Version of Record: https://www.sciencedirect.com/science/article/pii/S0040580919301728
Manuscript_746e466e126e5b2ea0ecbf9695455628

A multi-epoch model for the number of species within genera

Olivier François*

University Grenoble-Alpes

Abstract

An early question in evolutionary theory asked why frequency distributions of taxonomic group sizes exhibit "hollow curves" so frequently. An answer to this question was provided by G. Udny Yule's seminal contribution introducing a discrete model for those distributions. But Yule observed that the fit of his model to observed distributions was sometimes imperfect, in particular for the class of reptiles. The present study introduces a multi-epoch extension of the discrete Yule model that accounts for unobserved extinction of ancient lineages. The multi-epoch model is described as a Pòlya urn embedded in a continuous-time branching process with an harmonic sequence of diversification rates. The main results include equivalent descriptions of multi-epoch models, their probability distributions, expected values, tail behavior and a self-similarity property. As an illustration of the theory, the multi-epoch model is applied to study the taxonomic diversity of reptile species, and provides a much better fit to the observed distribution of species than the original discrete Yule model.

Keywords: Yule distribution, Branching processes, Urn models, Tree embedding, Evolution theory

*Corresponding author: TIMC-IMAG, CNRS UMR 5525, Univ. Grenoble-Alpes, Grenoble INP, 38000 Grenoble, France.

Email address: olivier.francois@univ-grenoble-alpes.fr (Olivier François)

© 2019 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

1. Introduction

The distribution of the number of species within genera follows one of the oldest laws in evolution (Willis, 1922). This distribution exhibits a "hollow curve" for which most genera contain only one species and a few genera contain a large number of species. The observed pattern has been long-recognized, and the hollow curve was summarized by Charles Darwin's note in *The Origin of Species* (1859) – "Rarity is the attribute of vast numbers of species in all classes". Not restricted to a particular taxonomic level, hollow curves occur for example in ecology where they represent relative species abundance in various habitats (McGill *et al.*, 2007).

The discrete Yule model was one of the earliest attempt at characterizing the shape of the hollow curve in a mathematical way (Yule, 1925). The model derives from a continuous-time pure-birth branching process with a constant rate of diversification, and explains taxonomic diversification as a consequence of pure randomness. More precisely, a constant rate pure-birth branching process is a continuous-time model of a tree in which each lineage splits at a constant rate. In the branching process, the distribution of the number of descendants at a particular time point is geometric, and its mean value grows exponentially with the rate of diversification (Yule, 1925). The discrete Yule distribution is obtained by considering the number of descendants after a random period of evolution having an exponential distribution of mean equal to one.

While the simplicity of the discrete Yule distribution makes it interesting as a null-model to test hypotheses about evolution, it may not fully reflect the next words of Darwin's quotation: "If we ask ourselves why this or that species is rare, we answer that something is unfavourable in its conditions of life" (Darwin, 1859). In fact, the discrete model belongs to a class of skewed distributions that find applications far beyond evolutionary theory, describing various processes such as the size of cities, scientific citations, superstardom, or species abundance (Simon, 1955; Chung and Cox, 1994; Chu and Adami, 1999). In an ecological context, Nee (2003) also showed that the discrete distribution provides a good fit

to species abundance distributions. Considering higher taxonomic levels, Yule acknowledged, however, that the fit of the model to observed distributions was imperfect, in particular for the family of snakes.

An alternative mathematical description of the discrete Yule distribution is as a Pòlya urn model, a discrete-time version of the branching process in which lineages correspond to balls drawn from an urn (Simon, 1955; Athreya and Karlin, 1968; Mahmoud, 2008). Probabilistic urn models also arises in random processes with reinforcement, in which previously visited states see their probability of visit increased (Pemantle, 2007). In processes with reinforcement, the tail of the distribution of states is often equivalent to the tail of a power law (Newman, 2005). In addition, urn models have connections with the sampling theory of neutral alleles in population genetics (Crane, 2016; Hoppe, 1987). In all cases, continuous trees or discrete urn models lead to statistically identical histories of diversification events. The connection between urns and trees provides a natural approach to extend the discrete Yule distribution, and to improve its fit to observed frequency spectra.

This study introduces a multi-epoch model for describing the distribution of the number of species within genera. In the multi-epoch model, the most ancient epochs correspond to lower effective diversification rates, decreasing according to the harmonic sequence, and representing the unobserved extinction of some ancient lineages. The multi-epoch model could be defined either as an urn model or as a non-constant rate branching process. The main results for this model are explicit formulas for the distribution of the number of species, first moment and the tail of the distribution. In addition, the multi-epoch model exhibits an interesting self-similarity property at its critical rate. An application of the multi-epoch model to the taxonomic diversity of reptile species (Uetz, 2000) is studied. For this class, the multi-epoch model fits the observed distribution of species better than the original discrete Yule distribution.

2. A multi-epoch model

The discrete Yule distribution. The discrete Yule distribution can be described as the following urn model, also called Polya's urn or reinforcement model (Newman, 2005). The process starts with an urn containing two balls, a black one and a white one. The white ball may be viewed as representing an ancestral species in a tree-like evolutionary process. The black ball has weight one, and the white ball has weight $\lambda > 0$. The parameter λ can be interpreted as a speciation or diversification rate for "white" lineages. Balls are drawn from the urn with probability proportional to their respective weights. If the color resulting from a drawing is white, then the white ball is replaced in the urn, and an exact copy of it is added to the urn content. This event corresponds a speciation event, *i.e.*, the occurrence of a new species in the urn. The sampling process is continued until the black ball is drawn. Note that the tree topology corresponding to the series of speciation events could be made explicit by labelling each new ball with a distinct label. This is not useful for the purpose of describing the discrete Yule distribution, and the labels will be ignored in the derivation of the result.

The discrete Yule model describes the probability distribution of the number of white balls (species) resulting from the sequence of drawings. Because the number of white balls corresponds to the waiting time until the black ball is drawn, the distribution can be formulated as a product of sampling probabilities. Considering the inverse speciation rate, $\rho = 1/\lambda$, the probability of having n species is indeed given by

$$p(n|\rho) = \frac{1}{(1+\rho)}\frac{2}{(2+\rho)}\cdots\frac{n-1}{(n-1+\rho)}\frac{\rho}{(n+\rho)}, \quad n \geq 1.$$

In the above product, the first $(n-1)$ terms represent the probabilities of drawing white balls, whereas the n^{th} term represents the probability that the last draw is a black ball. By definition of the Beta Euler function,

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 u^{a-1}(1-u)^{b-1}du, \quad a,b>0\,,$$

the probability of having n species can be rewritten as follows

$$p(n|\rho) = \rho\,\frac{\Gamma(n)\Gamma(\rho+1)}{\Gamma(\rho+1+n)} = \rho\,B(\rho+1,n), \quad n \geq 1.$$

Remarkably, the result entails a simple integral representation for the distribution

$$p(n|\rho) = \rho \int_0^1 u^{\rho}(1-u)^{n-1}du, \quad n \geq 1\,.$$

While the above integral representation results from simple calculus, it is also connected to the embedding of the Pòlya urn into a pure-birth branching process with birth rate λ (Athreya and Karlin, 1968; Athreya and Ney, 1972). The corresponding random tree is often called the equal rate Markov model or the *Yule* model of speciation (Aldous, 2001). The random tree is rooted with a unique ancestor and evolves for a period of time T, a random variable having an exponential distribution of rate one. In the continuous tree, the number of white balls corresponds to the number of branching events. The unit period T is called a random *epoch*, so that the Yule model corresponds to a single epoch. The unit rate for the period T corresponds to the evolutionary time scale for a genus occurrence. This rate was equal to $\mu > 0$ in Yule's original study. With rate μ, the model parameter ρ rescales as $\rho = \mu/\lambda$, and the probabilities $p(n|\rho)$ are unchanged. For this reason, only $\mu = 1$ is considered in the rest of this study.

The N-epoch model. Here, an extension of the urn model considers that the urn initially contains N black balls and a single white ball ($N \geq 1$). Again, the white ball is viewed as representing an ancestral species in a tree-like evolutionary process, and its weight is still $\lambda > 0$. Each black ball has weight one. Balls are drawn from the urn with probability proportional to their weights. When drawn from the urn, white balls are replaced in the urn, and an exact copy of them is added to the urn content. In contrast, when a black ball is drawn, it is removed from the urn. The last event could be called a transition event. The sampling process is continued until all N black balls are removed from the

urn. The Yule(N, ρ) distribution describes the law of the number of white balls resulting from the sequence of drawings.

Before describing the Yule(N, ρ) distribution mathematically, some remarks may help to better understand the connection between urns, trees and their timescales. A first remark is that the Yule$(1, \rho)$ distribution is the discrete Yule distribution. The Yule$(1, \rho)$ distribution is equivalently described by giving a weight one to the white ball and a weight ρ to the black ball. This operation is equivalent to changing the time-scale and speciation rate in the continuous tree model. A second remark is that, in a model with N epochs, the waiting time for the first transition event follows the Yule$(1, N\rho)$ distribution. This can be shown by considering an equivalent model for the first epoch in which there is a single black ball of weight one, and the weight of the white ball is equal to λ/N. At the end of the first epoch, there are $N-1$ black balls, and the weight of each white ball increases to $\lambda/(N-1)$. The above argument can be repeated for each period. During epoch k, which corresponds to having k black balls in the urn, the speciation rate for white balls is equal to λ/k. The N-epoch model is thus a model with unequal diversification rates in which the rates decrease backward in time according to the harmonic sequence. In this model, each ancestor species leaves more descendant species in recent epochs than in ancient epochs.

3. Main results

This section describes an integral representation of the probability distribution of the number of white balls, or equivalently the number of extant species within genera, for the Yule(N, ρ) distribution. It uses this representation to compute first order moments and tails, and provides an efficient simulation algorithm for the N-epoch model.

The Yule(N, ρ) distribution. Theorem 1 extends the formula available for the discrete Yule distribution. The result can be stated as follows.

Theorem 1. *Let N be a positive integer, $\lambda > 0$ and $\rho = 1/\lambda$. The probabil-*

ity distribution of the number of species in the N-epoch model is given by the following formula

$$p(n\,|N,\rho) = N\rho \int_0^1 u^\rho (1-u^\rho)^{N-1}(1-u)^{n-1} du\,, \quad n \geq 1\,.$$

Like the discrete Yule distribution, the N-epoch model has a natural embedding in a pure-birth branching process. In the continuous tree, the rate of diversification varies from an epoch to the next one. The continuous tree has N independent epochs T_1, ..., T_N, where T_1 corresponds to the most recent epoch and T_N corresponds to the most ancient one. The T_k's are defined as exponentially distributed random variables with rate one. During epoch k, the speciation rate is λ/k. To derive an expression for the Yule(N,ρ) distribution, this is equivalent to consider a tree of height $T = T_1 + T_2/2 + \cdots + T_N/N$ in which the speciation rate is constant and equal to λ. The Theorem is then a consequence of Yule's result for the equal rate Markov model, for which the distribution of the number of species is geometric. This remark leads to

$$p(n\,|N,\rho) = \mathbb{E}[e^{-\lambda T}(1-e^{-\lambda T})^{n-1}]\,, \quad n \geq 1,$$

which corresponds to the integral in Theorem 1. See appendix for mathematical proofs.

By using Newton's binomial formula and the Beta function, integrals in Theorem 1 can be computed accurately for all n as follows

$$p(n|N,\rho) = N\rho \sum_{k=1}^{N} (-1)^{k-1} \binom{N-1}{k-1} B(k\rho+1, n)\,, \quad n \geq 1. \tag{1}$$

Note that for $N = 1$, the discrete Yule distribution is recovered.

First moment and critical value. Like the standard Yule$(1,\rho)$ distribution, the multi-epoch model has a critical value at $\rho = 1$. For $\rho \leq 1$, the expected values predict an infinite number of species, whereas for $\rho > 1$, the expected values are finite and can be expressed in a closed form.

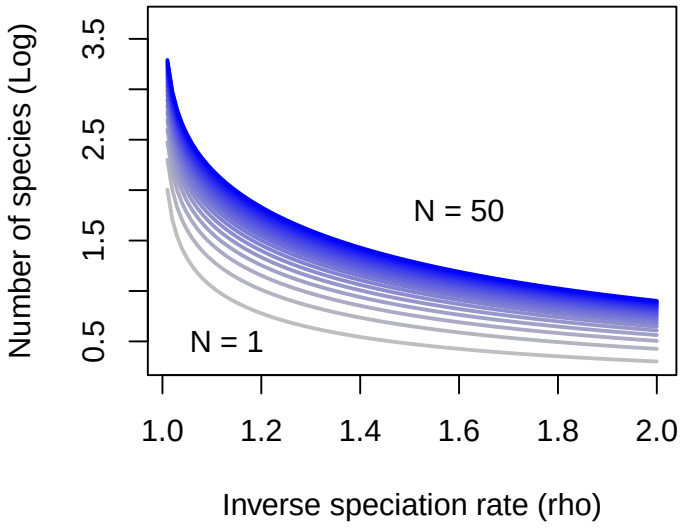


Figure 1. Expected number of species in the super-critical multi-epoch model ($\rho > 1$). The number of epochs, N, is varied in the range $[1, 50]$. Expected values are computed according to the formula given in Theorem 2, and displayed on a logarithmic scale (base 10).

Theorem 2. *Let X_N be a random variable having the Yule(N, ρ) distribution. Let $\rho = 1/\lambda$. For all $N \geq 1$, the expected number of species within genera is equal to*

$$\mathbb{E}[X_N|N, \rho] = \infty, \qquad \text{if } \rho \leq 1\,,$$

and

$$\mathbb{E}[X_N|N, \rho] = N\,B(1-\lambda, N)\,, \quad \text{if } \rho > 1\,.$$

In the super-critical case ($\rho > 1$), the expected number of species is a finite number. Its expression coincides with the result known for the discrete Yule distribution for $N = 1$. For general values of N, explicit values can be obtained for integer values of ρ. In the particular case of $\rho = 2$, the expected number of species is given by the ratio of the even numbers to the odd numbers between 1 and $2N$

$$\mathbb{E}[X_N|N, \rho = 2] = \frac{2}{1}\frac{4}{3}\cdots\frac{2N}{(2N-1)}\,.$$

For other super-critical values of ρ, the formulas are less explicit, but an asymptotic result for the large N can be described according to Stirling's formula as follows ($\lambda < 1$)

$$\mathbb{E}[X_N|N, \rho] \sim \Gamma(1-\lambda)N^{\lambda}\,, \quad N \to \infty,$$

where the $\sim$ symbol means that the sequences are mathematically equivalent (Abramowitz and Stegun, 1970). For $\rho = 2$ and $N = 100$, the asymptotic approximation is equal to $\sqrt{\pi N} \approx 17.72$, while the exact value is around 17.74. The result shows that the expected number of species grows like a power of the number of epochs, and that growth is slower for smaller diversification rates (Figure 1).

The results for the multi-epoch model can be compared to those for a constant rate model. In the urn process, a constant rate model would keep the number of black balls at their initial value, N, by replacing them in the urn each time they are drawn from the urn. The replacement rules for white balls are unchanged, and the process is stopped after drawing N black balls. Like for the multi-epoch model, results for a constant rate model could be obtained by applying a continuous tree embedding approach. In the continuous time model, N independent epochs are considered, and epochs have exponential distributions of rate one. The probability distribution of the number of white balls after drawing N black balls is then given by the following formula

$$p_{\text{constant}}(n\,|N,\rho) = \rho^{N-1}\int_0^1 u^{\rho}(1-u)^{n-1}\log(1/u)^{N-1}du\,,\quad n\geq 1\,.$$

The constant rate model is also critical at $\rho = 1$, and, for $\rho > 1$, its expected value is given by the following formula

$$\mathbb{E}_{\text{constant}}[X_N|N,\rho] = \left(\frac{1}{1-\lambda}\right)^N\,.$$

In the constant rate model, the expected number of species grows exponentially with N, whereas the growth is much slower in the Yule(N,ρ) model.

Tail of the Yule(N,ρ) distribution. Next, the tail of the distribution in the N-epoch model is studied, and the next result shows that it is described by a power law.

Theorem 3. *Let N be a positive integer, $\rho > 0$ and consider X_N, the number of species within genera, having a Yule(N,ρ) distribution. This distribution satisfies*

$$p(n|N,\rho) \sim \frac{N\rho\Gamma(\rho+1)}{n^{\rho+1}},\quad n\to\infty\,,$$

and

$$\mathbb{P}(X_N > n|N,\rho) \sim \frac{N\rho\Gamma(\rho)}{n^{\rho}},\quad n\to\infty\,,$$

where $\Gamma(\rho)$ denotes the Gamma Euler function. The $\sim$ symbol means that the sequences are mathematically equivalent.

The result shows that the tail probabilities are asymptotically larger by a factor N in the multi-epoch model compared to the single epoch model. Thus reinforcement is stronger in the multi-epoch model than in the standard discrete Yule model. The power exponents are, however, independent on the number of epochs, and the tail of the distribution for larger N does not differ very much from those with smaller N. The difference with the discrete Yule model arises in a more subtle way, as the distribution tends to flatten with large N's and puts more probability on larger clades.

Computer simulation of the Yule(N, ρ) distribution. By definition of the model, the Yule(N, ρ) distribution can be simulated as the number of white balls resulting from the Pòlya urn scheme. Changing weights as described in section 2.2, the Yule(N, ρ) distribution can thus be defined recursively as follows. Let us denote by Polya(x, ρ) the distribution of the number of white balls when the urn process is started with x white balls of weight ρ and $N = 1$ black ball of weight one. Then define a sequence of N random variables as follows

$$Y_N = \text{Polya}(1, \rho N)$$

and, for $k = N - 1, \ldots, 1$,

$$Y_k = \text{Polya}(Y_{k+1}, \rho k).$$

According to the Pòlya urn scheme, Y_1 follows the Yule(N, ρ) distribution. Simulation results for $m = 100,000$ samples confirmed that the distribution of Y_1 was described by equation (1) (see Figure 2, for $N = 3$, $\rho = 0.9$ and $\rho = 1.2$).

The above forward equations reproduce the urn scheme mechanistically, but they are generally costly for large N and values of $\rho \leq 1$. A much more efficient (backward) simulation algorithm can be obtained from the integral representation presented in Theorem 1. From this representation, the Yule(N, ρ) distribution falls in the category of Beta-geometric distributions (Johnson and Kotz, 1969). It can be interpreted as a mixture of geometric distributions, Geom(U^λ), where U is sampled from a Beta distribution with shape parameters 1 and N,

$$U \sim_d \text{Beta}(1, N)$$

and

$$X_N \sim_d \text{Geom}(U^\lambda)\,.$$

In the above formulas, the symbol $\sim_d$ means that the variable on the left hand-side is sampled according to the distribution on the right hand-side. This alternative representation of the probability distribution does not simulate any tree

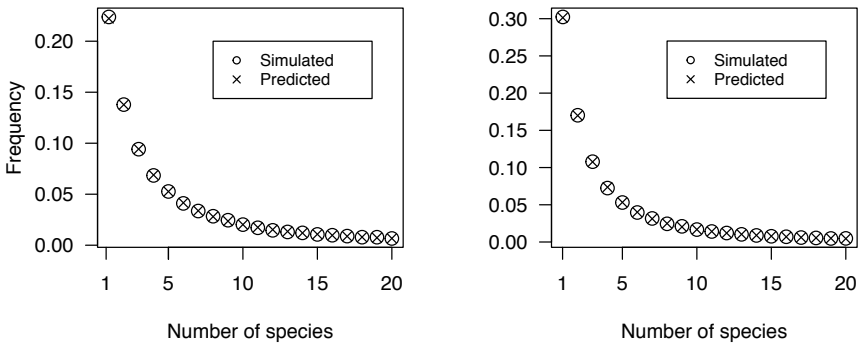


Figure 2. Frequencies from the Pòlya urn process for the Yule(N, ρ) and theoretical values predicted by Theorem 1. The multi-epoch model had $N = 3$ epochs with $\rho = 0.9$ (left) and $\rho = 1.2$ (right). Number of simulations, $m = 100,000$.

or urn models. It is considerably faster than simulating the urn process from the set of recursive equations.

Critical case. The critical case $\rho = 1$ exhibits a self-similarity property. Theorem 4 shows that the critical N-epoch Yule distribution rescaled by the number of epochs, N, corresponds to the critical Yule distribution with $N = 1$. A more precise statement of the property is as follows.

Theorem 4. *Consider the critical case $\rho = 1$. For $N \geq 1$, let X_N be a random variable having Yule$(N, 1)$ distribution. Let $\lceil x \rceil$ denote the least succeeding integer of x. Then the distribution of the renormalized random variable $\lceil X_N/N \rceil$ is the critical discrete Yule distribution,*

$$\lceil \frac{X_N}{N} \rceil \sim_d X_1 \, .$$

Here X_1 is a random variable following the discrete Yule distribution, and the symbol $\sim_d$ means that the variables share the same distribution.

4. Application to reptiles

In this section, a likelihood function of the parameters of the multi-epoch model is proposed and used to adjust the Yule(N, ρ) distribution to the observed frequencies of species in reptile genera.

Maximum-likelihood estimation. Consider a sample with m observations (n_i), $i = 1, \ldots, m$, from the Yule(N, ρ) distribution. Since the distribution is explicit, a maximum-likelihood method can be used for fitting the N-epoch model to the observed data. To apply the maximum-likelihood approach, the sampled data can be summarized by an histogram of observed frequency, assuming that observations from a very large tree are independent. Let us define

$$m_k = \#\{i : n_i = k\}\,, \quad 1 \leq k \leq K,$$

the number of genera having k species, where $K = \max n_i$ is the largest value observed in the sample. Considering a multinomial distribution for the observed frequency data, the log-likelihood function is defined as follows

$$L_N(\rho) = \sum_{k=1}^{K} \frac{m_k}{n} \log p(k|\rho, N)\,.$$

The log-likelihood corresponds to the cross-entropy of observed and predicted probability distributions. For each N, the parameter ρ can estimated by maximizing the log-likelihood function over a range of values.

Next, a series of experiments was performed to evaluate the bias and variance of the maximum likelihood estimate (MLE) for N = 1-8 and ρ in the range $[0.7, 1.6]$. To keep in order with the sample size in the real data ($m = 1196$), samples of size $m = 1,200$ were considered. Samples were simulated by using the Beta-geometric representation of the Yule(N, ρ) distribution. First, a random value, u_i, was sampled from the Beta$(1, N)$ distribution, and n_i was then sampled from a geometric distribution of rate u_i^{λ}, for $i = 1, \ldots, m$.

The squared bias of the MLE of the inverse speciation rate (ρ) was small, around 1.45e-05, but differed from zero significantly (t-test $P = 0.003$). Analysis

Table 1: Inverse speciation rate estimates (MLEs) for the reptile data

N	estimate	lower	upper	log-likelihood
1	0.548	0.468	0.579	-277.9
2	0.872	0.813	0.920	-217.8
3	1.069	1.009	1.112	-203.8
4	1.206	1.154	1.257	-200.3
5	1.317	1.257	1.376	-200.4
6	1.411	1.342	1.462	-201.8
7	1.479	1.419	1.547	-203.7
8	1.547	1.479	1.616	-205.8

N: Number of epochs, estimate: inverse speciation rate, lower and upper: 95% confidence interval.

of variance did not detect significant difference among the squared bias values for different N (ANOVA $P = 0.34$) or different ρ (ANOVA $P = 0.1$). The variance of the MLE of the inverse speciation rate varied in the range [2.34e-04, 5.56e-03]. The logarithm of the variance exhibited a linear trend in which larger variance was observed for smaller ρ and larger N. Overall those results indicated that the MLEs of ρ were accurate (low bias) and precise (low variance) in the range of values considered.

Reptile data. Reptiles (*Reptilia*) form a highly diverse class that consists of 10,885 species classified in $m = 1,196$ genera (Uetz, 2000). In reptiles, the average number of species within genera is around 9.101. The empirical distribution of the number of species within genera exhibits a typical hollow curve (Figure 3). About 21 percent (20.81%) genera contain more that 10 species, and 0.85% genera contain more that 100 species. In his original study, Yule (1925) used the family of snakes to illustrate its mathematical theory of evolution, and acknowledged that the constant rate branching process provided a

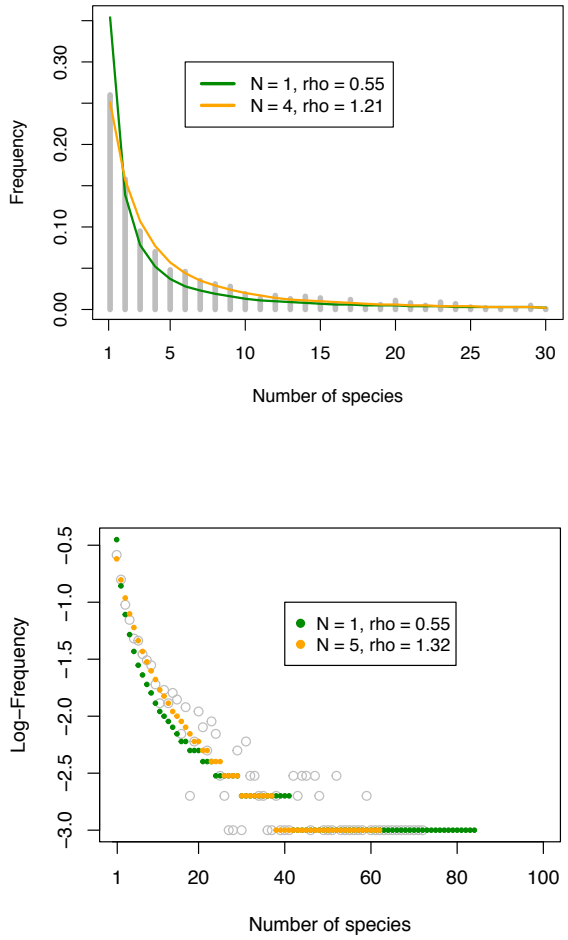


Figure 3. Observed frequencies of reptile species within their genera for less than 30 and 100 species (grey bars and circles). The green line and points show the values predicted by the discrete Yule model ($N = 1$). **Top:** The brown line shows the values predicted by a model with $N = 4$ epochs. **Bottom:** The brown line shows the values predicted by a model with $N = 5$ epochs (Log-scale). Values for ρ are MLEs.

poor fit to the snake data. After fitting the discrete Yule model to the reptile data, the MLE for the inverse speciation rate was equal to $\hat{\rho}_1 = 0.548$ (Table 1). A confidence interval for this estimate was computed by using the bootstrap method, and was equal to CI = (0.468,0.579). Figure 3 shows that the resulting Yule(1, $\hat{\rho}$) distribution provided an imperfect fit to the observed frequencies of species in genera. Considering nine histogram bins, the chi-squared goodness-of-fit statistic was equal to 196.36, and the discrete Yule model was rejected by the chi-squared test (df = 8, $P < 10^{-10}$). In addition, a subcritical estimate, $\hat{\rho} < 1$, predicting an infinitely large expected value for number of species was biologically implausible.

Estimates for the N-epoch model reached their maximum likelihoods for $N = 4, 5$ and MLEs of the inverse speciation rate were equal to $\hat{\rho}_4 = 1.257$ and $\hat{\rho}_5 = 1.376$ respectively (Table 1, Figure 3). The likelihood functions exhibited a large plateau and the confidence intervals overlaped significantly (Table 1, Figure S1). For this reason, providing a unique choice for N was difficult. For $N = 4$, the chi-squared statistic was equal to 16.05, and the multi-epoch model was weakly rejected (df = 8, $P = 0.041$). For $N = 5$, the statistic was equal to 15.00, and the test was not significant ($P = 0.059$).

MLEs for $N = 7$ and $N = 8$ reached values close to those obtained for $N = 4, 5$. Models with $N = 7$ and $N = 8$ predicted the frequency of singleton species less accurately than with $N = 5$, but the deeper models provided a better fit to the mean observed value and to the tail of the observed distribution (Table 2). For $N = 8$, the multi-epoch model predicted an average number of 9.797 species per genera (for 9.101 observed species), and the probability to have more that 100 species in a genus was 0.9% for an observed value of 0.85%. Thus larger number of epochs predicted the tail of the observed distribution more accurately that smaller N's.

Table 2: Observed and predicted values for the reptile data.

	Observed	$N = 5$	$N = 6$	$N = 7$	$N = 8$
One species	0.260	0.240	0.232	0.225	0.220
More than 10 species	0.208	0.197	0.194	0.196	0.194
More than 100 species	0.008	0.013	0.011	0.010	0.009
Average number of species	9.101	13.068	11.193	10.457	9.797

Observed: empirical values, $N = k$: predicted values for the k-epoch model.

5. Discussion

Hollow curves are one of the most frequently observed patterns in ecology and evolution. These curves, describing the distribution of species occurring within a community, at a trophic or at a taxonomic level, exhibit a general shape for which many species are rare and few species are abundant. Which models provide the best fit to the data, and the resulting implications for the mechanistic processes structuring the data have been an active field of investigation since the discovery of these curves. The discrete Yule distribution is one of those models, and it provides a useful null-model for testing hypotheses about diversity (Yule, 1925; Mooers and Heard, 1997; Nee, 2003, 2006).

Since Yule's contribution, models attempting to explain the causes of the hollow curve in species abundance or specie/genus distributions have proliferated to a very large degree. Yule's paper initiated a series of works on birth-death processes which yield similar distributions (Kendall, 1948; Raup, 1973; Foote *et al.*, 1999; Aldous, 2001). In their review of species abundance distributions, McGill *et al.* (2007) identified five families of models with over forty members. For those distributions, the first theories attempting to explain mechanisms underlying the curve used a stick breaking analogy of niche partitioning (Motomura, 1932). The stick breaking model likely inspired further works on splitting tree distributions (Aldous, 2001). Fisher *et al.* (1943) argued for a logseries distribu-

tion as the limit of a Poisson sampling process, and Kendall (1948) derived the logseries from branching models. However, empirical data show lack of fit to the Yule and other distributions, including hyperbolic (Chamberlin, 1924), logseries (Fisher *et al.*, 1943), broken stick (Dial and Marzluff, 1999) distributions. The statistical similarities of hollow curve distributions even led to the hypothesis that they were due not to mechanisms, but rather to pure randomness in combination with the branching nature of speciation and extinction (Bokma *et al.*, 2013).

Partly because of the unsatisfactory fit of theoretical distributions to empirical data on species over taxa, alternative models of the shape of phylogenetic clades have been proposed, differing from Yule's model substantially (Aldous, 2001; Blum and François, 2006). But it has also been suggested that the observed hollow curve distributions are affected by the criteria used by biologists to define taxa, which may not reflect evolutionary history (Scotland and Sanderson, 2004). By introducing a multi-epoch model for the observed data, the present study is closer to this last vein of thought, reconsidering the evolutionary time scale of a genus definition. The multi-epoch model is a natural extension of the discrete Yule distribution. The number of evolutionary epochs, N, corresponds to the tree height parameter. Higher trees have lower diversification rates in their more ancient epochs, and may capture extinction of ancient species in a better way than constant rate models. Compared to the discrete Yule distribution, the topology of the underlying tree is left unchanged, only the date of the root is reconsidered. In the discrete Yule model, the date of origin is a random variable having an exponential distribution of rate one. The multi-epoch model birth data corresponds to a maximum of N independent copies of this random variable. For reptiles, the fact that $N = 4-8$ copies fit the observed frequencies better than a single copy could reflect uncertainty in the definition of a genus not captured by the single copy.

The main contribution of the present study was to describe mathematical properties of the multi-epoch model. To cope with the combinatorics, an embedding of the urn scheme in a continuous branching process was introduced.

The method could likely be extended to other urn processes with consideration of births and deaths. The distribution of the number of species in birth and death processes would yield less explicit representations of probability distributions than those obtained from pure birth processes (Lansky *et al.*, 2014). In addition, Pòlya's urn theory most often consider the equilibrium distributions of tenable urns – a remarkable exception is a gunfight model studied by Kingman (1999). The models considered here do not satisfy the tenability condition (Mahmoud, 2008), and are less amenable to analysis by standard combinatorial techniques. Nevetheless, we feel that integral representations could also be obtained in complexified models, and open a future avenue of research on hollow curve distributions.

Acknowledgements. This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

Appendix

This appendix provides analytic arguments for the proofs of Theorems 1-4, given for sake of completness. Let X_N be a random variable having a Yule(N, ρ) distribution.

Proof of Theorem 1. The proof uses an embedding of the urn scheme into a pure-birth branching process. According to the description of the urn process, the tree has N independent epochs of duration T_1, ..., T_N, having exponential distribution of rate one, during which the speciation rates are λ, $\lambda/2$, $\cdots$, λ/N, respectively.

The urn process is equivalently described by a rescaled version of the tree having N independent epochs of duration T_1, $T_2/2$, ... ,T_N/N, during which the speciation rates are equal to λ. In this representation, the random variable T_k/k is exponentially distributed with rate k, for each k. Following Yule (1925), the probability of observing n extant species after a period T is equal to

$$p(n\,|N,\rho) = \mathbb{P}(X_N = n\,|N,\rho) = \mathbb{E}[e^{-\lambda T}(1-e^{-\lambda T})^{n-1}]\,, \quad n \geq 1.$$

In the rescaled model, the height of the tree, T, is equal to $T = T_1 + T_2/2 + \cdots + T_N/N$. According to this result, the above probabilities are equal to

$$p(n\,|N,\rho) = \int_0^\infty e^{-\lambda t}(1-e^{-\lambda t})^{n-1} p_{T_1+T_2/2+\cdots+T_N/N}(t)dt\,, \quad n \geq 1.$$

Using the loss-of-memory property of the exponential distribution, one can show that T has the same distribution as the maximum of the N variables $T_1, T_2, \cdots, T_N$ (Ross, 2013). For all $n \geq 1$, the probability of observing n extant species is thus equal to

$$p(n\,|N,\rho) = N\int_0^\infty e^{-\lambda t}(1-e^{-\lambda t})^{n-1} e^{-t}(1-e^{-t})^{N-1}dt\,, \quad n \geq 1.$$

The proof of Theorem 1 follows from the change of variable $u = e^{-\lambda t}$ in the above integrals. □

Proof of Theorem 2. Theorem 2 follows from the integral representation of the Yule(N,ρ) distribution and straightforward calculus. The expected value $\mathbb{E}[X_N|N,\rho]$ is defined as

$$\mathbb{E}[X_N|N,\rho] = N\rho\int_0^1 \left(\sum_{n=1}^\infty nu(1-u)^{n-1}\right) u^{\rho-1}(1-u^\rho)^{N-1}dt\,.$$

Recognizing the mean of a geometric distribution and changing variable $u = v^\lambda$ ($\lambda = 1/\rho$), the integral rewrites as

$$\mathbb{E}[X_N|N,\rho] = N\int_0^1 v^{-\lambda}(1-v)^{N-1}dv\,.$$

The result is equal to

$$\mathbb{E}[X_N|N,\rho] = \infty, \qquad \text{if } \lambda \geq 1\,,\ (\rho \leq 1)\,,$$

and

$$\mathbb{E}[X_N|N,\rho] = N\,B(1-\lambda,N)\,,\quad \text{if } \lambda<1\,,\ (\rho>1)\,.$$

□

Proof of Theorem 3. The proof of Theorem 3 is based on a version of Watson's lemma (Watson, 1922). The lemma considers the function $h(x)$ defined by

$$h(x) = x^\alpha \sum_{n=0}^{\infty} c_n x^n\,,\quad \alpha>0,$$

for $x>0$, close to zero, and $g(x)=\sum_{n=0}^{\infty} c_n x^n$ a real analytic function. Then there is an asymptotic equivalent for the Laplace transform of $h(x)/x$

$$\int_0^\infty e^{-nx} h(x)\frac{dx}{x} \sim \frac{\Gamma(\alpha)c_0}{t^\alpha}\,,\quad n\to\infty.$$

After the change of variable $1-u=e^{-x}$ in the integrals defining $p(n\,|N,\rho)$, one has

$$p(n\,|N,\rho) = N\rho\int_0^\infty e^{-nx}(1-e^{-x})^\rho\left(1-(1-e^{-x})^\rho\right)^{N-1}dx\,,\quad n\geq 1\,.$$

The function $h(x)$ in Watson's lemma can thus be obtained as

$$h(x) = x(1-e^{-x})^\rho\left(1-(1-e^{-x})^\rho\right)^{N-1} = x^{\rho+1}\left(1-\rho\frac{x}{2}+\cdots\right)\,.$$

The coefficients α and c_0 in Watson's lemma can be identified as $\alpha=\rho+1$ and $c_0=1$. The result for the cumulative distribution function can be obtained with similar arguments. □

Proof of Theorem 4. For $\rho=1$, one has

$$\mathbb{P}(X_N=n\,|N,\rho=1) = N\int_0^1 (1-u)^{N-1}u(1-u)^{n-1}du\,,\quad n\geq 1\,.$$

Then, for all $n\geq 0$, one has

$$\mathbb{P}(X_N>n\,|N,\rho=1) = N\int_0^1 (1-u)^{N-1}(1-u)^{n}du\,.$$

and

$$\mathbb{P}(X_N > n \,|N, \rho = 1) = NB(1, N + n) = \frac{N}{N + n}\,.$$

Thus one obtains

$$\mathbb{P}(X_N > nN \,|N, \rho = 1) = \frac{1}{1 + n}\,,$$

and

$$\mathbb{P}(X_N/N > n \,|N, \rho = 1) = B(1, 1 + n) = \mathbb{P}(X_1 > n \,|N = 1, \rho = 1)\,.$$

□

References

Abramowitz, M., Stegun, I. A. 1970. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Vol. 55. Dover publications, 1970.

Aldous, D. J., 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to Today. Stat. Sci. 16, 23-34.

Athreya, K.B., Karlin, S., 1968. Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. Ann. Math. Statist. 39, 1801-1817.

Athreya, K.B., Ney, P.E., 1972. Branching Processes. Springer-Verlag, Berlin.

Blum, M. G.,B., François, O., 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55(4), 685-691.

Bokma, F., Baek, S. K., Minnhagen, P., 2013. 50 years of inordinate fondness. Syst. Biol. 63(2), 251-256.

Chamberlin, J.C., 1924. Concerning the hollow curve of distribution. Am. Nat. 58, 350-374.

Chu, J., Adami, C., 1999. A simple explanation for taxon abundance patterns. Proc. Natl. Acad. Sci. U.S.A. 96(26), 15017-15019.

Chung, K.H., Cox, R.A.K., 1994. A stochastic model of superstardom: An application of the Yule distribution. Rev. Econ. Stat. 76(4), 771-775.

Crane, H., 2016. The ubiquitous Ewens sampling formula. Stat. Sci. 31(1), 1-19.

Darwin, C., 1859. On the Origin of Species, Murray, London, UK.

Dial K.P., Marzluff J.M., 1989. Nonrandom diversification within taxonomic assemblages. Syst. Zool. 38, 26-37.

Fisher, R. A., A. S. Corbet, Williams, C. B., 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. 12, 42-58.

Foote M., Hunter J.P., Janis C.M., Sepkoski, J.J.J., 1999. Evolutionary and preservational constraints on origins of biological groups: divergence times of eutherian mammals. Science 283, 1310-1314.

Hoppe, F.M., 1987. The sampling theory of neutral alleles and an urn model in population genetics. J. Math. Biol. 25(2), 123-159.

Johnson, N.L., Kotz, S., 1969. Distributions in Statistics – Discrete Distributions. John Wiley and Sons, New York.

Kendall, D.G., 1948. On the generalized birth-and-death process. Ann. Math. Stat. 19, 1-15.

Kingman, J. F. C., 1999. Martingales in the OK Corral. B. Lond. Math. Soc. 31(5), 601-606.

Lansky, P., Polito, F., Sacerdote, L., 2014. The role of detachment of in-links in scale-free networks. J. Phys. A: Math. Theor. 47(34), 345002.

Mahmoud, H., 2008. Pòlya Urn Models. Chapman and Hall/CRC.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J. L., He, F., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B. A., Ostling, A., Soykan, C. U., Ugland, K. I., P. White, E. P., 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol. Lett. 10(10), 995-1015.

Mooers, A.O., Heard, S.B., 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol 72(1), 31-54.

Motomura, I., 1932. On the statistical treatment of communities. Zool. Mag. 44, 379-383.

Nee, S., 2003. The unified phenomenological theory of biodiversity. In: Macroecology: Concepts and Consequences(eds Blackburn,T.M. and Gaston, K.J.). Blackwell Science, Oxford, pp. 31-44.

Nee, S., 2006. Birth-death models in macroevolution. Annu. Rev. Ecol. Evol. Syst. 37, 1-17.

Newman, M.E.J., 2005. Power laws, Pareto distributions and Zipf's law. Contemp. Phys. 46(5), 323-351.

Pemantle, R., 2007. A survey of random processes with reinforcement. Probab. Surv. 4, 1-79.

Raup, D. M., Gould, S. J., Schopf, T. J., Simberloff, D. S., 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol 81(5), 525-542.

Raup, D. M., 1985. Mathematical models of cladogenesis. Paleobiology 11(1), 42-52.

Ross, S. M., 2013. Applied Probability Models with Optimization Applications. Dover Publications, New York, USA.

Scotland, R.W., Sanderson, M.J., 2004. The significance of few versus many in the tree of life. Science 303, 643.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika* 42(4):425-440.

Uetz, P. (2000) How many reptiles species? Herpetol. Rev. 31(1), 13-15.

Watson, G.N., 1922. Treatise on the Theory of Bessel Functions. Cambridge University Press, Cambridge, UK.

Willis, J. C., 1922. Age and Area: A Study in Geographical Distribution and Origin of Species. Cambridge University Press, Cambridge, UK.

Yule, G.U., 1925. A Mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B Biol Sci 213 (402-410), 21-23.

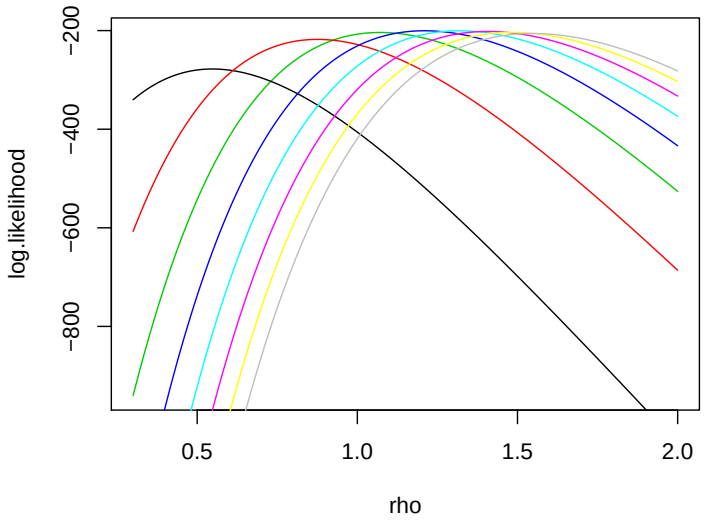


Figure S1. Log-likelihood functions for the reptile data and N-epoch models. The curve for $N = 1$ has black color (left), and the curve for $N = k$ is the kth curve from the left ($N = 8$ at the right, grey color).