Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 249 (2007) 262–270
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

# Prediction of group patterns in social mammals based on a coalescent model

Eric Durand[a,*], Michael G.B. Blum[a,b], Olivier François[a]

[a]TIMC, University Joseph Fourier, INP Grenoble, CNRS, Grenoble, France
[b]Department of Human Genetics, University of Michigan, Ann Arbor, USA

Received 20 January 2007; received in revised form 29 May 2007; accepted 6 July 2007
Available online 19 July 2007

## Abstract

This study describes a statistical model which assumes that mammal group patterns match with groups of genetic relatives. Given a fixed sample size, recursive algorithms for the exact computation of the probability distribution of the number of groups are provided. The recursive algorithms are then incorporated into a statistical likelihood framework which can be used to detect and quantify departure from the null-model by estimating a clustering parameter. The test is then applied to ecological data from social herbivores and carnivores. Our findings support the hypothesis that genetic relatedness is likely to predict group patterns when large mammals have few or no predators.

*Keywords:* Group patterns; Genetic relatedness; Coalescence; Statistical tests; Social mammals

## 1. Introduction

Group formation is a widespread phenomenon throughout the social mammals, and the problem of animal grouping is one of the most fundamental ones in biology. The nonuniform aggregation patterns of organisms have both ecological and evolutionary significance, and the tendency to aggregate is under strong evolutionary control (Rubenstein, 1978). Individuals might derive various kinds of benefit from living in groups. Many authors have proposed the idea that social carnivores live in groups because group hunting facilitates their acquisition of large prey (Mech, 1981; Nudds, 1978; Pulliam and Caraco, 1978). Parallel to this argument, the most widely studied advantage of large herbivore grouping is lowered predation risk (Hamilton, 1971; Pulliam, 1973; Bertram, 1978; Inman and Krebs, 1987). There might, however, exist several other advantages to group living as animals may huddle together to keep warm, learn from one another about good feeding sites, or gain access to mates.

Mathematical models of group formation have often restricted attention to relating observed grouping patterns to the processes of aggregation (fusion) and splitting (fission) (Gueron and Levin, 1995). Perhaps one of the most representative among the animal aggregation models is the one proposed in Bonabeau and Dagorn (1995) which was inspired by a model of particle aggregation (Takayasu, 1989). Other examples of fusion/fission models were also proposed by Bonabeau et al. (1999) and Niwa (2003). These models rely on the basic assumption that groups are randomly moving units in a spatial domain, and whenever two groups meet, they aggregate. Each group may also split in two subgroups with a given probability. Although these models are very simple, they are flexible enough to capture the power-law distributions of large animal group size observed in many surveys (Bonabeau et al., 1999).

Separately a process called *kin-selection* was suggested by Hamilton (1964) as a mechanism for the evolution of altruistic behavior, and as one of the mechanisms that may explain group formation in social animal species (Dawkins, 1989; Foster et al., 2006). Since identical copies of genes

[*]Corresponding author. TIMB Department of Mathematical Biology, TIMC UMR CNRS 5525, Fac. Méd., Grenoble Universités, F38706 La Tronche, France. Tel.: +33 456 520 070; fax: +33 456 520 055.
*E-mail address:* eric.durand@imag.fr (E. Durand).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.07.012

may be carried by relatives, a gene that favors altruism may become successful provided that the reproductive benefit gained by the recipient of the altruistic act compares favorably to the reproductive cost to the individual performing the act. In such a comparison, the reproductive benefit gained by the recipient is weighted by the genetic relatedness of the two animals, defined as the percentage of genes that they share by common descent. This theory suggests that genetic relatedness contributes to group formation and cohesion as a major actor.

Related model-based approaches to animal grouping usually include game-theoretic aspects, which, loosely speaking, put cooperation and competition into balance to determine group size. For example, aggregation may be explained from the recruiters and joiners point of view (Hamilton, 2000). In such models group size may be determined by optimizing fitness functions which express the advantages and flaws for individuals to aggregate. Animals may join the group as long as it improves their access to resource; they leave otherwise. Models in this category, however, contain many parameters, and are not primarily intended to perform statistical testing. While some models may include genetic relatedness (e.g., Giraldeau and Caraco, 1993), the full set of parameters may hardly be inferred from the ecological data, and this limitation prevents their use in data analyses.

Ecological studies of animal group size usually report direct observations of a few summary statistics like species census (or sample) size $n$, the number of groups $N_n$ in the sample and the average group size $n/N_n$. A challenging objective is to use these data to formulate general hypotheses about the way by which evolution shapes group patterns. Formulating and testing such hypotheses require a quantitative theory which may include within-group genetic relatedness as a basic principle.

Yet a similar quantitative theory has been developed in population genetics, where a stochastic process known as 'the coalescent' has played a central role since the 1980s (Nordborg, 2003). In its basic form, the coalescent is an approximation of the genealogy of a very simple evolutionary dynamical model—the Wright–Fisher model—which assumes random mating and selective neutrality. Statistical tests of selection typically attempt to reject this null-model based on the value of a particular summary statistic computed from genetic data (Tajima, 1989).

This study parallels the traditional coalescent approach to population genetics in order to devise a statistical model of mammal group patterns which incorporates genetic relatedness as the major factor explaining these patterns. Examples that motivate the theory include large carnivores (wolves and lions) and social herbivores (buffalos, gazelles, elephants). Wolves are pack-living animals with a complex social organization in which packs are primarily family groups (Mech, 1981). African lions live in prides in which females are usually related to one another and are group members for life (Schaller, 1972). As well feral cattle are organized in societies that reflect social structure with many levels of organization (Lazo, 1994). At the high level, animals may form stable social subgroups within a herd. These "subherds" are often collections of matrilineal groups (Reinhardt and Reinhardt, 1981; Lazo, 1994).

Social organization of wild African elephants is another remarkable example of matrilineal structure. Recently, Archie et al. (2006) have documented how genetic relatedness, and in particular mtDNA relatedness, predicts the organization of social groups in wild African elephants. In brief, social groups in African elephants consist of genetic units. Female elephants are matrilocal and remain in their native group. Such a striking example of matrilocal organization advocates for the development of models of group patterns that are based on genetic relatedness. Our objective consists of introducing such a model and especially comparing the prediction of this model to the data available for social mammals.

In this study, we describe a statistical model for group patterns based on a coalescent genealogy, and we introduce a new parameter which measures the degree of extra clustering in mammal populations compared to the predictions of a neutral model. Our model is based on the minimal assumption that mammals live in group of relatives, and we call this model *random or neutral aggregation*. Deviations from the model will indicate that selective pressures like benefit from group hunting or avoidance of predators may interplay with genetic relatedness to shape group patterns.

The probability distribution of the number of groups in a sample is then described, and used to devise new statistical tests of random aggregation. This approach is illustrated using published social mammal data extracted from the recent sociobiology and ecology literature.

## 2. Models

Statistical models of group formation usually rely on fission and fusion of randomly moving animals. Here we introduce a random model of group patterns relying on the sole assumption that genetic relatedness is higher within groups than between groups. The model may be viewed as a deliberate simplification of the actual process of animal grouping, built on the principle that departure from minimal assumptions is always easier to interpret than departure from complex null hypotheses (see Discussion). Because genetic relatedness can be computed on the basis of a neutral gene tree, the basic model is referred to as the *neutral* model. In order to capture and measure the degree at which a group pattern may deviate from the neutral model, we also introduce a model incorporating *extra clustering* which allows us to estimate a natural clustering parameter. At this stage, there is an important distinction that must be made between fission/fusion models which try to explain the clustering process itself, and the neutral model which focus on the resulting group patterns, ignoring the processes by which they evolve.

### 2.1. The neutral model

Consider $n$ individuals sampled from an arbitrary population. The following assumptions are the building blocks of the neutral model.

(a) Groups have size greater or equal than two, i.e., every sampled individual must be grouped with at least another individual in the sample.
(b) Groups result from a random clustering process which reflects the random genealogical relationships within the sample, and in which each individual is attached to its closest relatives.

Point (a) states that communities have size greater or equal than two which is a natural assumption. Every sampled individual must be grouped with at least another individual in the sample when the sample is large enough. In species with strong matriarchal organization, it may happen that males are solitary. For such species, only the female grouping structure may well be captured by the model.

The groups obtained by the above process derive from an underlying tree which may be thought of as a genealogy of the sample. The sample genealogy is approximated by a fully dichotomic tree motivated by the need to avoid more complex pedigrees. To connect the model with a definition of relatedness, it may be convenient to view the hidden tree as a gene tree. The tree has internal branches that link its internal nodes, and external branches that start from the tips and end at an internal node. The definition of genetic groups consists in moving backward along each external branch until an ancestor is met. The tip that corresponds to this external branch is then aggregated to all the descendants of the first ancestor encountered (see Fig. 1). Although the aggregation process does not account for sexual reproduction, we can restrict the model to mtDNA which is inherited maternally, and we may think of relatedness as being measured from an mtDNA tree in a natural way. In the African elephant example, Archie et al. (2006) have investigated to which extent genetic relatedness predicts the pattern of fusion–fission events. Their analysis brought several lines of evidence that genetic relatedness and the level of association between individuals are strongly correlated. First, all individuals that are part of core social groups share the same mtDNA haplotype. Second, a Mantel test clearly established that the average pairwise level of genetic relatedness within a group is a good predictor of animal association within that group.

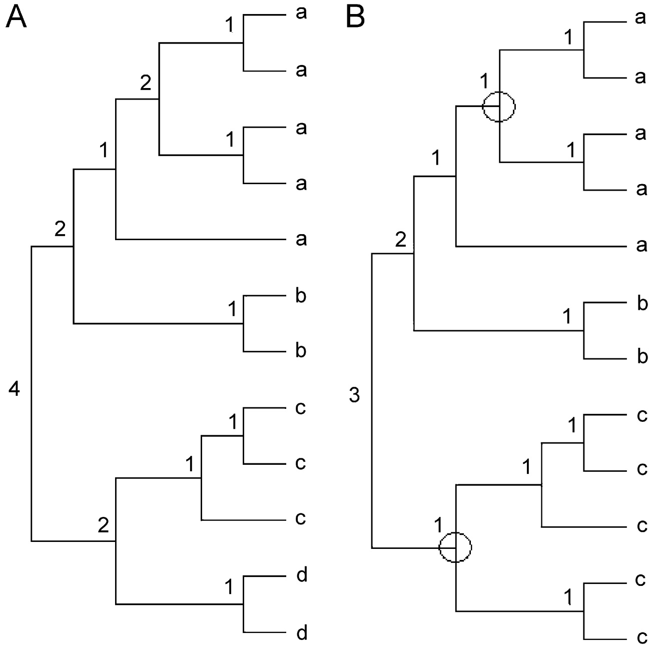

Fig. 1. Recursive computations of the number of groups, $N_n$ for $n = 12$. The letters at the tips stand for the community labels. (A) Neutral model: the sample has four groups denoted a–d. (B) Extra clustering: two extra-clustering events occur and are symbolized by circles. The sample has three groups denoted a–c.

### 2.2. Extra clustering

In order to account for the fact that genetic relatedness may not be the sole factor that contributes to shape group patterns, we consider an extension of the basic model that tolerates extra clustering without modifying the underlying tree model. In the extra-clustering model, groups may sometimes arise from the random coalescence of clusters created from the neutral process. We assume that the extra-clustering events occur during the construction process at rate $p$, called the *clustering rate* (See Fig. 1). More specifically, at each internal node of the binary tree all subgroups may be aggregated with probability $p$; otherwise the neutral rules are applied with probability $q = 1 - p$. Groups in social mammals sometimes consist of two or several subgroups often organized in a hierarchical way. In the wild African elephant example, Archie et al. (2006) noticed that the merging of core social groups to form larger groups occurs predominantly between individuals sharing the same mtDNA haplotype. The extra-clustering model incorporates this kind of properties which are partly missing in the neutral model. The parameter $p$ quantifies the amount of clustering present in the data which cannot be explained by genetic relatedness. Large $p$'s may indicate that external forces like predator avoidance or increased access to mates are involved in group formation. When estimated from the data, the parameter $p$ can provide an intuitive and useful measure of how much the data deviate from the neutral model.

## 3. The number of groups

The simplicity of the neutral and extra-clustering aggregation models allows exact theoretical predictions about the number of groups in a sample of mammals. Describing the probability distribution of this summary statistic under both models provides a mean to assess the significance of genetic relatedness as the main factor explaining group patterns.

### 3.1. A distributional recursion

The derivation of a probability distribution for the number of groups in a sample of size $n$ strongly relies on a recursive definition of this statistic. Such a definition uses mathematical properties of coalescent genealogies (Kingman, 1982). Starting with $n$ tips, these genealogies have the particular property that the size $L_n$ of the "left" sister clade at the basal split of the tree has a uniform distribution over the set $\{1, \ldots, n-1\}$ (Aldous, 2001)

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

and this property also holds within each subtree.

Regarding the number of groups $N_n$, we have $N_2 = N_3 = 1$. We can split the tree at the root so that two sister clades of sizes $L_n$ and $R_n = n - L_n$ are obtained, and then let $I_n = \min(L_n, R_n)$ be the minimum of $L_n$ and $R_n$. The number of groups can be involved in a set of recursive distributional equations as follows:

$$N_n = \begin{cases} 1 & \text{if } I_n = 1, \\ N_{L_n} + N^*_{R_n} & \text{otherwise,} \end{cases} \tag{2}$$

where $N^*_n$ denotes an independent copy of $N_n$. In this definition, the replicates of $L_n$ are recursively sampled from the uniform distribution over $\{1, \ldots, n-1\}$. The above set of equations also provides an efficient simulation algorithm for the number of groups of $N_n$ that avoids generating the trees themselves.

Sets of recursive distributional equations such as those described in Eq. (2) appear in theoretical computer science, and are natural in the analysis of divide-and-conquer algorithms (Rösler, 2001; Hwang and Neininger, 2002; Blum and François, 2005a). Using results obtained by Blum and François (2005b, p. 649), we can check that the "outdegree" of an arbitrary tip (i.e., the number of its closest relatives) in the neutral process has a power-law distribution with exponent $\alpha = 3$.

Turning to the model with extra clustering, the equations for $N_n$ change as follows. We now have $N_n = 1$ if $I_n = 1$, and otherwise,

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

Group patterns and the recursive computations of $N_n$ are illustrated in Fig. 1 where examples with $n = 12$ tips are displayed in the neutral and extra-clustering models.

### 3.2. Statistical tests and P-values

In order to perform statistical tests, we need to compute the probability distribution of the number of groups under the neutral and extra-clustering models. $P$-values can be directly deduced from this distribution when $N_n$ is used as a test statistic. Determining this distribution exactly is also useful in order to devise a more powerful likelihood-ratio (LR) test.

The recursive equations allow us to describe the probability distribution of $N_n$ by solving a triangular system. To see this, we let $\pi_n(x) = \text{Prob}(N_n = x)$ denote the probability distribution of $N_n$ for all integer $x \geqslant 1$. In the neutral model, we have $\pi_n(1) = 2/(n-1)$, and

$$\pi_n(x) = \frac{1}{n-1} \sum_{\ell=2}^{n-2} \sum_{y=1}^{x-1} \pi_\ell(y)\pi_{n-\ell}(x-y), \quad 1 \leqslant x \leqslant \lfloor n/2 \rfloor. \tag{3}$$

Examples of this distribution for $n = 200$ individuals are displayed in Fig. SM2.

Given $x$ groups in the sample, the one-sided $P$-value can be computed as $P = \text{Prob}(N_n \leqslant x)$. We further refer to the computation of this $P$-value as the $N_n$-test. Extra clustering at rate $p$ modifies the recursions for $N_n$ and the probability distribution of this statistic. Again we denote $\pi_n(x) = \text{Prob}(N_n = x)$ for all $x \geqslant 1$. Then the new distribution $\pi_n$ can be calculated using triangular induction as follows:

$$\pi_n(1) = p + \frac{2q}{n-1} \tag{4}$$

and

$$\pi_n(x) = \frac{q}{n-1} \sum_{\ell=2}^{n-2} \sum_{y=1}^{x-1} \pi_\ell(y)\pi_{n-\ell}(x-y), \quad 2 \leqslant x \leqslant \lfloor n/2 \rfloor. \tag{5}$$

We checked that these equations can provide numerical results up to sample sizes greater than $n \geqslant 1000$ in short running times.

Because we find reasonable to discard samples for which we observe a unique group configuration ($N_n = 1$, no obvious grouping), the $P$-values may also be computed by using the conditional distribution given that at least two groups are observed ($N_n \geqslant 2$). For the neutral model ($p = 0$) and the extra-clustering model, the conditional distribution can be obtained as

$$\pi^1_n(x) = \frac{\pi_n(x)}{1 - \pi_n(1)} = \frac{n-1}{q(n-3)} \pi_n(x)$$

if $2 \leqslant x \leqslant \lfloor n/2 \rfloor$, and 0 otherwise.

### 3.3. Clustering rates and the LR test

Given that $x$ groups are observed in a sample of $n$ individuals, the recursive equations enable us to compute a likelihood function

$$L(p) = \text{Prob}(N = x; p) = \pi_n(x)$$

and a maximum-likelihood estimate $\hat{p}$ of the clustering rate $p$ can be obtained. In practice these estimates can be computed by using basic grid search. Rejecting the unique group configuration leads to a distinct estimate

$$\hat{p} = \underset{0 \leqslant p \leqslant 1}{\arg\max} \frac{n-1}{q(n-3)} L(p).$$

For example a sample of $n = 200$ individuals with $x = 10$ groups has maximum-likelihood estimate equal to $\hat{p} = 0.30$, and this estimates becomes $\hat{p} = 0.32$ after removing the unique group configuration. The bias and the variance of $\hat{p}$ are displayed in Fig. SM3. The main contribution to the statistical error of $\hat{p}$ comes from the peak in the distribution at $\hat{p} = 1$ which also corresponds to the peak at $x = 1$ in $\pi_n$. Removing the peak at $x = 1$ (i.e., conditioning) improves the efficiency of the estimator significantly.

To test the null-hypothesis $\mathrm{H}_0 : p = 0$ against $\mathrm{H}_1 : p > 0$ and obtain a test with optimal power, the LR test is a standard approach. The LR test is based on the following test statistic:

$$s(x) = \frac{L(0)}{L(\hat{p})}$$

which can be calculated using Eqs. (3) and (5). Using this statistic, we are now able to compute a second *P*-value called the *LR test P-value*. In practical applications, this *P*-value was obtained using 10,000 Monte-Carlo replicates of $N_n$ obtained thanks to Eq. (2).

Table SM1 reports the powers of the $N_n$ test and of the LR test of the neutral model ($p = 0$) against the extra-clustering model at various levels of the clustering rate $p = p_1$. The highest powers were achieved for the LR test. Removing the unique group configuration also yielded increased power (data not reported), and all computed *P*-values removed the unique group configuration. Reasonable powers (greater than 75%) were obtained for sample sizes larger than 100 and $p_1$ larger than 0.2. The $N_n$ test generally lacks power (below 15%).

## 4. Data analysis

To illustrate our approach, we analyzed data collected from a number of previously published ecological studies. The data used here contained census sizes and observed numbers of groups in either herbivore or carnivore social species. The data sets were selected on the basis of large census sizes ($n \geqslant 50$). For each sample, the departure from the neutral model was tested using the two methods introduced in the previous section: the $N_n$ test and the LR test. The alternative hypothesis was $\mathrm{H}_1 : p > 0$. Although the $N_n$ test generally lacked power, we found useful to report the corresponding *P*-values because they had a simple interpretation as the probability of obtaining less groups than observed. Maximum-likelihood clustering rates were estimated according to the method described in the previous section. In order to reduce statistical errors and increase the power of the tests, all estimates and *P*-values were computed from the conditional distribution removing the unique group configuration. Fig. 2 displays the number of groups $N_n$ as a function of the sample size $n$ for the different populations of herbivores and carnivores.

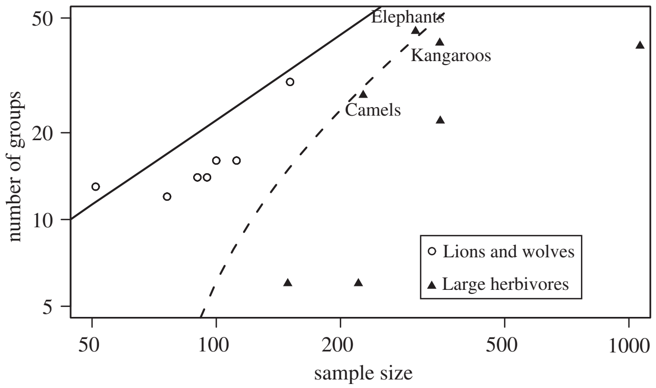

Fig. 2. The number of groups in different herbivore and carnivore populations as a function of the sample size. The solid and dashed lines represent, respectively, the expected value and the 5% quantile of the number of groups in the neutral model.

### 4.1. Social carnivores

*Wolf packs*: Gray wolves (*Canis lupus*) are pack-living animals with a complex social organization. Packs are primarily family groups. Packs include up to 30 individuals, but smaller sizes (between 8 and 12) are more common. A review of wolf social behavior and ecology can be found in Mech (1981). We used data from three sources: The Wolf project of Yellowstone national park which annually publishes accurate data on wolf pack sizes (Smith et al., 2002), and studies of wolf population recovery after quasi-extinction in Scandinavia (Wabakken et al., 2001) and in Alaska (Ballard et al., 1987). When available, the total sample size was given as the number of sampled adults (in wolves the number of pups per packs is usually small). In 2002, $n = 90$ adult wolves were sampled in Yellowstone, living in 14 packs. Using the neutral model, we obtained that $\mathrm{Prob}(N_{90} \leqslant 14) = 0.18$. The LR *P*-value was equal to 0.15. The clustering rate $\hat{p}$ was equal to 0.11. Table 1 report similar results for the year 2004. In Alaska, $n = 151$ wolves were sampled, living in 30 packs (number of pups not known). We obtained that $\mathrm{Prob}(N_{151} \leqslant 30) = 0.34$. The LR test *P*-value was equal to 0.34 as well. The clustering rate was estimated at $\hat{p} = 0.02$. In Scandinavia 76 wolves were sampled, living in 12 packs (number of pups not known). From the neutral model, we obtained $\mathrm{Prob}(N_{76} \leqslant 12) = 0.21$ and the LR *P*-value was 0.16. The clustering rate was estimated at $\hat{p} = 0.12$. The *P*-values might underestimate the true values because the pups were included in the sample.

*Lion prides*: African lions (*Panthera leo*) live in prides that typically consist of two males, 4–10 females and their offspring. The adult females are usually related to one another and are group members for life. A review of Serengeti lion behavior and ecology can be found in Schaller (1972). We used recent data from three sources: Selous Game reserve Tanzania (Spong et al., 2002), Serengeti Tanzania (Packer et al., 2005), and Kafue Park

Table 1
Data on group structure for social animals: (A) Social herbivores; (B) Social carnivores

| | Sample size | Number of herds | Rate $\hat{p}$ | $N_n$ test $P$ | LR test $P$ |
|---|---|---|---|---|---|
| (A) | | | | | |
| Springboks (browsers) | 149 | 6 | 0.40 | 0.009 | 0.008 |
| Springboks (graze) | 1064 | 40 | 0.24 | 0.001 | 0.000 |
| Fallow deers | 349 | 22 | 0.23 | 0.007 | 0.005 |
| Grant's gazelles | 221 | 6 | 0.44 | 0.004 | 0.003 |
| Wild camels | 227 | 27 | 0.14 | 0.043 | 0.042 |
| Kangaroos | 348 | 41 | 0.12 | 0.028 | 0.023 |
| African savannah elephants | 304 | 45 | 0.08 | 0.071 | 0.063 |
| | Sample size | Number of packs/prides | Rate $\hat{p}$ | $N_n$ test $P$ | LR test $P$ |
| (B) | | | | | |
| Yellowstone Wolves 2002 | 90 | 14 | 0.11 | 0.18 | 0.15 |
| Yellowstone Wolves 2004 | 112 | 16 | 0.12 | 0.13 | 0.12 |
| Alaska Wolves | 151 | 30 | 0.02 | 0.34 | 0.34 |
| Scandinavian wolf | 76 | 12 | 0.11 | 0.21 | 0.16 |
| Zambia Kafue lions | 95 | 14 | 0.12 | 0.15 | 0.13 |
| Selous Game lions | 51 | 13 | 0.00 | 0.73 | 1.00 |
| Serengeti lions | 100 | 16 | 0.10 | 0.19 | 0.16 |

The one-sided $P$-value of the $N_n$ test and the estimated clustering rate $\hat{p}$ were computed using the conditional distribution of the number of groups $N_n$ given $N_n \geqslant 2$. The one-sided $P$-values of the LR test were computed using the likelihood ratio statistic $s(x)$ and 10,000 Monte-Carlo replicates from the neutral model.

Zambia (Carlson et al., 2004). A study of social and genetic structure of Selous Game reserve lions (Spong et al., 2002) reported the presence of 14 prides, with an average number of 5.6 adults (range 2–9) and two males in each pride. These data were turned into an estimate of 51 females in the sample. A recent survey of Serengeti lions reported the presence of about 100 lionesses in the park (Packer et al., 2005). Based on an average of six females per pride, a number of 16–17 prides in Serengeti was consistent with the current data. At least 95 adult lions resided in the northern sector of Kafue National Park, either living in one of 14 prides or roaming as solitary males (Carlson et al., 2004). Among the adult lions, there were 31 males and 64 females (a sex ratio of 1:2). Nine of the 14 prides did not have a sexually mature male residing with them. Pride sizes ranged from 2 to 14 adult animals (mean = 6.4 animals per pride). Of the 17 sexually mature males that were identified, six of them were associated with prides of females while 11 lived either alone or in all-male dyads. Table 1 report results for the three lion samples. As for wolves, the lion samples exhibited high $P$-values, and low estimates of the clustering rates were obtained ($\hat{p} = 0.12, 0.00, 0.1$, respectively). Estimates of clustering rates for Zambia might be biased upward because we included males in the sample. Actual values of female counts would exhibit larger $P$-values, lower clustering rates, and an even stronger agreement with the neutral model.

### 4.2. Social herbivores

*Springbok*, *Fallow deer*, *Grant gazelle and Kangaroo*: In a study of springbok viligance in the Etosha National Park in Namibia, Burger et al., 1999 measured the time that the animals devoted to vigilance. They reported that this time differed for browsers and grazers. They also reported the number of groups observed in each species. They observed 149 browsing springboks (*Antidorcas marsupialis*) in six groups, and 1064 grazing springboks in 40 groups. Gerard and Loisel (1995) studied the variation of group size with habitat openness in large herbivores. They reported 349 Fallow deers (*Dama dama*) in 22 groups, 221 Grant gazelles (*Gazella granti*) in six groups and 348 kangarooes (*Macropus giganteus*) in 41 groups. Under the neutral model, these configurations were highly improbable with the probabilities computed as $\text{Prob}(N_{149} \leqslant 6) = 0.01$ (browsers) and $\text{Prob}(N_{1064} \leqslant 40) = 0.001$ (grazers). The clustering rates were estimated at $\hat{p} = 0.4$ for browsers and $\hat{p} = 0.24$ for grazers. The LR test rejected the neutral model ($P = 0.008$ for browsers, and $P = 0.000$ for grazers). All the data led to a strong reject of the neutral model. We obtained that $\text{Prob}(N_{349} \leqslant 22) = 0.007$ (Fallow deer), $\text{Prob}(N_{221} \leqslant 6) = 0.004$ (Grant gazelle) and $\text{Prob}(N_{348} \leqslant 41) = 0.028$ (Kangaroo). The clustering rates were $\hat{p} = 0.23$ (Fallow deer), $\hat{p} = 0.44$ (Grant gazelle), $\hat{p} = 0.14$ (Kangaroo). The LR test $P$-values were equal to 0.005, 0.003 and 0.023, respectively.

*Wild camels*: In an aerial survey of known and suspected wild camels (*Camelus bactrianus*) habitat, Reading et al. (1999) estimated group density and population size of large ungulates in the south-western Gobi Desert in Mongolia. They observed 277 Wild camels in 27 groups. Like the other large herbivores, the clustering rate was rather high ($\hat{p} = 0.12$) and the LR test as well as the $N_n$ test rejected the neutral model ($P = 0.042$).

*African savannah elephants*: In contrast to many social animals which live in stable groups, African elephants (*Loxondonta africana*) live in groups that vary according to fusion and fission events. However, core social groups that are composed of well defined individuals can be determined. These groups may temporarily split into smaller units or merge into bigger units over the course of days. Elephants living in and around Amboseli National Park in Kenya are individually known and have been studied since 1972 by the Amboseli Elephant Research Project (AERP). As count data, we considered the 304 adult female elephants that have been listed in the Supplementary Material given by Archie et al. (2006). The adult female elephants live in 45 core groups. The mean number of individuals in each group is 6.75 and the standard deviation is 3.92. Both the LR and the $N_n$ test give small $P$-values (respectively, 0.06 and 0.07) that are nonetheless bigger than 0.05 indicating that the neutral model cannot be rejected based on these summary statistics. The clustering rate was estimated at $\hat{p} = 0.08$.

## 5. Discussion

This study introduced new statistical models of mammal group patterns that are based on genetic relatedness. The great advantage of these models is their simplicity, their parsimony and the fact that statistical theory can be carried out in order to assess deviations from a neutral model.

Before giving interpretations of results, an important point to recall is that the neutral model does not serve as an explanation for the group formation process itself, but rather aims at assessing the power of genetic relatedness to solely explain the observed group patterns. Genetic relatedness is usually computed from kinship coefficients and genealogical relationships. We used the coalescent to model a genealogy of genes and thus to describe genetic relationships. Then it enabled us to define groups for which genetic relatedness is stronger within than between. Here a straightforward parallel can be drawn with population genetics theory where simplified and unrealistic random mating models like the Hardy–Weinberg or the Wright–Fisher models are used in tests of selective neutrality (Hartl and Clark, 1997). Testing departures from these models using test statistics such as the Tajima $D$ are nevertheless informative, as they yield clues about the past demographic or genetic events which occurred in a population (Tajima, 1989; Hein et al., 2005).

Population genetics approaches generally use summary statistics, and describe the probability distribution of these statistics based on genealogical tree models. Philosophically our approach is then close to population genetics, since the coalescent trees also serve as a null-model for testing the assumption that genetic relatedness explains group patterns without resorting to other evolutionary pressures, like benefit from group hunting or lowered predation risk.

In social carnivore examples, the neutral model was generally not rejected at the 5% level, and the clustering rates remained at low values ($\hat{p} < 0.12$) (see Figs. 2, 3 and Table 1). A perhaps remarkable fact is that the elephant group pattern also displayed a similar agreement with the neutral model, classifying the elephants together with the carnivores in Fig. 2. The case of wild camels and kangaroos is also instructive. Although the test is significant at the 5% level, the clustering rates are similar to those of carnivores and elephants and the $P$-values are significantly larger than those obtained for the other herbivores. The agreement of carnivore and elephant data (and, to a lesser extent, wild camels and kangaroos) supports the hypothesis that hunting in groups carries only weak explanatory power compared to a theory that predicts that mammals group with their relatives. These observations are consistent with summaries which have argued that communal hunting actually has little power to explain group patterns in felids (Packer et al., 1990) and across social carnivores in general (Caro, 1994).

Social herbivores like gazelles, springboks, or deers live in large herds which convert into large clustering rates $\hat{p} \approx 0.23$–$0.44$. In these examples, the poor agreement with the neutral model predictions was an expected result because the typical group sizes are by far larger than the carnivore group sizes. When compared to the carnivore results, the departure from the null-model suggests that genetic relatedness may explain herbivores group patterns only partially, and that external evolutionary forces are probably involved. Back to the data with low clustering rates, one can observe that wild camels habitat corresponds to steppic areas with no known predators. The same phenomenon occurs in Australia for kangaroos and in Africa for large carnivores and elephants. Keeping in mind that we have used a restricted sample of available ecological surveys, one can observe that high clustering rates correlate well with a factor like predator avoidance. Indeed, the results displayed in Figs. 2 and 3 support the hypothesis that genetic relatedness is likely to predict group patterns when mammals have few or no predators. Of course, further confirmation of this hypothesis would be worthwhile. This would require a thorough scan of the biological literature to increase the number of observations of group patterns in social species. For example, dolphins are frequently reported to live in pods of 5–7 adults (average size adjusted for bias in the detectability of large vs small pods in aerial surveys), and this value is consistent with the average value predicted by the neutral model ($n/E[N_n] \approx 5$). In contrast, many fishes live in very large schools which are usually believe to confer protection against larger predators. Further works would also be needed to combine fission/fusion models and genetic relatedness to better explain variations of group size with time, breeding seasons or migration epochs (Archie et al., 2006).

Predictions from the null-model can also be confirmed when more data than the mere number of groups are available. The elephant study in Archie et al. (2006) actually provides additional data on the group pattern, which enables us to study the distribution of the number of elephants in each group. The group distribution of the 304 elephants surveyed in Archie et al. (2006) agrees with a power-law distribution of exponent $\alpha = 1.57$ estimated for

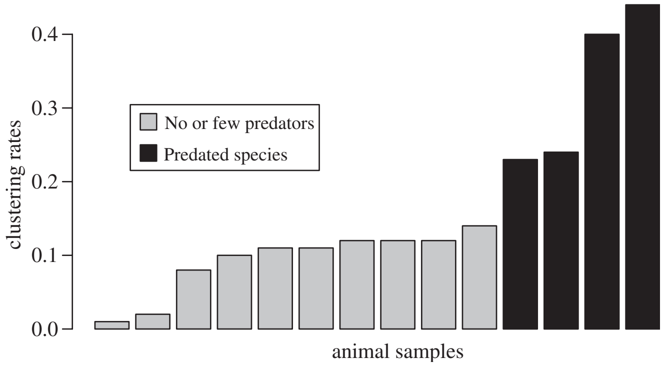

Fig. 3. Clustering rates for 14 animals species (see Table 1). Low clustering rates, colored in gray, are associated with species that have few or no predators.

groups of size larger than 5 (log–log linear regression, $R$-squared $= 0.82$, $P = 0.0002$). We performed Monte-Carlo simulations of group size distribution for 304 individuals under the neutral model (10,000 replicates). The simulated replicates provided an excellent fit to power-law distributions ($R$-squared greater than 0.8), and for each replicate we computed an exponent for the power law. The lower and upper quartiles of the distribution of estimated exponents were equal to $q_\ell = 1.01$ and $q_u = 1.90$, which shows that $\alpha = 1.57$ lies within the interval predicted by the neutral model.

Network theory has recently become increasingly important in ecology and evolution (Proulx et al., 2005). In particular social networks have emerged as a paradigm of the complexity of human or animal interactions (Wasserman and Faust, 1994; Frank, 1998; Scott, 2000). Ruling the basic principles of network formation is an highly difficult task, and there is a large tradition for extracting community structure by cluster analysis which is generally represented by a binary tree structure. Communities are usually inferred by cutting the tree at an appropriate height. A novelty of the present study is that networks and their associated trees are considered as unobserved or hidden data, and no attempts are made to reconstruct them. Instead, the networks are viewed as random objects which enable us to validate biological assumptions about the observed patterns in a statistical way. A premise of this study is that network and coalescent approaches can fruitfully be integrated in specific statistical approaches which reveal themselves useful for analyzing the available masses of ecological data.

## Acknowledgments

The authors are grateful to Noah Rosenberg and to Mattias Jakobsson for interesting comments on an early version of the manuscript.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://www.10.1016/j.jtbi.2007.07.012.

## References

Aldous, D.J., 2001. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statist. Sci. 6, 23–34.

Archie, E.A., Moss, C.J., Alberts, S.C., 2006. The ties that bind: genetic relatedness predicts the fission and fusion of social groups in wild african elephants. Proc. Roy. Soc. B Biol. Sci. 273, 513–522.

Ballard, W.B., Whitman, J.S., Gardner, C.L., 1987. Ecology of an exploited wolf population in south-central alaska. Wildlife Monographs, vol. 98.

Bertram, B.C.R., 1978. Living in Groups: Predators and Prey, Behavioural Ecology: An Evolutionary Approach. Blackwell Scientific Publications, Oxford, pp. 64–96.

Blum, M.G.B., François, O., 2005a. On statistical tests of phylogenetic imbalance: the sackin and other indices revisited. Math. Biosci. 195, 141–153.

Blum, M.G.B., François, O., 2005b. Minimal clade size and external branch length under the neutral coalescent. Adv. Appl. Prob. 37, 647–662.

Bonabeau, E., Dagorn, L., 1995. Possible universality in the size distribution of fish schools. Phys. Rev. E 51 (6), 220–223.

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

Burger, J., Safina, C., Gochfeld, M., 1999. Factors affecting vigilance in springbok: importance of vegetative cover, location in herd, and herd size. Acta ethologica 2, 97–104.

Carlson, A.A., Carlson, R., Bercovitch, F.B., 2004. Kafue national park african wild dog conservation project. Technical Report, Kafue National Park, Zambia, 2004, Annual Report, Zoological Society of San Diego, CA.

Caro, T.M., 1994. Cheetahs of the Serengeti Plains. University of Chicago Press, Chicago.

Dawkins, R., 1989. The Selfish Gene. Oxford University Press, Oxford.

Foster, K.R., Wenseleers, T., Ratnieks, F.L.W., 2006. Kin selection is the key to altruism. Trends Ecol. Evol. 21 (2), 57–60.

Frank, S.A., 1998. Foundations of Social Evolution. Princeton University Press, Englewood Cliffs, NJ.

Gerard, J.F., Loisel, P., 1995. Spontaneous emergence of a relationship between habitat openness and mean group size and its possible evolutionary consequences in large herbivores. J. Theor. Biol. 176, 511–522.

Giraldeau, L.A., Caraco, T., 1993. Genetic relatedness and group size in an aggregation economy. Evol. Ecol. 7, 429–438.

Gueron, S., Levin, S.A., 1995. The dynamics of group formation. Math. Biosci. 128, 243–264.

Hamilton, W.D., 1964. The evolution of social behavior. J. Theor. Biol. 7, 1–52.

Hamilton, W.D., 1971. Geometry for the selfish herd. J. Theor. Biol. 31, 295–311.

Hamilton, I.M., 2000. Recruiters and joiners: using optimal skew theory to predict group size and the division of resources within groups of social foragers. Am. Nat. 155, 684–695.

Hartl, D.L., Clark, A.G., 1997. Principles of Population Genetics, third ed. Sinauer, Sunderland, MA.

Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution. Oxford University Press, Oxford.

Hwang, H.K., Neininger, R., 2002. Phase change of limit laws in the quicksort recurrence under varying toll functions. SIAM J. Comput. 31, 1687–1722.

Inman, A.J., Krebs, J.R., 1987. Predation and group living. Trends Ecol. Evol. 2, 31–32.

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13, 235–248.

Lazo, A., 1994. Social segregation and the maintenance of social stability in a feral cattle population. Anim. Behav. 48, 1133–1141.

Mech, L.D., 1981. The Wolf: The Ecology and Behavior of an Endangered Species. University of Minnesota Press, Minneapolis.

Niwa, H., 2003. Power-law versus exponential distributions of animal group sizes. ⟨http://arxiv.org/pdf/cond-mat/0305241⟩.

Nordborg, M., 2003. Coalescent theory. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics, second ed. Wiley, Chichester, UK, pp. 602–635.

Nudds, T.D., 1978. Convergence of group size strategies by mammalian social carnivores. Am. Nat. 112, 957–960.

Packer, C., Scheel, D., Pusey, A.E., 1990. Why lions form groups: food is not enough. Am. Nat. 136, 1–19.

Packer, C., Hilborn, R., Mosser, A., 2005. Ecological change, group territoriality, and population dynamics in serengeti lions. Science 307, 389–393.

Proulx, S.R., Promislow, D.E.L., Phillips, P.C., 2005. Network thinking in ecology and evolution. Trends Ecol. Evol. 20 (6), 345–353.

Pulliam, H.R., 1973. On the advantages of flocking. J. Theor. Biol. 38, 419–422.

Pulliam, H.R., Caraco, T., 1978. Living in Groups: Is There An Optimal Group Size? Blackwell Publishing, Incorporated, pp. 122–147.

Reading, R.F., Mix, H., Lhagvasuren, B., Blumer, E.S., 1999. Status of wild bactrian camels and other large ungulates in south-western mongolia. Oryx 33, 247–255.

Reinhardt, V., Reinhardt, A., 1981. Cohesive relationships in a cattle herd (*Bos indicus*). Behaviour 77, 121–151.

Rösler, U., 2001. On the analysis of stochastic divide and conquer algorithms. Algorithmica 29, 238–261.

Rubenstein, D.I., 1978. On predation, competition, and the advantages of group living. Perspect. Ethol. 3, 205–232.

Schaller, G.B., 1972. The Serengeti Lion. University of Chicago Press, Chicago.

Scott, J., 2000. Social Network Analysis: A Handbook, second ed. Sage Publications, London.

Smith, D.W., Stahler, D.G., Guernsey, D.S., 2002. Yellowstone wolf project annual report, Technical Report, Yellowstone National Park.

Spong, G., Stone, J., Creel, S., Björklund, M., 2002. Genetic structure of lions (*Panthera leo l.*) in the selous game reserve. J. Evol. Biol. 15, 945–953.

Tajima, F., 1989. DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 129, 229–240.

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

Wabakken, P., Sand, H., Liberg, O., Bjärvall, A., 2001. The recovery, distribution, and population dynamics of wolves on the scandinavian peninsula, 1978–1998. Can. J. Zool. 79 (4), 710–725.

Wasserman, S., Faust, K., 1994. Social Network Analysis. Cambridge University Press, Cambridge, UK.