J. Math. Biol.
DOI 10.1007/s00285-009-0270-y

Mathematical Biology

Probabilistic analysis of a genealogical model of animal group patterns

Eric Durand · Olivier François

Received: 26 October 2007 / Revised: 5 March 2009
© Springer-Verlag 2009

Abstract Many social animals live in stable groups, and it has been argued that kinship plays a major role in their group formation process. In this study we present the mathematical analysis of a recent model which uses kinship as a main factor to explain observed group patterns in a finite sample of individuals. We describe the average number of groups and the probability distribution of group sizes predicted by this model. Our method is based on the study of recursive equations underlying these quantities. We obtain asymptotic equivalents for probability distributions and moments as the sample size increases, and we exhibit power-law behaviours. Computer simulations are also utilized to measure the extent to which the asymptotic approximation can be applied with confidence.

Keywords Animal group patterns · Genealogical models · Distributional recursions · Power-law distributions

Mathematics Subject Classification (2000) 92B99

1 Introduction

The problem of predicting animal group sizes is one of the most stimulating in theoretical biology because the tendency to aggregate is under strong evolutionary control (Gueron and Levin 1995; Rubenstein 1978). In order to gain insight into the grouping patterns of animals, several models have been proposed. These models encompass fusion and fission processes (Gueron and Levin 1995; Takayasu 1989), kin

E. Durand (✉) · O. François
TIMB Group of Mathematical Biology, TIMC UMR 5525, Fac. Méd.,
Grenoble Universités, 38706 La Tronche Cedex, France
e-mail: eric.durand@imag.fr

Published online: 12 April 2009

selection (Yokoyama and Felsenstein 1978; Hamilton 2000) and game-theoretic models (Giraldeau and Caraco 1993). In fission-fusion models, animal groups are randomly moving units which aggregate or split with given probabilities. Although very simple, fission-fusion models are flexible enough to capture some aspects of group patterns in social animals, such as power-law distributions (Bonabeau et al. 1999). Kin selection was proposed by Hamilton (1964) in order to explain the evolution of altruistic behaviour. This theory postulates that a gene favorising altruistic behaviour in an individual is likely to be carried by its relatives, and then individuals who live in groups of parents benefit from an increased fitness (Dawkins 1989). In kin selection models, genetic relatedness plays a key role in determining the animal group structure. In game-theoretic models, the group patterns evolve by putting into balance cooperation and competition (Hamilton 2000). Game-theoretic models sometimes use genetic relatedness as one of their parameters in order to predict animal group patterns (Giraldeau and Caraco 1993).

In a recent study, we introduced a new model which incorporates genetic relatedness as the main factor explaining group patterns in social mammals (Durand et al. 2007). The model is based on a random genealogy that summarizes the ancestral relationships among the n individuals represented in the sample. Given the genealogy and assuming the simplest form of the model, a group pattern can be deduced by aggregating each individual to its closest relatives in the sample. In Durand et al. (2007), the latent genealogy assumed a coalescent tree model (Kingman 1982), and the resulting aggregation process was termed the *neutral model*. In fact, a random clustering process having the same group distribution as the neutral model can be defined in the following way. Start with the n individuals, and choose two of them at random. Then merge the two chosen individuals into one unit that replaces these two individuals. Repeat this process with the $n-1, n-2, \ldots$, entities progressively formed by adding units to the remaining original individuals. Stop the process when all the individuals are put into units. The N_n random units resulting from this process define the group structure of the sample. Examples of group patterns for $n = 12$ and 15 individuals are shown in Figs. 1a and 2b. In Fig. 1a, the individuals sort into the 4 groups $\{3, 5, 9\}$, $\{1, 4\}$, $\{10, 11\}$ and $\{2, 6, 7, 8, 12\}$.

Assuming the same model as in Durand et al. (2007), we present a mathematical analysis of the expected number of groups and the distribution of group sizes in a sample of n individuals. In Sect. 2, we recall the description of the model in mathematical terms, and we provide recursive defintions for the quantities of interest. In Sect. 3 we describe our main results, and we present a short data analysis. Section 4 contains the proofs of our results.

2 Distributional recursions

Assuming a sample of n individuals, we shall describe a model with extra-clustering, extending the neutral model presented in the introduction. In the extra-clustering model, larger groups can be formed by randomly aggregating two units at each step of the construction process. The extra-clustering events occur at rate p. This model encompasses the neutral model which corresponds to the particular case

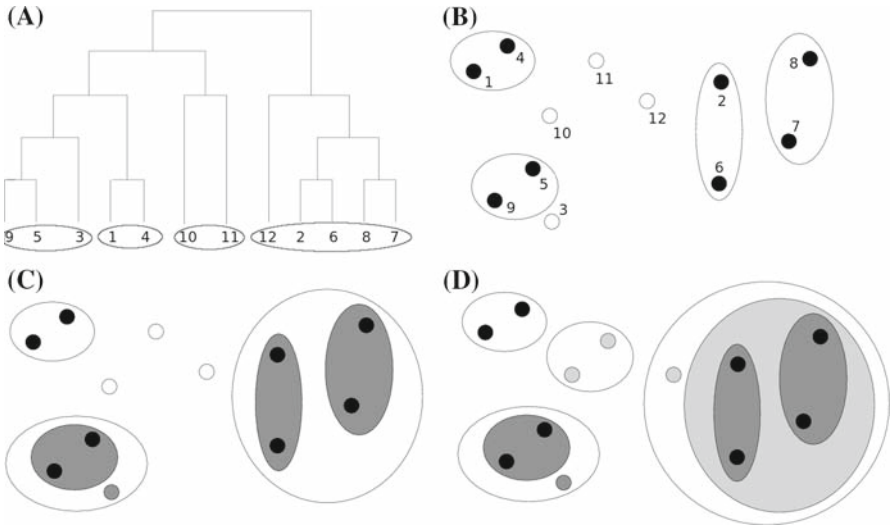

Fig. 1 Construction of group patterns illustrated. **a** $n = 12$ individuals are aggregated into four groups represented by ellipses. The tree reflects the group structure obtained for the $n = 12$ individuals. **b** Iterative construction—step 1. After four initial coalescent events, 4 provisional units are formed. **c** Iterative construction—step 2. We observe the clustering of an extra individual to a previously formed unit at the left and the aggregation of two units at the right. **d** The process is continued until all individuals are clustered to form the same groups as in **a**

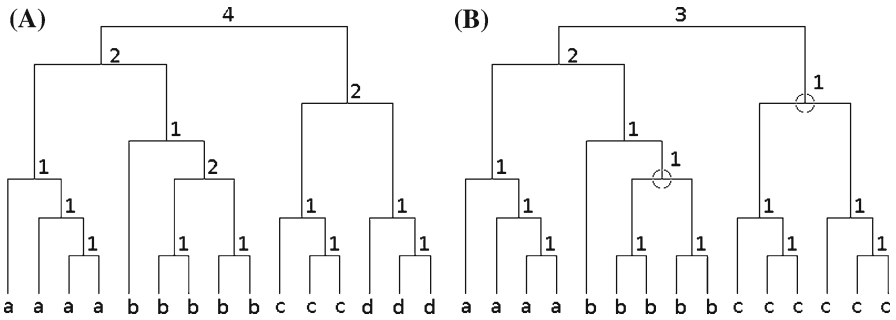

Fig. 2 Recursive computations of N_n for $n = 15$. The letters at the tips stand for the group labels. **a** Neutral model: the sample has 4 groups denoted a–d. **b** Extra clustering: Two extra-clustering events occur and are symbolized by circles. The sample has 3 groups denoted a–c

$p = 0$. A formal description of the partition of the n individuals into groups is given hereafter.

The definition of the extra-clustering model is strongly connected with the neutral *coalescent* (Kingman 1982; Tavaré 2004; Hein et al. 2005), which models the genealogy as a random binary tree. A property of coalescent tree topologies is that the size of the left sister clade at the root of the tree, L_n, has a uniform distribution over the set $\{1, \ldots, n-1\}$ (see Kingman 1982)

$$\text{Prob}(L_n = \ell) = \frac{1}{n-1}, \quad \ell = 1, \ldots, n-1. \tag{2.1}$$

Because this property also holds within each subtree, one can assume a recursive definition of the tree topology through the uniform split distribution replicated at each internal node (Aldous 1996; Blum and François 2005a). The extra-clustering model describes a partition of the n individuals, P_n, in the following way. Let us denote $R_n = n - L_n$ and $I_n = \min(L_n, R_n)$. We have $P_n =$ all n elements if $I_n = 1$. Otherwise, we have

$$\mathrm{P}_n = \begin{cases} \text{all } n \text{ elements} & \text{with probability } p \\ \mathrm{P}_{L_n} \cup \mathrm{P}^*_{R_n} & \text{with probability } 1-p\,, \end{cases} \tag{2.2}$$

where $\mathrm{P}^*_\cdot$ and $\mathrm{P}_\cdot$ are two independent copies of $\mathrm{P}_\cdot$, and are also independent from (L_n, R_n) (this assumption is true for all other quantities in the remainder of this section). The recursion is initialized with the values $\mathrm{P}_1 = \{$ a one-element subset $\}$ and $\mathrm{P}_2 = \{$ a two-elements subset $\}$ and L_n (and thus R_n) has a uniform distribution over the set $\{1, \ldots, n-1\}$.

The number of groups, N_n, can be recursively defined as follows. Let us denote $I_n = \min(L_n, R_n)$. We have $N_n = 1$ if $I_n = 1$. Otherwise, we have

$$N_n = \begin{cases} 1 & \text{with probability } p \\ N_{L_n} + N^*_{R_n} & \text{with probability } 1-p\,, \end{cases} \tag{2.3}$$

where $N^*_\cdot$ and $N_\cdot$ are two independent copies of $N_\cdot$. The recursion is initialized with the values $N_1 = N_2 = 1$, and L_n (and thus R_n) has a uniform distribution over the set $\{1, \ldots, n-1\}$. Equation (2.3) does not only provide a definition of N_n, but also an efficient way to simulate this random variable for arbitrary sample sizes. The recursive computation of N_n is illustrated in Fig. 2. In this figure, the value of the number of groups is propagated from the tips to the root of the tree.

Our second quantity of interest is the number of groups of size k, $N_n^{(k)}$, for $2 \leq k \leq n$. The collection of all $N_n^{(k)}$ forms the group pattern, or size spectrum, and we obviously have

$$N_n = \sum_{k=2}^{n} N_n^{(k)}.$$

In Fig. 2a, the size spectrum is $(0, 2, 1, 1, 0, \ldots)$, whereas in Fig. 2b, it is equal to $(0, 0, 1, 1, 1, 0, \ldots)$. For each k, the variable $N_n^{(k)}$ satisfies the following set of recursive equations. We have $N_n^{(k)} = \delta(n, k)$ if $I_n = 1$ where $\delta(n, k)$ is the Kronecker symbol. Otherwise, we have

$$N_n^{(k)} = \begin{cases} \delta(n,k) & \text{with probability } p \\ N_{L_n}^{(k)} + N_{R_n}^{(k)} & \text{with probability } 1-p, \end{cases} \tag{2.4}$$

where $N_{L_n}^{(k)}$ and $N_{R_n}^{(k)}$ denote two independent copies of the same process. The recursion starts for $n \geq k$. In addition, we have $N_{k+1}^{(k)} = 0$ for all $k \geq 2$. For $k > 2$, $N_k^{(k)}$ is a Bernoulli random variable of parameter $p + 2(1-p)/(k-1)$. Finally, $N_2^{(2)} = 1$.

Our third quantity of interest is the size, S_n, of the group which contains the leftmost individual in the hidden tree. We call this size-biased group a typical group because the individuals are exchangeable. The recursion for S_n is as follows. We have $S_n = n$ if $I_n = 1$. If $I_n > 1$, we have

$$S_n = \begin{cases} n & \text{with probability } p \\ S_{L_n} & \text{with probability } 1-p\,, \end{cases} \tag{2.5}$$

and $S_2 = S_3 = 2$. In Fig. 2, we have $S_{12} = 4$.

The connection between random trees and recursive structures have also been exploited many times in theoretical computer science (Knuth 1973). For example, this connection is the basis for the analysis of divide-and-conquer algorithms such as the quicksort algorithm (Rosler 1991; Hwang and Neininger 2002). Regarding coalescent trees, the connection has been used to derive the distribution of minimal clade sizes and other statistical measures of tree shape (Blum and François 2005a,b).

Note that other tree models share the same topology: the Yule tree (Yule 1924; Harding 1971), and the critical branching process (Popovic 2005; François and Mioland 2007). The results presented afterwards are also valid for these topologies.

3 Results

In this section, we present our results for the main quantities of interest: the number of groups, the group size spectrum and the size of a typical group. The results describe the limiting behaviours of these quantities when the sample size n grows to infinity. Our first result deals with the expected number of groups.

Theorem 3.1 *Let $0 \leq p < 1$. Denote $q = 1-p$. Let N_n be defined as in Eq. (2.3), and $e_n = E[N_n]$. The expected number of groups e_n satisfies the following properties. For $p < 1/2$, we have*

$$e_n \sim c(p) n^{1-2p}, \quad n \to \infty,$$

where

$$c(p) = \frac{e^{-2q}}{(1-2p)\Gamma(2q)} \left(1 + 2q^2 J(p)\right), \tag{3.1}$$

and

$$J(p) = \int_0^1 (1-y)^{1-2p} y^2 e^{2qy} dy.$$

When $p = 1/2$, we have

$$e_n \sim \frac{\log(n)}{2}, \quad n \to \infty.$$

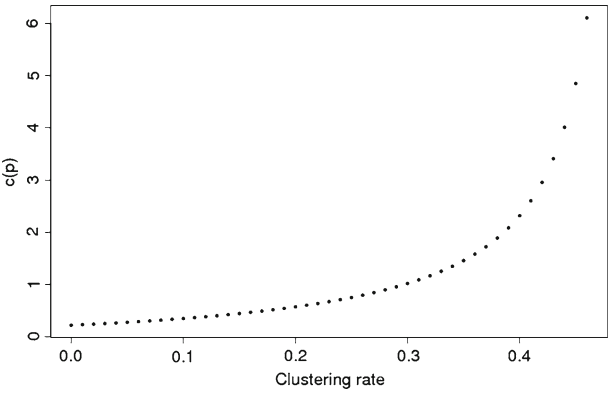

Fig. 3 Plot of $c(p)$ for $p = 0$ to $p = 0.5$ obtained from numerical integration. The exact value for $p = 0$ is $(1 - e^{-2})/4$

For $p > 1/2$, we have

$$e_n \sim \frac{p}{2p-1}, \quad n \to \infty.$$

Remark 1 In particular, Theorem 3.1 states that for $p = 0$, which corresponds to the neutral model, we have

$$e_n \sim \frac{1 - e^{-2}}{4} n,$$

when n grows to infinity. In a sample of size $n = 100$, one expects approximatively $N_{100} \approx 21 - 22$ groups under the neutral model.

Figure 3 shows the values of the term $c(p)$, obtained from the numerical integration of Eq. (3.1). The curve starts from $c(0) \approx 0.21$ and grows to infinity as p converges to 0.5.

Remark 2 We investigated the accuracy of the asymptotic equivalent of the expected number of groups for a sample size of $n = 100$ individuals. The average number of groups was computed numerically using the recursive equation given in Lemma 4.7 of Sect. 4. As shown in Fig. 4, the approximation of e_n is accurate for values of the clustering rate lower than ≈ 0.3 or greater than ≈ 0.7. The double phase transition at $p = 0.5$ may be responsible for the very high values of the relative error when p approaches 0.5. In contrast the value computed for $p = 0.5$ leads to a very good approximation.

Next, we present results for the group size spectrum, i.e. the collection of all $(N_n^{(k)})$, $2 \leq k \leq n$, where $N_n^{(k)}$ denote the number of groups of size k in a sample of n individuals.

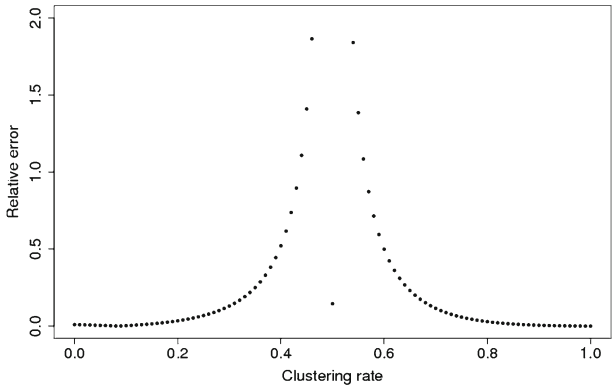

Fig. 4 Relative error when approximating the expected number of groups by its asymptotic equivalent for a sample size of $n = 100$ individuals

Theorem 3.2 *Let* $0 \leq p < 1$. *Denote* $q = 1 - p$. *Let* $N_n^{(k)}$ *be defined as in Eq.* (2.4), *and* $e_n^{(k)} = E[N_n^{(k)}]$. *For* $2 \leq k \leq n-2$, *the expected number of groups of size* k *satisfies*

$$e_n^{(k)} \sim d(k, p)(n-k)^{1-2p}, \quad n \to \infty, \tag{3.2}$$

where

$$d(k, p) = \frac{2qe^{-2q}(2 + (k-3)p)}{(k-1)\Gamma(2q)} J(k, p),$$

and

$$J(k, p) = \int_0^1 y^k (1-y)^{1-2p} e^{2qy} dy.$$

Remark Consider $f_n(k, p)$ the frequency spectrum of group size, defined as

$$f_n(k, p) = \frac{k e_n^{(k)}}{n}, \quad 2 \leq k \leq n.$$

The $f_n(k, p)$'s represent the frequency of groups of size k in a sample of size n. It can be computed numerically from Eq. (4.8) in Sect. 4.

Finally, we turn to our last statistic defined as the size of the clade which contains the left-most tip of the underlying tree, S_n.

Below, we provide a result describing the probability distribution of S_n.

Theorem 3.3 *Let $0 \leq p < 1$. Denote $q = 1 - p$. Let S_n be defined as in Eq. (2.5) and $p_n(x) = \text{Prob}(S_n = x)$, $2 \leq x \leq n$. For $x > 2$ we have*

$$p_n(x) \sim \frac{(p(x-3)+2)qe^{-q}}{(x-1)\Gamma(q)} I_p(x) n^{-p}, \quad n \to \infty, \tag{3.3}$$

where

$$I_p(x) = \int_0^1 y^x e^{qy} (1-y)^{-p} dy.$$

For $x = 2$ we have

$$p_n(2) \sim \frac{qe^{-q}}{\Gamma(q)} I_p(2) n^{-p}, \quad n \to \infty. \tag{3.4}$$

Remark 1 In the neutral model ($p = 0$), Theorem 3.3 shows that for large values of n, the distribution of the size of a typical group is defined by

$$p_n(2) \sim \frac{e-2}{e}, \quad n \to \infty, \tag{3.5}$$

and

$$p_n(x) \sim (-1)^{x+1} 2x! \frac{e^{-1} - e(x)}{x-1}, \quad x \geq 3, \quad n \to \infty, \tag{3.6}$$

where we denote

$$e(x) = \sum_{k=2}^{x} \frac{(-1)^k}{k!}.$$

When x grows to infinity it is not difficult to see that the right-hand side in formula (3.6) is equivalent to a power-law distribution of exponent 2

$$p_n(x) \sim \frac{2}{(x-1)(x+1)}, \quad x \to \infty. \tag{3.7}$$

Remark 2 We studied the quality of the approximation of the probability distribution of S_n for a sample of $n = 100$ individuals. In Fig. 5, the fit is almost perfect for a clustering rate of $p = 0.25$. For $p = 0$ the fit is almost perfect for small groups ($x \leq 15$). However, for larger values of x, the asymptotic approximation provides less accurate results.

Our last result concerns the expected values of the size of a typical group.

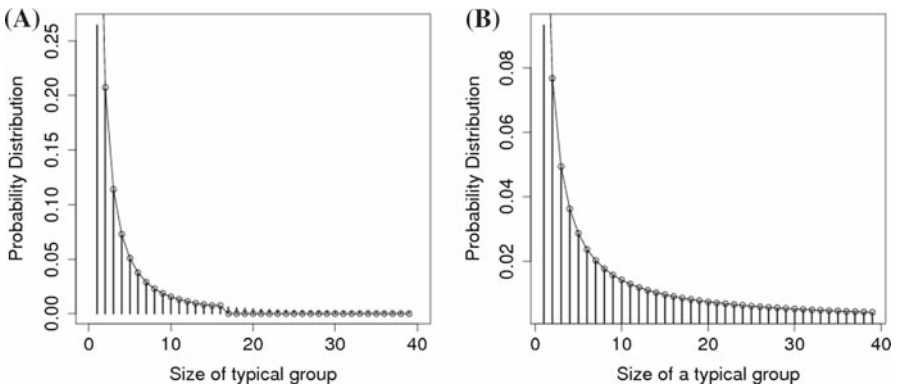

Fig. 5 Exact probability distribution of the size of a typical group for $n = 100$ and **a** $p = 0$ and **b** $p = 0.25$ (black bars) and its equivalent for large n as stated in Eq. 3.3 (*linked points*). We can see that for $n = 100$ the fit with the power-law distributed asymptotic is almost perfect as the asymptotic and the real distribution are roughly identicals

Theorem 3.4 *Let S_n be defined as in Eq. (2.5) and $s_n = E[S_n]$. Under the neutral model ($p = 0$), the expected size of a typical group satisfies*

$$s_n \sim 2\log(n), \quad n \to \infty. \tag{3.8}$$

Under the extra-clustering model ($0 < p < 1$), we have

$$s_n \sim \frac{2p}{1+p} n, \quad n \to \infty. \tag{3.9}$$

Remark 1 Interestingly, the mean size of a typical group undergoes a phase transition at $p = 0$. It means that the neutral model ($p = 0$) has small components, with a typical group size growing like $\log n$. In contrast, when $p > 0$, the model allows large extra-communities, and the typical group size is of the same order as the number of individuals.

Remark 2 We studied the numerical approximation of the expected size s_n. The expected group size was computed numerically thanks to Lemma 4.4. Approximating the average typical group size by its limit value for $n = 100$ leads to good results for $p \geq 0.3$, but it leads to important errors for small values of p (Fig. 6). This may be due to the phase transition observed at $p = 0$.

Finally, we focus on the higher moments of the size of a typical group.

Proposition 3.1 *Let S_n be defined as in Eq. (2.5) and assume that $p = 0$. Let us denote $s_n^{(k)} = E[S_n^k]$ the moment of order k of S_n. For $k > 1$, we have*

$$s_n^{(k)} \sim \frac{2k}{k-1} n^{k-1}, \quad n \to \infty. \tag{3.10}$$

In particular, for $k = 2$, we obtain $Var[S_n] \sim 4n$.

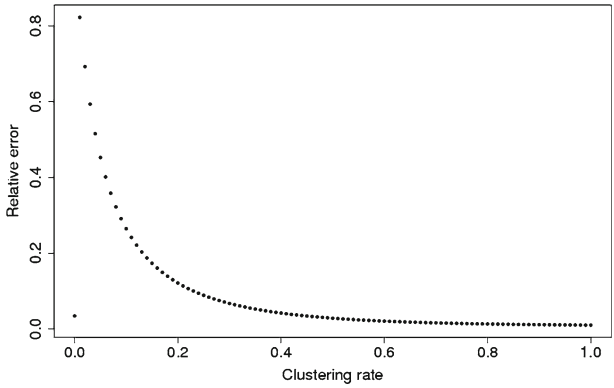

Fig. 6 Relative error when approximating the expected size of a typical group by its asymptotic equivalent for a sample size of $n = 100$ individuals

Data analysis To illustrate our results, we present a short data analysis using data collected by Archie et al. (2006) for wild African elephants. This data set was also briefly described in Durand et al. (2007). Archie et al. observed $n = 304$ African elephants (*Loxondonta africana*) living in 45 groups of related individuals. The size spectrum $(N_{304}^{(k)})$, $2 \leq k \leq 304$, is given by

$$(N_{304}^{k})_{2 \leq k \leq 304} = (4, 5, 5, 9, 4, 4, 1, 1, 4, 2, 3, 0, 0, 1, 1, 0, 1, 0, 0, \ldots).$$

In Durand et al. (2007), we tested the agreement of the elephant group pattern with the neutral model using the likelihood-ratio statistic $\text{Prob}(N_n | p = 0)/\text{Prob}(N_n | \hat{p})$. We were not able to reject the neutral model with this test at a significance level of 5% (p-value $= 0.063$). A test based on the full group size spectrum is more powerful than a test based on the N_n statistic. We performed a log-log linear regression of the elephant group size spectrum on group sizes. The estimated coefficient of the linear regression was $\alpha = 0.83$ (p-value $= 0.005$). We computed the coefficient of the log-log linear regression for the frequency spectrum under the neutral model ($p = 0$) for $n = 304$ using Eq. (2.4) and we found $\alpha_0 = 1.68$. We checked that the 2.5% and 97.5% quantiles of the distribution of estimated exponents were equal to $q_l = 1.01$ and $q_u = 1.90$ under the neutral model. The estimated value $\alpha = 0.83$ did not fall between those quantiles, which leads us to reject the neutral model for the elephant group pattern.

The fact that the above coefficient was higher under the neutral model than the coefficient estimated for the elephant group size distribution suggests that the extra-clustering model may better explain the data. In fact, using the maximum-likelihood estimate of the clustering rate (Durand et al. 2007), we found that the elephant group size distribution was best explained by the extra-clustering model with $p = 0.08$. The log-log coefficient was equal to $\alpha = 0.88$ under the extra-clustering model with $p = 0.08$ and $n = 304$, which is much closer to the estimate based on the elephant data ($\alpha = 0.83$).

4 Proofs

In this section we provide detailed proofs for the results stated in Sect. 3. First, we present a general strategy for the resolution of a particular recursive equation, based on a theorem by Darboux (Theorem 4.12 in Sedgewick and Flajolet (1996)). All the proofs are based on this method.

4.1 General method

Let f_n denote a sequence of real values, and suppose that f_n satisfies the following equation

$$(n+1)f_n = nf_{n-1} + af_{n-2} + b, \tag{4.1}$$

with a and b being real numbers, $a > 0$ and $a \neq 1$. The first terms f_0 and f_1 can take arbitrary values. One can easily check that $-b/(a-1)$ is a fixed point of Eq. (4.1). We then introduce the auxiliary sequence $v_n = f_n + b/(a-1)$ which satisfies the following equation

$$(n+1)v_n = nv_{n-1} + av_{n-2}, \quad n \geq 2. \tag{4.2}$$

Now we denote $h(x) = \sum_{k=2}^{\infty} v_k x^k$ the ordinary generating function of the sequence (v_n) and we denote r its radius of convergence. We can check that $h(x)$ satisfies the following ordinary differential equation

$$(1 - x - ax^2)y = (x^2 - x)y' + (2v_1 + av_0)x^2 + av_1x^3. \tag{4.3}$$

The unique solution of Eq. (4.3) with initial condition $y(0) = 0$ is given by

$$\hat{h}(x) = \frac{e^{-ax}}{x(1-x)^a} \int_0^x (2v_1 + av_0 + av_1y)y^2(1-y)^{a-1}e^{ay}dy, \quad x \geq 0. \tag{4.4}$$

The problem to determine the asymptotic behaviour of $\hat{h}(x)$ can be solved by using the following result.

Theorem 4.1 [Darboux (see Sedgewick and Flajolet 1996)] *Let $g(z)$ be a power series with a radius of convergence strictly greater than 1. Suppose that $g(1) \neq 0$. Then for any real number $a \notin \{0, -1, -2, \ldots\}$, we have*

$$[z^n]\frac{g(z)}{(1-z)^a} \sim g(1)\binom{n+a-1}{n}, \quad n \to \infty, \tag{4.5}$$

where $\binom{m}{n}$ denotes the binomial coefficient. $[z^n]\frac{g(z)}{(1-z)^a}$ denotes the nth coefficient of the power series $\frac{g(z)}{(1-z)^a}$.

In our general strategy, we apply Theorem 4.1 to the function $g(x) = \hat{h}(x)(1-x)^a$. First, we show that for $a > 0$, $g(x)$ satifies the hypotheses of Theorem 4.1. Let us denote $W(x) = \int_0^x (2v_1 + av_0 + av_1 y)y^2(1-y)^{a-1}e^{ay}dy$. We integrate $W(x)$ by parts, and we find that

$$W(x) = -\left((4v_1 + 2av_0)x + (5av_1 + a^2v_0)x^2 + a^2v_1x^3\right)(1-x)^a\frac{e^{ax}}{a} + \frac{1}{a}\int_0^x e^{ay}(1-y)^a((4v_1 + 2av_0)y + (5av_1 + a^2v_0)y^2 + a^2v_1y^3)dy$$

Because we have $a > 0$, $e^{ay}(1-y)^a((4v_1 + 2av_0)y + (5av_1 + a^2v_0)y^2 + a^2v_1y^3)$ can be written as a power series with a radius of convergence equal to infinity. Let us denote δ_n its coefficients for $n \geq 0$. We have

$$\begin{aligned} W(x) &= -\left((4v_1 + 2av_0)x + (5av_1 + a^2v_0)x^2 + a^2v_1x^3\right)(1-x)^a\frac{e^{ax}}{a} \\ &\quad + \frac{1}{a}\int_0^x \sum_{n=0}^{\infty}\delta_n y^n dy \\ &= -\left((4v_1 + 2av_0)x + (5av_1 + a^2v_0)x^2 + a^2v_1x^3\right)(1-x)^a\frac{e^{ax}}{a} \\ &\quad + \frac{1}{a}\sum_{n=0}^{\infty}\frac{\delta_n}{n+1}x^{n+1}. \end{aligned}$$

Because $g(x) = \frac{e^{-ax}}{x}W(x)$, we have

$$g(x) = \frac{\left((4v_1 + 2av_0) + (5av_1 + a^2v_0)x + a^2v_1x^2\right)(1-x)^a}{a} + \frac{e^{-ax}}{a}\sum_{n=0}^{\infty}\frac{\delta_n}{n+1}x^n.$$

Thus we can conclude that $g(x)$ is a power series with a radius of convergence equal to infinity for $a > 0$. Then, we can check that $g(1) \neq 0$, because at least one δ_n is different from 0. We can conclude that $g(x)$ satisfies the hypotheses of theorem (4.1). Finally, remark that, because $\frac{g(z)}{(1-z)^\alpha} = \hat{h}(z)$, we have

$$v_n = [z^n]\frac{g(z)}{(1-z)^\alpha}.$$

We can then conclude that

$$v_n \sim e^{-a}\int_0^1 (2v_1 + av_0 + av_1 y)(1-y)^{a-1}y^2e^{ay}dy\frac{n^{a-1}}{\Gamma(a)}, \quad n \to \infty. \tag{4.6}$$

To obtain f_n, we need to discuss four cases depending on the value of a, v_0 and v_1. If $v_n \sim b/(a-1)$, we cannot give conclusions regarding the asymptotic equivalent of f_n because $f_n = v_n - b/(a-1)$, but this case never happens in our equations. If $v_n \sim c$, c being constant and not equal to $\frac{b}{a-1}$, we have

$$f_n \sim c - \frac{b}{a-1}, \quad n \to \infty.$$

If v_n grows to infinity when n grows to infinity, we have $f_n \sim v_n$ and we can use Eq. (4.6). If v_n tends to 0 when n grows to infinity, we have $f_n \sim b/(1-a)$.

4.2 Number of groups

In this section we prove Theorem 3.1.

Lemma 4.1 *Let $n \geq 2$. Let $0 \leq p < 1$ and N_n be defined as in Eq. (2.3). The expected number of groups e_n satisfies the following equation*

$$(n+1)e_{n+2} = ne_{n+1} + 2(1-p)e_n + p, \tag{4.7}$$

with initial values $e_2 = e_3 = 1$.

Proof We obtain Lemma 4.1 from Eq. 2.3 by applying the law of total probability, remembering the fact that L_n is uniformly distributed over the set $\{1, \ldots, n-1\}$. □

Proof of Theorem 3.1 Equation (4.7) is of the same form as Eq. (4.1). We can apply the method described in Sect. 4.1 for $p \neq 0.5$ and $p < 1$. Let $v_n = e_{n+2} - p/(2p-1)$. Equation (4.6) tells us that

$$v_n \sim e^{-a} \int_0^1 (2v_1 + av_0 + av_1 y)(1-y)^{a-1} y^2 e^{ay} dy \frac{n^{a-1}}{\Gamma(a)}, \quad n \to \infty,$$

with $v_0 = 1 - p/(2p-1)$ and $v_1 = 1 - p/(2p-1)$ and $a = 2(1-p)$. Then, we need to discuss to cases. If $p < 0.5$, then v_n grows to infinity when n grows to infinity, and we have

$$e_n \sim v_n, \quad n \to \infty.$$

If $p > 0.5$, then $v_n \to 0$ when n grows to infinity, and we have $e_n \sim p/(2p-1)$. In both cases we recover the result stated in Theorem 3.1.

Finally, if $p = 0.5$, let $u_n = e_{n+1} - e_n$ for $n \geq 2$. By re-arranging equation (4.7) we can check that $nu_{n+1} = 1/2 - u_n$ which leads to $u_n \sim 1/(2n)$ when n grows to infinity. It is then obvious that $e_n \sim \log(n)/2$. □

4.3 Group size spectrum

Here we prove Theorem 3.2.

Lemma 4.2 *Let $n \geq 4$. Let $0 \leq p < 1$. Let $q = 1 - p$. For $2 \leq k \leq n-2$, let $N_n^{(k)}$ be defined as in Eq. (2.4) and $e_n^{(k)} = E[N_n^{(k)}]$. We have*

$$e_n^{(k)} = \frac{n-2}{n-1} e_{n-1}^{(k)} + \frac{2q}{n-1} e_{n-2}^{(k)}, \tag{4.8}$$

with initial values $e_k^{(k)} = (2 + (k-3)p)/(k-1)$ and $e_{k+1}^{(k)} = 0$.

Proof We obtain Lemma 4.2 from Eq. (2.4) by applying the law of total probability, remembering the fact that L_n is uniformly distributed over the set $\{1, \ldots, n-1\}$. □

Proof of Theorem 3.2 We introduce $k_n = e_{n+k}^{(k)}$ which satisfies the following equation

$$(n+k-1)k_n = (n+k-2)k_{n-1} + 2(1-p)k_{n-2}, \quad n \geq 4, \tag{4.9}$$

which is similar to Eq. (4.2). We can then apply the methodology described in Sect. 4.1 to the sequence (k_n). The generating function $h_k(x)$ of k_n satisfies the following differential equation

$$(k-1-(k-1)x-2(1-p)x^2)y = (x^2-x)y' + 2(1-p)\frac{2+(k-3)p}{k-1}x^2. \tag{4.10}$$

The solution of this equation is given by

$$h_k(x) = \frac{x^{1-k}(x-1)^{2-2p}e^{-2(1-p)x}}{k-1} \times \int_0^x 2(p-1)y^k(y-1)^{1-2p}(2+(k-3)p)e^{2(1-p)y}dy, \quad x \geq 0.$$

We can conclude the proof by applying Theorem 4.1 with $a = 2(1-p)$ which leads to the result stated in Theorem 3.2. □

4.4 Size of a typical group

This section deals with the results concerning the size of a typical group, S_n. First, we prove Theorem 3.3 on the probability distribution of the size of a typical group. We let $p_n(x) = \mathrm{P}(S_n = x)$ denote the probability distribution of the size of a typical group, $2 \leq x \leq n$.

Lemma 4.3 *Let $n \geq 2$. Let $0 \leq p < 1$. Let $2 \leq x \leq n-1$. Let S_n be defined as in Eq. (2.5) and $p_n(x) = \text{Prob}(S_n = x)$. The probability $p_n(x)$ satisfies the following equation*

$$np_{n+1}(x) = (n-1)p_n(x) + (1-p)p_{n-1}(x). \tag{4.11}$$

Proof We obtain Lemma 4.3 from Eq. (2.5) by applying the law of total probability, remembering the fact that L_n is uniformly distributed over the set $\{1, \ldots, n-1\}$. □

Proof of Theorem 3.3 For a given $x \in [2, n-1]$, Eq. (4.11) has the same form as Eq. (4.1). Thus we can apply the method described in Sect. 4.1. We need to distinguish two cases for x because the initial values of p_n depend on it. If $x = 2$, the initial values of the sequence p_n are $p_2 = 1$ and $p_3 = 0$. We apply the result stated in Eq. (4.6) and we find the expected result as stated in Theorem 3.3. If $x > 2$, the initial values of $p_n(x)$ are $p_x = 2(1-p)/(x-1) + p$ and $p_{x+1} = 0$. Denoting $f_n = p_{n+x}(x)$ we have $f_0 = 2(1-p)/(x-1) + p$ and $f_1 = 0$. f_n satisfies the following recursive equation

$$(n+x-1)f_n = (n+x-2)f_{n-1} + (1-p)f_{n-2}, \quad x \geq 2. \tag{4.12}$$

The end of the proof of Theorem 3.3 follows the same lines as the proof of Theorem 3.2. □

Now we prove Theorem 3.4 on the expected size of a typical group, $s_n = E[S_n]$.

Lemma 4.4 *Let $n \geq 2$. Let $0 \leq p < 1$. Let S_n be defined as in Eq. (2.5) and $s_n = E[S_n]$. The expected size of a typical group satisfies the following recursive equation*

$$s_{n+1} = \left(1 - \frac{1}{n}\right) s_n + \frac{1-p}{n} s_{n-1} + 2p + \frac{2q}{n}, \tag{4.13}$$

with initial values $s_2 = s_3 = 1$.

Proof We obtain Lemma 4.4 from Eq. (2.5) by applying the law of total probability, remembering the fact that L_n is uniformly distributed over the set $\{1, \ldots, n-1\}$. □

Proof of Theorem 3.3 In the case $p = 0$, we introduce the sequence $d_n = s_{n+1} - s_n$, which satisfies the following recursive equation

$$d_n = -\frac{1}{n} d_{n-1} + \frac{2}{n}, \quad n \geq 3, \tag{4.14}$$

with $d_2 = 0$. It is then obvious that $d_n \sim \frac{2}{n}$ when n grows to infinity. Then, by summing over all values of n, we find that $s_n \sim 2\log(n)$. In the case $p > 0$, we denote

$a_n = 2pn/(1+p)$ and we let $w_n = s_n - a_n$. Remark that a_n satisfies the following equation

$$a_{n+1} = \left(1 - \frac{1}{n}\right) a_n + \frac{q}{n} a_{n-1} + 2p + \frac{2pq}{n(1+p)}, \quad n \geq 3, \tag{4.15}$$

where $q = 1 - p$ and $a_2 = 4p/(1+p)$ and $a_3 = 6p/(1+p)$. By rearranging the terms in Eq. (4.11), we find that

$$n w_{n+1} = (n-1) w_n + q w_{n-1} + \left(2q - \frac{2pq}{1+p}\right), \quad n \geq 3, \tag{4.16}$$

where $w_2 = 1 - 4p/(1+p)$ and $w_3 = 1 - 6p/(1+p)$. We can apply the same method as described in Sect. 4.1 to study the sequence w_n. In any of the cases listed for Eq. 4.6 (w_n converging to a constant, or $w_n \sim n^{-p}$ when n grows to infinity), we find that $s_n \sim a_n$ because a_n grows to infinity as n grows to infinity. □

Finally we prove Proposition 3.1 about the higher moments of the size of a typical group. In this paragraph we fix $p = 0$. Let us denote $\phi_n(t) = E[e^{tS_n}]$ for $t \geq 0$ the moment generating function of S_n. We can check, thanks to the law of total probability, that

$$\phi_{n+1}(t) = \left(1 - \frac{1}{n}\right) \phi_n(t) + \phi_{n-1}(t) + \frac{2}{n} e^{tn}(e^t - 1) \qquad n \geq 3, \quad t \geq 0. \tag{4.17}$$

Now let us denote $f_n(t) = 2e^{tn}(e^t - 1)/n$ for $t \geq 0$. We can check by using a direct recursion that for $k \geq 1$, we have

$$f_n^{(k)}(0) = \frac{2(n+1)^{k-1}}{n} + 2((n+1)^{k-1} - n^{k-1}), \quad n \geq 3. \tag{4.18}$$

The kth moment of S_n, denoted $s_n^{(k)}$, equals $\phi_n^{(k)}(0)$. Thus it solves the following recursive equation

$$s_{n+1}^{(k)} = \left(1 - \frac{1}{n}\right) s_n^{(k)} + \frac{s_{n-1}^{(k)}}{n} + f_n^{(k)}(0), \quad n \geq 3, \tag{4.19}$$

where $s_2^{(k)} = s_3^{(k)} = 1$. Let us denote $z_n^{(k)} = s_{n+1}^{(k)} - s_n^{(k)}$. This quantity satisfies the following recursion for $k \geq 1$

$$z_n^{(k)} = -\frac{1}{n} z_{n-1}^{(k)} + f_n^{(k)}(0), \quad n \geq 3, \tag{4.20}$$

with $z_2 = 0$. We can check, using Newton's binomial formula, that $f_n^{(k)}(0) \sim 2kn^{k-2}$ for large n. So we have $z_n^{(k)} + \frac{1}{n} z_{n-1}^{(k)} = f_n^{(k)}(0) \sim 2kn^{k-2}$ and then $z_n^{(k)} \sim 2kn^{k-2}$.

Finally, for $k \geq 1$, we obtain

$$s_n^{(k)} = \sum_{i=2}^{n} z_i^{(k)} + s_2^{(k)} \sim \frac{2k}{k-1} n^{k-1}, \quad n \to \infty. \tag{4.21}$$

□

The same proof scheme can also be used to prove Theorem 3.4 for the neutral model.

Acknowledgments We are grateful to Michael G. B. Blum for inspiring discussions and many usefull comments on a previous draft of the manuscript. We thank two anonymous referees whose constructive remarks and suggestions significantly improved the original manuscript. This work was supported by grants from the Agence Nationale de la Recherche, project MAEV "Modèles Aléatoires pour l'Evolution du Vivant".

References

Archie EA, Moss CJ, Alberts SC (2006) The ties that bind: genetic relatedness predicts the fission and fusion of social groups in wild african elephants. Proc R Soc B: Biol Sci 273:513–522

Aldous DJ (1996) Probability distributions on cladograms. Random discrete structures. In: Aldous DJ, Pemantle R (eds) IMA volumes math appl, vol 76. Springer, Heidelberg, pp 1–18

Bonabeau E, Dagorn L, Fréon P (1999) Scaling in animal group size distribution. Proc Natl Academy Sci 96:4472–4477

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

Blum MGB, François O (2005b) On statistical tests of phylogenetic imbalance: the sackin and other indices revisited. Math Biosci 195:141–153

Dawkins R (1989) The selfish gene. Oxford University Press, London

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

François O, Mioland C (2007) Gaussian approximation for phylogenetic branch length statistics under stochastic models of biodiversity. Math Biosci 209:108–123

Gueron S, Levin SA (1995) The dynamics of group formation. Math Biosci 128:243–264

Giraldeau LA, Caraco T (1993) Genetic relatedness and group size in an aggregation economy. Evol Ecol 7:429–438

Hamilton IM (2000) Recruiters and joiners: Using optimal skew theory to predict group size and the division of ressources within groups of social foragers. Am Nat 155:684–695

Hamilton WD (1964) The evolution of social behavior. J Theor Biol 7:1–52

Hein J, Schierup MH, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, USA

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

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 3:44–77

Knuth DE (1973) The art of computer programming. Fundamental algorithms, vol I, 2nd edn. Addison-Wesley, Reading

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

Popovic L (2005) Asymptotic genealogy of a critical branching process. Ann Appl Probab 14:2120–2148

Rubenstein DI (1978) On predation, competition, and the advantages of group living. Perspect Ethol 3:205–232

Rosler U (1991) A limit theorem for quicksort. Theor Appl Inform 25:85–100

Sedgewick R, Flajolet P (1996) An introduction to the analysis of algorithms. Addison-Wesley Longman Publishing Co., Inc., Boston

Takayasu H (1989) Steady-state distribution of generalized aggregation system with injection. Phys Rev Lett 63:2563–2565

Tavaré S (2004) Ancestral inference in population genetics. In: Picard J (ed) Lectures on probability theory and statistics. Ecole d'Etés de Probabilité de Saint-Flour XXXI, 2001. Lecture notes in mathematics, vol 1837. Springer, New York, pp 1–188

Yokoyama S, Felsenstein J (1978) A model of kin selection for an altruistic trait considered as a quantitative character. Proc Natl Academy Sci 75:420–422

Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. Philos Trans R Soc Lon Ser B 213:21–87