I_n \notin \{1, n-1\}; \\ 1, & \text{if } K_n = 1 \text{ or } I_n \in \{1, n-1\}, \end{cases} \tag{1}$$

where N_n^* is an independent copy of N_n, the sequences of random variables K_n, I_n, N_n, N_n^* are independent, K_n is a Bernoulli random variable with $\mathbb{P}(K_n = 1) = p$ and $0 \leq p < 1$, and I_n has throughout this note the *Catalan distribution*:

$$\mathbb{P}(I_n = j) = \frac{C_{j-1}C_{n-j-1}}{C_{n-1}}, \qquad 1 \leq j \leq n-1, \tag{2}$$

where $C_n = \frac{1}{n+1}\binom{2n}{n}$ denotes the n-th Catalan number. (That this is indeed a random distribution will become clear below.)

We give a brief description of the aforementioned extra-clustering model which then will also explain the above recurrence for the number of groups.

First, consider $p = 0$, where the model is called the *neutral model*. In this case, the model is based on the assumption that the main (and in fact) only driving force behind the group formation process

*All three authors are partially supported by the grants MOST-104-2923-M-009-006-MY3 and MOST-107-2115-M-009-010-MY2

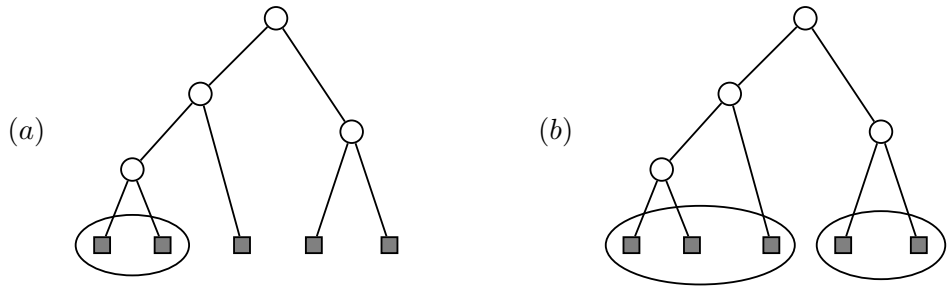

Figure 1: A phylogenetic tree representing the interrelationship of 5 animals (grey leaves; labels are omitted). On the left, the two encircled nodes are the clade of the first or second leaf from the left; note that this clade is not maximal since it is strictly contained in the clade of the third leaf from the left. On the right, the two maximal clades of the tree (which arise from the third and forth or fifth leaf from the left). So, if the the interrelationship of 5 animals is represented by this tree, then $N_n = 2$, $N_n^{[2]} = N_n^{[3]} = 1$, $N_n^{[m]} = 0$ for $m \geq 3$, and $M_n = 3$.

is genetic relatedness; see [4]. Thus, we first need to understand the interrelationship between the n animals which is done via *phylogenetic trees*, i.e., rooted, binary, leaf-labeled trees where we do not consider a left-right order of the children of nodes and leaves represent the n animals; see Semple and Steel [8] for a comprehensive introduction into properties of such trees and Figure 1 for an example (where labels of leaves are omitted). A *clade* of a leaf of such a tree is the set of leaves contained in the tree which is rooted at the parent of the leaf; see Figure 1, (a). The reason for considering clades is that the leaves (resp. animals) from a clade can be considered to be all closely related. Of particular interest are *maximal clades*, i.e., clades which are maximal under set inclusion in the poset of all clades; see Figure 1, (b). The set of all maximal clades is taken to be the set of groups formed by the n animals under the neutral model and its cardinality is denote by N_n (which so far is not random).

Of course, we usually do not have the phylogenetic tree representing the interrelationship of the n animals and thus we need to resort to probabilistic tree reconstruction methods. More precisely, we will consider random models on the set of all phylogenetic trees of size n. The most simple and basic models for such *random phylogenetic trees* are the Yule-Harding model and the uniform model (also called PDA model in the biological literature); see [8]. Properties of the (now random) N_n if the former model is used where studied by Durand and Françoıs [5] and Drmota et al. [2, 3]. In this paper, we are interested in the latter model which assumes that each phylogenetic tree with n leaves is equally likely.

Note that the distribution of N_n for a random phylogenetic tree of size n does not change if one considers a left-right order of the children of the nodes in trees and also if one ignores the labels of the leaves; see for instance the discussion in Blum et al. [1] where this was also used. Thus, we will from now on (with a slight abuse of notation) consider phylogenetic trees as rooted, binary with children of nodes having a left-right order and leaves having no labels. It is a basic combinatorial fact that the number of such trees with n leaves is given by C_{n-1}. Thus, under the uniform model, each tree with n leaves has probability $1/C_{n-1}$ and the probability that the left subtree has size j is given by (2) since there are exactly $C_{j-1}C_{n-j-1}$ trees whose left subtree has size j. (This now also shows that (2) is indeed a random distribution.)

We can now explain the above distributional recurrence for the number of groups N_n. Recall, that N_n is the number of maximal clades of a random phylogenetic tree of size n under the uniform model. It is immediate that this number can be computed as the sum of the number of maximal clades of the left and right subtree unless all leaves are in one maximal clade. This, together with the fact that if the left subtree size equals j, then left and right subtree are independent random phylogenetic trees of size j and $n - j$, respectively, explains the above distributional recurrence for $p = 0$. (In particular, note that

in this case, we have $K_n \equiv 0$.)

Next, we are going to explain the more general extra-clustering model. Recall that, as just explained, the neutral model is based on the assumption that the only reason for animals to form groups is genetic relatedness. Whereas for some types of social animals this assumption is reasonable, for others it is not; see [4] where this was discussed with real-world data. In order to take into account other factors which cause animals to form groups (and also in order to devise statistical tests to test whether or not the neutral model is appropriate), the authors in [4] introduced the more general extra-clustering model. Here, one in addition has a probability p which measures the degree of which other factors are decisive in the group formation process. According to this probability, in each step of the recursive procedure to compute N_n, it may happen independently from everything else that an extra-clustering event occurs which means all the remaining animals are in one cluster. These extra-clustering events are modeled via the random variable K_n in (1) which was the last un-explained piece in (1). Thus, the distributional recurrence for N_n is now fully explained.

In [2, 3, 5], moments and limit laws of N_n for the Yule-Harding model were studied. Here, we will prove corresponding results for N_n as well as for more refined characteristics of the group formation process under the extra-clustering model with uniformly chosen random phylogenetic trees, i.e., for the uniform model.

The paper is organized as follows. In the next section, we introduce the so-called cluster tree and will associate two important generating functions with it. This will then be used in Section 3 to derive limit distribution results for N_n and the number of groups containing exactly m animals where $m \geq 2$. Finally, in Section 4 we will study moments and the limit distribution of the largest group size. We will conclude in Section 5 by comparing the results from this paper with those for the Yule-Harding model from previous works.

2 Cluster Trees and Weights

In order to find moments of N_n, one could work with the distributional recurrence (1). However, we will use a more combinatorial method which will turn out to be advantageous when dealing with more refined properties of the group formation process.

First, note that the definition of the extra-clustering model can be broken into two probabilistic stages: (i) a phylogenetic tree of size n is picked uniformly at random and (ii) the picked tree is traced starting from the root and one stops if either a node is encountered whose left or right subtree is a leaf or an extra clustering event has occurred. In the second step, we replace the subtrees at the places where one has stopped by leaves and call the resulting tree a *cluster tree* of the picked tree. Note that cluster trees are again rooted, binary trees with children having a left-right order and leaves not labeled. Moreover, note that they are not unique but rather depend on the outcome of the probabilistic procedure in Step (ii) above; see Figure 2 for all the possible cluster trees associated with the tree from Figure 1.

Now, in order to keep track of the probabilities attached to cluster trees, we associate two generating functions with them. First, since no extra-clustering event has occurred at any internal node of a cluster tree, we attach the probability $q := 1 - p$ to these nodes, i.e., we consider

$$G(z) := \sum_{n \geq 1} q^{n-1} C_{n-1} z^n = zC(qz),$$

where $C(z)$ is the ordinary generating function of the Catalan numbers (see, e.g., Page 34 and 35 in [6]):

$$C(z) = \frac{1 - \sqrt{1-4z}}{2z}.$$

Next, for the leaves of the cluster tree, they either resulted from an extra-clustering event (in which case we have to attach the weight p to them and there are C_{n-1} possible trees) or they have been nodes

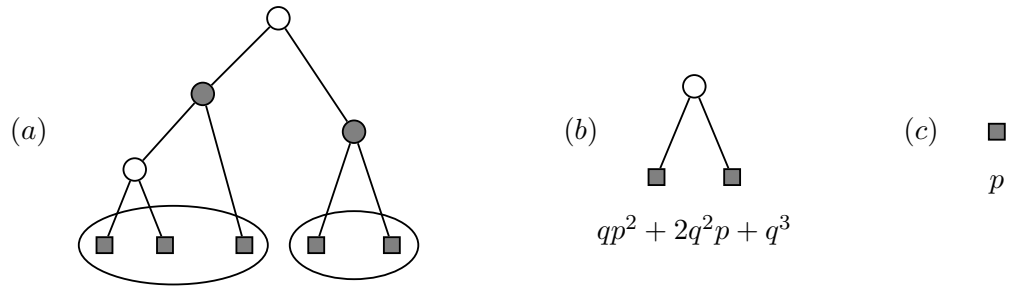

Figure 2: The phylogenetic tree from Figure 1 together with its cluster trees in (b) and (c). The tree is traced from the root until extra-clustering events occur and/or the grey internal nodes are reached (which are the parents of the leafs whose clades are maximal). The shapes of all possible cluster trees are in (b) and (c): the one in (b) occurs 4 times (depending on whether there are extra-clustering events at the leaves or not) and the one in (c) only occurs if there is an extra-clustering event at the root. The probabilities are indicated below the shapes (in (b) it is the sum of the four probabilities) and as explained in the paragraph preceding Lemma 1 they indeed sum up to 1.

in the original phylogenetic tree with either one or two children as leaves (in this case, we use the weight q and the number of trees is either $2C_{n-2}$ in the former case or 1 in the latter case). Thus, for each single leaf of the cluster tree, we consider the generating function

$$\begin{aligned} H(z) :&= (pC_1 + q)z^2 + \sum_{n\geq 3}(pC_{n-1} + 2qC_{n-2})z^n \\ &= z^2 + pz(C(z) - 1 - z) + 2qz^2(C(z) - 1). \end{aligned}$$

Now, the composition of these two generating functions, namely, $G(H(z))$ generates for any phylogenetic trees all its associated cluster trees with their corresponding probabilities. In particular, since for each phylogenetic tree the probabilities of its cluster trees sum up to 1, we have

$$[z^n]G(H(z) = C_{n-1}, \qquad (n \geq 2)$$

and $[z]G(H(z)) = 0$ because all the phylogenetic trees have at least two leaves. We formulate this as a lemma.

Lemma 1. *For all* $0 \leq p < 1$*, we have*

$$G(H(z)) = z(C(z) - 1).$$

3 Number of Groups and Number of Fixed-Size Groups

The two generating functions from the previous section become really useful only if one introduces a second variable, say u, which keeps track of the number of leafs of the cluster trees which by definition is the number of groups under the extra-clustering model. More precisely, we consider now $G(uH(z))$. By the above description, this generating function is related to the distribution of N_n via

$$\mathbb{P}(N_n = k) = \frac{[u^k z^n]G(uH(z))}{C_{n-1}},$$

where the denominator incorporates the probability from Step (i) of the stochastic description of the extra-clustering model from the beginning of the last section and $G(uH(z))$ incorporates the probabilities from Step (ii).

Limit laws for random variables arising in the above way from a composition of generating functions have been studied before in the literature; see Flajolet and Sedgewick [6]. We recall here one such result which we will use in the sequel. To state the result, we need some notations.

Assume that $g(z)$ and $h(z)$ are two generating functions with non-negative coefficients and $h(0) = 0$. Denote the radii of convergence of $g(z)$ and $h(z)$ by ρ_g and ρ_h, respectively. Moreover, set

$$\tau_g := \lim_{z \to \rho_g -} g(z) \qquad \text{and} \qquad \tau_h := \lim_{z \to \rho_h -} h(z).$$

Then, the following result holds.

Theorem 1 (Proposition IX.1 in [6])**.** *Assume that $\tau_h < \rho_g$. Moreover, assume that ρ_h is finite and that $z = \rho_h$ is the only singularity of $h(z)$ on the circle of convergence. Finally, assume that*

$$h(z) = \tau_h - c\left(1 - \frac{z}{\rho_h}\right)^{\lambda} + o\left(\left(1 - \frac{z}{\rho_h}\right)^{\lambda}\right),$$

where c is a positive real number, $0 < \lambda < 1$ and the above asymptotics holds for z in

$$\Delta := \{z \ : \ |z| < r \text{ and } |\arg(z - \rho_h)| > \phi\},$$

where $r > 1$ and $0 < \phi < \pi/2$.

Then, for the sequence of random variables defined by

$$\mathbb{P}(X_n = k) := \frac{[u^k z^n] g(uh(z))}{[z^n] g(h(z))},$$

we have the limit distribution result

$$X_n \xrightarrow{d} X$$

with convergence of all moments, where X is a discrete random variable with probability generating function:

$$P_X(u) = \frac{ug'(u\tau_h)}{g'(\tau_h)}.$$

The proof of this result follows from singularity analysis, where the dominant singularity (the one closest to the origin) of $g(h(z))$ comes from the dominant singularity of $h(z)$ since $\tau_h < \rho_g$; for details see [6] and the proof of Theorem 3 below. The condition $\tau_h < \rho_g$ is the so called *subcritical case* and one usually refers to $f(ug(z))$ as *subcritical composition schema*.

In fact, $G(uH(z))$ is also a subcritical composition schema and thus the limit distribution of N_n follows from the above result.

Theorem 2. *We have the limit distribution result*

$$N_n \xrightarrow{d} N$$

with convergence of all moments, where

$$N \stackrel{d}{=} \mathrm{NB}\left(\frac{1}{2}, \frac{3 - 2p - p^2}{4}\right) + 1.$$

Here, $\mathrm{NB}(r, p)$ denotes the negative binomial distribution.

Remark 1. NB in the above theorem is more precisely the (standard) generalization of the negative binomial distribution to the case where the first parameter is allowed to be any positive real number.

Proof. First, note that

$$\rho_H = \frac{1}{4}, \rho_G = \frac{1}{4q} \qquad \text{and} \qquad \tau_H = \frac{3+p}{16},\ \tau_G = \frac{1}{2q}.$$

Since

$$\tau_H = \frac{3+p}{16} < \frac{1}{4} \le \frac{1}{4q} = \rho_G,$$

$G(uH(z))$ is indeed a subcritical composition schema. Moreover, by a straightforward expansion

$$H(z) = \frac{3+p}{16} - \frac{1+p}{4}\sqrt{1-4z} + o(\sqrt{1-4z})$$

in a suitable Δ-domain.

Thus, by applying the proposition, we obtain the claimed result with the probability generating function of N given by

$$P_N(u) = u\sqrt{\frac{1-4q\tau_H}{1-4q\tau_H u}}.$$

From this, it is clear that N has the claimed distribution. ∎

Remark 2. The previous theorem can also be proved by deriving the asymptotics of all moments of N_n which can be done in a recursive way since it follows from (1) that all moments satisfy the same type of recurrence. Then, the above result also follows since the negative binomial distribution is uniquely determined by its moment sequence; see, e.g., [2, 3] where such a recursive approach was employed to prove limit distribution results (but with other limiting distributions).

As a corollary, we obtain the following.

Corollary 1. *We have,*

$$\mathbb{E}(N_n) \sim \frac{5+2p+p^2}{2+4p+2p^2}.$$

Thus, on average, there are only a finite number of groups.

Next, we fix $m \ge 2$ and consider the number of groups of size m which we denote by $N_n^{[m]}$; see the description of Figure 1 for an example. In order to understand the distribution of this random variable we can again use the two generating functions $G(z)$ and $H(z)$. However, this time we only mark with u those leaves of the cluster tree which correspond to groups of size m, i.e., only the coefficient of $[z^m]$ in $H(z)$. Thus, we consider

$$G((pC_{m-1} + (2-\delta_{2,m})qC_{m-2})(u-1)z^m + H(z)),$$

were $\delta_{2,m}$ is the Kronecker delta function. Then,

$$\mathbb{P}(N_n^{[m]} = k) = \frac{[u^k z^n]G((pC_{m-1} + (2-\delta_{2,m})qC_{m-2})(u-1)z^m + H(z))}{C_{n-1}}.$$

In order to find the limit distribution of $N_n^{[m]}$, we cannot directly apply Theorem 1. However, the method of proof of Theorem 1 can be applied and yields the following result.

Theorem 3. *We have the limit distribution result*

$$N_n^{[m]} \xrightarrow{d} N^{[m]}$$

with convergence of all moments, where

$$N^{[m]} \stackrel{d}{=} \mathrm{NB}\left(\frac{1}{2}, \frac{4^{2-m}(1-p)(pC_{m-1} + (1-p)(2-\delta_{2,m})C_{m-2})}{1+2p+p^2+4^{2-m}(1-p)(pC_{m-1} + (1-p)(2-\delta_{2,m})C_{m-2})}\right).$$

Proof. Let

$$H_m(u,z) = (pC_{m-1} + (2-\delta_{2,m})qC_{m-2})(u-1)z^m + H(z)$$

which has dominant singularity at $z = 1/4$. By a straightforward expansion, as $z \to 1/4$,

$$H_m(u,z) = c_m(u) - \frac{1+p}{4}\sqrt{1-4z} + o(\sqrt{1-4z}),$$

where

$$c_m(u) = (pC_{m-1} + (2-\delta_{2,m})qC_{m-2})(u-1)4^{-m} + \frac{3+p}{16}.$$

Note that for u close to 1, we have

$$|c_m(u)| < \frac{1}{4q}$$

and the upper bound is the dominant singularity of $G(z)$. Thus, $G(H_m(u,z))$ has also dominant singularity at $z = 1/4$. Moreover, as $z \to 1/4$,

$$G(H_m(u,z)) = \frac{1-\sqrt{1-4qc_m(u)}}{2q} - \frac{1+p}{4\sqrt{1-4qc_m(u)}}\sqrt{1-4z} + o(\sqrt{1-4z}).$$

Now, by the transfer theorems of singularity analysis (see Chapter VI in [6]),

$$[z^n]G(H_m(u,z)) \sim \frac{1+p}{8\sqrt{\pi}\sqrt{1-4qc_m(u)}} \cdot \frac{4^n}{n^{3/2}}$$

and by using the well-known expansion of the Catalan numbers

$$C_n = \frac{4^n}{\sqrt{\pi}n^{3/2}}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right), \tag{3}$$

we obtain that

$$P_{N_n^{[m]}}(u) \sim \frac{1+p}{2\sqrt{1-4qc_m(u)}},$$

where $P_{N_n^{[m]}}(u)$ denotes the probability generating function of $N_n^{[m]}$. From this the claimed result follows by standard results from probability theory. ∎

Remark 3. Again this result can alternatively be proved via the asymptotics of moments since $N_n^{[m]}$ satisfies the following distribution recurrence: for $n > m$,

$$N_n^{[m]} = \begin{cases} N_{I_n}^{[m]} + (N_{n-I_n}^{[m]})^*, & \text{with probability } 1-p \text{ and } I_n \notin \{1, n-1\}; \\ 0, & \text{otherwise,} \end{cases}$$

where notation is as in (1) and initial conditions are given by $N_n^{[m]} = 0$ if $n < m$ and

$$N_m^{[m]} = \begin{cases} 0, & \text{with probability } 1-p \text{ and } I_m \notin \{1, m-1\}; \\ 1, & \text{otherwise.} \end{cases}$$

As a consequence, we again obtain the asymptotics of the mean.

Corollary 2. *We have,*

$$\mathbb{E}(N_n^{[m]}) \sim 2\frac{4^{1-m}(1-p)(pC_{m-1} + (1-p)(2-\delta_{2,m})C_{m-2})}{1+2p+p^2}.$$

Corollary 1 and Corollary 2 now imply the following proposition.

Proposition 1. *We have,*

$$\mathbb{E}(N) = 1 + \sum_{m \geq 2} \mathbb{E}(N^{[m]}).$$

Proof. This is proved by a straightforward computation (probably best done with mathematical software such as Maple). ∎

This suggests that there is only one big group and all other groups are small. That this is indeed the case will be proved in the next section.

4 Largest Group Size

Denote by M_n the largest size of the groups (i.e. largest size of the maximal clades) of a random phylogenetic tree of size n under the uniform model; e.g., for the tree in Figure 1 we have $M_n = 3$. Due to the above observation that there should be one big group, we set $X_n := n - M_n$.

In order to find the distribution of X_n, we again make use of the above two generating functions for the cluster tree. The main observation is that for $0 \le k < n/2$, we have

$$\mathbb{P}(X_n = k) = \frac{[z^k]G'(H(z))[z^{n-k}]H(z)}{C_{n-1}}$$

which is explained as follows: since the largest group size is equal to $n-k$, we have to replace one leaf of the cluster tree by a group of size $n-k$ (this is the factor $[u^{n-k}]H(z)$), whereas all other leaves are replaced by arbitrary groups (this is the factor $[z^k]G'(H(z))$); note that the restriction $0 \le k < n/2$ is essential here, because it ensures that all other groups are indeed of size smaller than $n-k$. Moreover, the range $0 \le k < n/2$ is expected to be sufficient for our purpose since we expect that the largest group size is close to n.

We start with the following lemma.

Lemma 2. *Uniformly for $0 \le k < n/2$, we have*

$$\mathbb{P}(X_n = k) = \frac{1+p}{2} 4^{-k} [z^k] G'(H(z)) \left(1 - \frac{k}{n}\right)^{-3/2} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Proof. Note that

$$[z^{n-k}]H(z) = pC_{n-k-1} + 2qC_{n-k-2}.$$

The result follows from this by a standard computation using (3). ∎

From the last lemma, we obtain the limit distribution of X_n.

Theorem 4. *We have the limit distribution result*

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

where X is a discrete random variable with probability generating function

$$P_X(u) = \sum_{k \ge 0} p_k u^k = \frac{1+p}{2F(u/4)}.$$

Here,

$$F(u) = \sqrt{1 - 2p + 2p^2 - 4(1-2p)(1-p)z + 4(1-p)^2 z^2 - 2(1-p)(p - 2(1-p)z)\sqrt{1-4u}}. \tag{4}$$

Proof. From Lemma 2, we have for fixed k

$$p_k := \lim_{n \to \infty} \mathbb{P}(X_n = k) = \frac{1+p}{2} 4^{-k} [z^k] G'(H(z)).$$

Thus,

$$P_X(u) = \sum_{k \ge 0} p_k z^k = \frac{1+p}{2} G'(H(u/4))$$

and the claimed form follows now by plugging into this the expressions for $G(z)$ and $H(z)$ and straightforward computation. ∎

Remark 4. Note that $F(u)$ has dominant singularity at $u = 1/4$. Moreover, as $u \to 1/4$,

$$P_X(u) = 1 - \frac{2(1-p)}{1+p}\sqrt{1-u} + o(\sqrt{1-u}).$$

From this, we obtain by the transfer theorems of singularity analysis,

$$p_k = \frac{1-p}{(1+p)\sqrt{\pi}k^{3/2}}\left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right), \qquad (k \to \infty). \tag{5}$$

Remark 5. Note that all moments of X are infinite. Thus, in contrast to Theorem 2 and Theorem 3, we do not have moment convergence in the above limit theorem for the largest group size.

Due to the latter remark, it is interesting to compute moments of X_n (and thus of M_n). We will do this next with the help of Lemma 2, (5) and the Euler-Maclaurin summation formula (for the latter see, e.g., Chapter 9 of Graham et al. [7]). We first need the following (crucial) lemma.

Lemma 3. *We have,*

$$\sum_{0\le k<n/2} \mathbb{P}(X_n = k) = 1 + o(n^{-1/2}) \tag{6}$$

and for $\ell \ge 1$

$$\sum_{0\le k<n/2} k^\ell \mathbb{P}(X_n = k) \sim d_\ell n^{\ell-1/2} \tag{7}$$

where

$$d_\ell = \frac{1-p}{(1+p)\sqrt{\pi}}\int_0^{1/2} x^{\ell-3/2}(1-x)^{-3/2}\mathrm{d}x.$$

Proof. We will derive the asymptotics of the sum in (6) by splitting it into two parts:

$$\sum_{0\le k<n/2} \mathbb{P}(X_n = k) = \sum_{0\le k<n^\rho} \mathbb{P}(X_n = k) + \sum_{n^\rho\le k<n/2} \mathbb{P}(X_n = k), \tag{8}$$

where $\rho > 0$ will be chosen as the proof proceeds.

For the first part, we have by Lemma 2,

$$\sum_{0\le k<n^\rho} \mathbb{P}(X_n = k) = \sum_{0\le k<n^\rho} p_k(1 + \mathcal{O}(n^{\rho-1})) = \sum_{0\le k<n^\rho} p_k(1 + o(n^{-1/2})),$$

where p_k was defined in Theorem 4 and $\rho < 1/2$ so that the last equality holds. Note that

$$\sum_{0\le k<n^\rho} p_k = 1 - \sum_{k\ge n^\rho} p_k = 1 - \frac{1-p}{(1+p)\sqrt{\pi}}\sum_{k\ge n^\rho} k^{-3/2}(1 + \mathcal{O}(1/k)),$$

where we used (5) in the last step. Combining the two equations above, we get

$$\sum_{0\le k<n^\rho} \mathbb{P}(X_n = k) = 1 - \frac{1-p}{(1+p)\sqrt{\pi}}\sum_{k\ge n^\rho} k^{-3/2}(1 + \mathcal{O}(1/k)) + o(n^{-1/2}). \tag{9}$$

The asymptotic of the sum on the right-hand side of the equation can be derived by using the Euler-Maclaurin summation formula:

$$\sum_{k\ge n^\rho} k^{-3/2} = \int_{n^\rho}^\infty x^{-3/2}\mathrm{d}x + \mathcal{O}(n^{-3\rho/2}) = 2n^{-\rho/2} + o(n^{-1/2}),$$

where the last step holds whenever $\rho > 1/3$. The asymptotic of the $\mathcal{O}$-term in (9) can be derived in a similar manner. Thus, we obtain that

$$\sum_{0\le k<n^\rho} \mathbb{P}(X_n = k) = 1 - \frac{2(1-p)}{(1+p)\sqrt{\pi}} n^{-\rho/2} + o(n^{-1/2}). \tag{10}$$

Now, we turn to the second part of the decomposition of (8) for which we use the expansions from Lemma 2 and (5):

$$\sum_{n^\rho\le k<n/2} \mathbb{P}(X_n = k) = \frac{1-p}{(p+1)\sqrt{\pi}} \sum_{n^\rho\le k<n/2} k^{-3/2}(1-k/n)^{-3/2}(1+\mathcal{O}(1/k)). \tag{11}$$

Using again Euler-Maclaurin summation formula,

$$\sum_{n^\rho\le k<n/2} k^{-3/2}(1-k/n)^{-3/2} = \int_{n^\rho}^{n/2} x^{-3/2}(1-x/n)^{-3/2}\mathrm{d}x + o(n^{-1/2}).$$

Note that

$$\int x^{-3/2}(1-x/n)^{-3/2}\mathrm{d}x = \frac{2(2x-n)}{\sqrt{nx(n-x)}}$$

and thus

$$\sum_{n^\rho\le k<n/2} k^{-3/2}(1-k/n)^{-3/2} = 2n^{-\rho/2} + o(n^{-1/2}).$$

Together with a similar treatment of the $\mathcal{O}$-term in (11), we obtain that

$$\sum_{n^\rho\le k<n/2} \mathbb{P}(X_n = k) = \frac{2(1-p)}{(1+p)\sqrt{\pi}} n^{-\rho/2} + o(n^{-1/2}). \tag{12}$$

Finally, substituting (10) and (12) into (8) gives the desired result.

Next, we proceed to the proof of (7). In a similar manner, we split the sum into

$$\sum_{0\le k<n/2} k^\ell \mathbb{P}(X_n = k) = \sum_{0\le k<n^\rho} k^\ell \mathbb{P}(X_n = k) + \sum_{n^\rho\le k<n/2} k^\ell \mathbb{P}(X_n = k), \tag{13}$$

where ρ is again chosen as the proof proceed.

For the first term on the right-hand side of (13):

$$\sum_{0\le k<n^\rho} k^\ell \mathbb{P}(X_n = k) \le n^{\rho\ell} = o(n^{\ell-1/2}),$$

where the last step holds when $\rho < 1/2$.

For the second term on the right-hand side of (13), we again apply the expansions in Lemma 2 and (5):

$$\sum_{n^\rho\le k<n/2} k^\ell \mathbb{P}(X_n = k) = \frac{1-p}{(1+p)\sqrt{\pi}} \sum_{n^\rho\le k<n/2} k^{\ell-3/2}(1-k/n)^{-3/2}(1+\mathcal{O}(1/k)). \tag{14}$$

Using once more Euler-Maclaurin summation formula yields

$$\begin{aligned}
\sum_{n^\rho\le k<n/2} k^{\ell-3/2}(1-k/n)^{-3/2} &= \int_{n^\rho}^{n/2} x^{\ell-3/2}(1-x/n)^{-3/2}\mathrm{d}x + o(n^{\ell-1/2}) \\
&= \int_{0}^{n/2} x^{\ell-3/2}(1-x/n)^{-3/2}\mathrm{d}x + o(n^{\ell-1/2}) \\
&= \left(\int_0^{1/2} x^{\ell-3/2}(1-x)^{-3/2}\mathrm{d}x\right) n^{\ell-1/2} + o(n^{\ell-1/2}).
\end{aligned}$$

The $\mathcal{O}$-term in (14) is treated similarly.

Finally, substituting the above two equations into (13) gives the desired result. ∎

From this lemma, we obtain now the asymptotics of all moments of X_n.

Theorem 5. *For $\ell \geq 1$, we have*

$$\mathbb{E}(X_n^\ell) \sim d_\ell n^{\ell-1/2},$$

where d_ℓ is as in Lemma 3.

Proof. Since

$$\mathbb{E}(X_n^\ell) = \sum_{0\leq k\leq n} k^\ell \mathbb{P}(X_n = k) = \sum_{0\leq k<n/2} k^\ell \mathbb{P}(X_n = k) + \sum_{n/2\leq k\leq n} k^\ell \mathbb{P}(X_n = k)$$

we only need to show that the second term is $o(n^{\ell-1/2})$. This follows directly from

$$\sum_{n/2\leq k\leq n} k^\ell \mathbb{P}(X_n = k) \leq n^\ell \left(1 - \sum_{0\leq k<n/2} \mathbb{P}(X_n = k)\right) = o(n^{\ell-1/2}),$$

where (6) is used in the last estimate. ∎

As a corollary, we obtain the asymptotics of moments of the maximal group size M_n.

Corollary 3. *We have,*

$$\mathbb{E}(M_n) = n - \frac{2(1-p)}{(1+p)\sqrt{\pi}} n^{1/2} + o(n^{1/2})$$

and for $\ell \geq 2$

$$\mathbb{E}(M_n - \mathbb{E}(M_n))^\ell \sim (-1)^\ell d_\ell n^{\ell-1/2},$$

where d_ℓ is as in Lemma 3.

5 Conclusion

In this paper, we considered the number of groups, number of fixed-size groups and the largest group size of the extra clustering model with uniformly distributed phylogenetic trees. For all these random variables, we derived limit laws and computed moments. Our results show that on average, there is only a finite number of groups and that one of these groups contains almost all animals (and thus all the others are small). This holds for all p with $0 \leq p < 1$.

Our results have to be compared with those for the extra clustering model where the phylogenetic trees are generated by the Yule-Harding model; see [5] and [2, 3]. In particular, in [5], the following asymptotics for the mean of number of groups (again denoted by N_n) was proved:

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

where

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

Thus, for the Yule-Harding model, the number of groups is on average finite if and only if $p > 1/2$. In all other cases, the number of groups is growing as n tends to infinity.

Higher moments and limit laws of N_n where discussed in [2, 3], where the authors proved that the limit law for $p = 0$ is continuous, for $0 < p < 1/2$ it is a mixture of a continuous and discrete random variables and only for $p \geq 1/2$ it becomes discrete. On the other hand, for the uniform model we proved in this paper that it is always discrete. Moreover, one also has convergence of all moments which in the Yule-Harding model was only the case for $0 < p < 1/2$ and $1/2 < p < 1$.

For the number of fixed-sized groups in the Yule-Harding model, only the mean was considered so far. For example, in [5], the authors showed that for $0 \leq p < 1/2$, the mean is again of order n^{1-2p}. Using the tools from [2, 3], higher moments and limit laws for the number of fixed-sized groups could be added as well (also for the range $p \geq 1/2$).

However, of possible greater interest would be a study of the largest group size in the Yule-Harding model, in particular, because it was claimed in [5] that the "typical" group size is of order $\log n$ in the neutral model ($p = 0$) and of order n in the extra clustering model with $p > 0$. Whether or not a similar sharp transition also holds for the maximal group size is an open problem.

References

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

[2] M. Drmota, M. Fuchs, Y.-W. Lee (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 Math. Theor. Comput. Sci. Proc.*, 73–85.

[3] M. Drmota, M. Fuchs, Y.-W. Lee (2016). Stochastic analysis of the extra clustering model for animal grouping, *J. Math. Biol.*, **73:1**, 123–159.

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

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

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

[7] R. L. Graham, D. E. Knuth, O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*, Addison-Wesley Publishing, Company, Reading, Massachusatts, 1994.

[8] C. Semple and M. Steel. *Phylogenetics*, Oxford University Press, Oxford, 2003