Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today

David J. Aldous
Department of Statistics
367 Evans Hall
U.C. Berkeley CA 94720
aldous@stat.berkeley.edu

Abstract

Yule (1924) observed that distributions of number of species per genus were typically long-tailed, and proposed a stochastic model to fit these data. Modern taxonomists often prefer to represent relationships between species via phylogenetic trees; the counterpart to Yule's observation is that actual reconstructed trees look surprisingly unbalanced. The imbalance can readily be seen via a scatter diagram of the sizes of clades involved in the splits of published large phylogenetic trees. Attempting stochastic modeling leads to two puzzles. First, two somewhat opposite possible biological descriptions of what dominates the macroevolutionary process (adaptive radiation; "neutral" evolution) lead to exactly the same mathematical model (*Markov* or *Yule* or *coalescent*). Second, neither this nor any other simple stochastic model predicts the observed pattern of imbalance. This essay represents a probabilist's musings on these puzzles, complementing the more detailed survey of biological literature by Mooers and Heard (1997).

Short title. Modeling and describing phylogeny

Key words and phrases. Descriptive statistics, phylogenetic tree, stochastic model, tree balance, Yule process.

Readers determined to get straight to the main point of this essay (phylogenetic tree imbalance) may skip to section 3, but we hope that the historical context provided in sections 1 and 2 will be helpful to most readers.

1 Yule's 1924 paper

Textbooks on introductory stochastic processes (e.g. Karlin and Taylor (1975) sec. 4.1, Ross (1983) sec. 5.3, Lawler (1995) sec. 3.3) often have a paragraph like the following.

> A population starts at time 0 with one individual. As time increases, individuals may give birth to a new individual, the chance of any particular individual giving birth during time $[t, t+dt]$ being $\lambda\, dt$. This is called the *linear birth process* or *Yule process.*

As often happens, textbooks fail to mention the original motivation for a mathematical innovation, and in this case there is an interesting story.

TABLE 1

Snakes: observed and calculated numbers of genera of each size. (Table VII of Yule (1924)).

Number of species in genus	Number of genera: Observed	Number of genera: Calculated
1	131	130.9
2	35	47.2
3	28	25.2
4	17	16.0
5	16	11.2
6	9	8.3
7	8	6.5
8	8	5.2
9 to 11	13	11.1
12 to 14	3	7.2
15 to 20	7	8.8
21 to 34	14	9.2
35 upwards	4	6.2
Total	293	293.0

Yule (1924) had data on the number of species in genera (recall that *genus* is the taxonomic rank immediately above *species*: a genus consists of a number of closely related species) for various biological groups. Here is one of his tables (the data are of course the *observed* column). The immediately striking feature of this data (confirmed by other such data sets) is that the distribution is long-tailed and that the most frequent number of species is 1. How can this be explained? Of course there is subjectivity in how a taxonomist is to judge "closely related" – how broadly to cast the net of a single genus – and it would be unreasonable to expect consistency of judgement between widely different families, but one can hope for some local consistency. Thus what Yule was looking for was a one-parameter family of distributions with which to compare data, the one parameter reflecting in part the judgements of "closely related" by taxonomists in a particular area of biology. Instead of just invoking some mathematically convenient family, Yule sought to derive distributions from some not-too-implausible model of evolution. Here's what he did. Assume

(1) a genus starts with a single species; new species appear according to the (Yule) process above, with parameter λ, and all these species are in the same genus;

(2) separately, from within each genus a new species of a novel genus appears, at constant rate μ, and thereafter the new genus behaves as in (1).

A simple (now textbook) calculation with the Yule process shows that the number $N(\lambda, t)$ of species in a genus at time t after its first appearance has geometric distribution with mean $e^{\lambda t}$:

$$P(N(\lambda, t) = n) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}, \ n = 1, 2, 3, \dots .$$

From assumption (2) the number of genera will grow exponentially at rate μ and so, in picking a random genus extant today, the time since its first appearance will have exponential(μ) distribution. Thus the distribution of number N of species in a random genus will be

$$p(n) = \int_0^\infty \mu e^{-\mu t} \ e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} \ dt$$

and setting $\rho = \lambda/\mu$ we recognize the integral as a beta integral and can calculate (Yule (1924) page 39)

$$P(N_\rho = n) = \frac{\Gamma(1 + \rho^{-1})}{\rho} \ \frac{\Gamma(n)}{\Gamma(n + 1 + \rho^{-1})}, \ n = 1, 2, 3, \dots . \qquad (1)$$

So this is the family (now called the *Yule distribution*: Kotz et al (1989)) that Yule devised to fit his data. The snake data-set reproduced in table 1 has the worst fit of the four data-sets he considered, and he writes (p. 58)

> I think it must be admitted that the formula given is capable of representing the facts with considerable precision, more closely indeed than we have any right to expect.

Mathematically, it is clear from the formula (1) that $P(N_\rho = n)$ is decreasing (and so maximized by $n = 1$) and asymptotically proportional to $n^{-1-\rho^{-1}}$ (and so long-tailed for suitable ρ).

2 Fast forward to today

In modern jargon Yule was modeling *macroevolution*, that is evolution at the level of species as opposed to within-species changes. While there has been sporadic interest in stochastic modeling of macroevolution since Yule's era, it has not become the kind of "standard subject" to which textbooks are devoted. Before focusing on our particular topic (phylogenetic tree balance) let us provide more context by briefly reviewing four related areas which *have* become standard subjects.

(a) *Population genetics.* To quote Maddox (1998, p. 248)

> (population genetics is) one of the few uses of mathematics in biology to which all biologists are reconciled, often with unaccustomed enthusiasm.

Ewens (1979) is the standard reference. Two aspects of the theory are relevant to this essay. Suppose $f_1, f_2, f_3 \ldots$ are frequencies of alleles at a locus, so that the parameter $F = \sum_i f_i^2$ measures the genetic diversity or lack of diversity at that locus (recall *alleles* are possible forms of a gene; a *locus* is a position on a chromosome.) To study whether observed frequencies provide evidence of selective advantage for some alleles, one needs a null model which predicts values for the f_i in terms of F under the neutral (non-selective) hypothesis, and there is a standard model which leads to predictions called the *Ewens sampling formula* (Ewens 1972, 1990; Kingman 1980). The details (constant mutation rate; mutations produce novel alleles) need not concern us, but the conceptual point (cf. section A.1) is that there does exist an accepted mathematical model for "what should happen just by chance" in population genetics.

A second aspect is a stochastic process called *the coalescent* (Kingman 1982, Tavaré 1984) which plays a central role in modern mathematical population genetics. In particular, sample n alleles from the current population by sampling n individuals, and trace back their line of descent until their most recent common ancestor. Under weak assumptions on the historical population process, the resulting family tree converges (as population size tend to infinity, with n fixed) to a particular *coalescent tree* on n leaves, and this stochastic model will reappear in section 5.1 as a model for phylogenetic trees.

(b) *Branching processes.* On can view the Yule process as the simplest continuous-time branching process. There is a huge mathematical theory of random branching processes, represented by texts such as Harris (1963), Athreya and Ney (1972), Jagers (1975) and Asmussen and Hering (1983). Abstractly, a branching process is equivalent to a random tree, though the bulk of the branching process literature has a different focus from the random tree literature. Recent research in stochastic modeling of phylogenetic trees (see section 5.2) implicitly uses more elaborate branching Markov process models.

(c) *Biological systematics and phylogenetic trees.* The classical Linnaean hierarchy (originally *species, genus, order, class, kingdom* but subsequently extended to many more ranks) remains widely used in practice, but modern systematic biologists regard it as conceptually preferable to describe relationships between species via phylogenetic trees, this being more informative and less arbitrary than the verbal hierarchy. Figure 1 shows an artificial phylogenetic tree on 11 species.

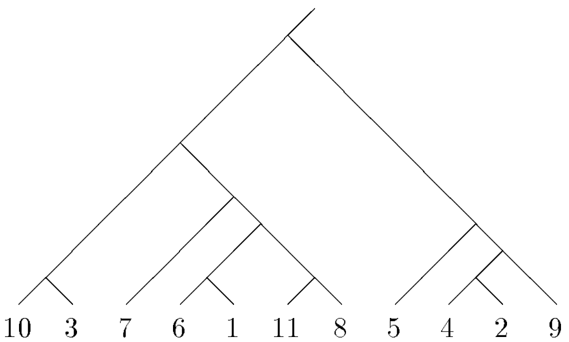

FIG. 1. *A cladogram on* 11 *species.*

This type of tree, where only the combinatorial structure is asserted, is more precisely called a *cladogram*: see Eldredge and Cracraft (1980) for biological discussion. As useful terminology, a *clade* (or *monophyletic* group) is a set of species consisting of all extant descendants of some ancestral species. Thus if the 11 species in figure 1 form a clade then the subset $\{7, 6, 1, 11, 8\}$ is a subclade but the complementary subset $\{10, 3, 5, 4, 6, 2\}$ is not. Mathematically, a subclade can be identified with an edge of the tree.

(d) *Computational algorithms for reconstructing phylogenetic trees from molecular data.* Over the last 20 years, the traditional methods of taxonomy based on physical morphology have been supplemented by methods based on molecular biology, and these more quantitative methods use computer algorithms to produce some "best fit" tree to given data. The mathematical side of this work can be found under *Math Reviews* classification 92B10. Holmes (1998) gives an overview aimed at statisticians. The actual trees reconstructed by biologists can be found in journals such as *Molecular Biology and Evolution*, *Molecular Phylogenetics and Evolution*, *Systematic Biology*, *Systematic Zoology*, *Systematic Botany*, *Evolution* and *Cladistics*. Moreover there is an online database called TreeBASE (www.herbaria.harvard.edu/treebase) intended as a repository of phylogenetic trees and currently containing over 1,000 trees. Let us (arbitrarily) call trees *small* if the number of taxa is less than 10, *medium* if 10 - 100, and *large* if over 100. Currently published trees are almost all medium or small.

3 Questions suggested by today's data

3.1 Phylogenetic trees

Reconstructing phylogenetic trees is a large-scale project, within which many challenging technical statistical questions arise. However, let me emphasize that my concern in this essay is not with technical issues in reconstruction but with the conceptual interpretation of the results. As an analogy, taking the U.S. census is a large-scale project in which technical statistical questions arise (Breiman 1994); but one can also just take published census data at face value and proceed to describe the demographic changes they reveal. So let us assume that published phylogenetic trees are broadly accurate and ask, in the spirit of Yule, what questions do they raise about patterns of macroevolution.

Of course, the first thing to do is to look at data. The figure shows a typical tree which can be viewed and downloaded from TreeBASE.

FIG. 2. *A typical tree from TreeBASE: phylogeny of 25 species of seals (and two outside species, at left), figure 5A.1 of Bininda-Emonds and Russell (1996).*

Here's one way to study the "shape" of such a tree. As mentioned before, each internal edge of a tree specifies a clade. So each branchpoint (internal vertex) of a tree specifies a split of a parent clade into two daughter clades, and we can record the sizes

$(m, i) =$ (size of parent clade, size of smaller daughter clade).

So $1 \leq i \leq m/2$, because the size of the parent clade is the sum of the sizes of the two daughter clades. Before looking at actual trees one might anticipate roughly even splits (i around $m/2$), or uniform random splits (i roughly uniform on $[1, m/2]$), but actual phylogenetic trees show a strong tendency for the smaller daughter clade to be much smaller than the parent clade. See table 2 for the splits in the tree from figure 2.

TABLE 2

Size of splits with parent size ≥ 6 *in the tree of figure 2.*

size of parent clade	size of smaller daughter clade
25	1
24	1
23	1
22	1
21	2
19	9
10	1
9	2
9	1
8	1
7	1
7	1
6	3
6	1

After a moment's thought one sees that this "imbalance" feature of trees is exactly the counterpart of Yule's observation that, in the traditional hierarchical taxonomy, number of species per genus has mode 1 and a long-tailed distribution. To see why, suppose that in the setting of figure 2 one wished to assign species to genera so that each genus is a clade. Then one is forced to assign each small clade branching off from the top of the tree to a separate genus, so that (cf. table 2) one might finish with genera of sizes

$\{1, 1, 1, 1, 2, \ldots\ldots\}$.

The central observation (amplified in section 4.1) is that this type of imbalance is pervasive in published phylogenetic trees. So what does it signify? One could start by asking for a biological explanation of the observed imbalance, but to a statistician that begs the question: maybe imbalance is just what one would expect to happen "by chance" and so requires no biological explanation. That begs the further question: is there a canonical stochastic model to say what trees would occur "just by chance"? But unfortunately (see section 5.1) it seems hard to justify any particular model of "just by chance". It seems to me better to re-order the questions as follows

1. What is a useful way to describe balance/imbalance in a general phylogenetic tree?

2. Is there some particular region of the balance - imbalance spectrum containing most actual phylogenetic trees?

3. If so, is there some mathematically simple and biologically plausible stochastic model for phylogenetic trees whose realizations mimic actual trees?

This ordering deliberately downplays the stochastic modeling aspect, because it is possible to be philosophically opposed to the whole idea of stochastic models of macroevolution, while it is scarcely possible to be philosophically opposed to descriptive statistics! Asking such questions is hardly original, of course. We are fortunate that Mooers and Heard (1997) have written an accessible, detailed survey of the technical biological literature on the subject of phylogenetic tree shape, and discussion of whether conclusions concerning macroevolutionary process may legitimately be inferred from tree shape. That survey, to which we shall refer often, covers substantially more topics than does this essay. However, our particular three questions have a distinctly different emphasis from much of the biology literature, which tends to start with speculation on biological mechanism underlying macroevolution (sections 5.2 and 5.3).

3.2 The simplest stochastic models

Though we intend to downplay stochastic models and emphasize descriptive statistics, it is useful to mention here the two simplest models for n-species cladograms. The Yule process itself, run until there are n species, defines a random n-species phylogenetic tree: following biologist's terminology, call this tree (regarded as a cladogram) the *Markov model.* Under the Markov model, different possible trees are not equally likely (when $n \geq 4$), so for comparison one can also consider the uniform distribution. Again following biologist's terminology, we call the uniform distribution on cladograms the *PDA model* (Proportional to Different Arrangements). It turns out (section A.1) that both models are particular cases (Markov is $\beta = 0$; PDA is $\beta = -1.5$) of a one-parameter *beta-splitting family* of distributions on cladograms, whose parameter varies over the range $-2 < \beta \leq \infty$, with increasing β corresponding to greater balance.

4 Phylogenetic tree balance

4.1 Visualizing balance via scatter diagrams

A binary tree contains a set of splits

(m, i) = (size of parent clade, size of the smaller daughter clade)

which can simply be plotted as a scatter diagram. My proposal for studying tree balance is that, given a large phylogenetic tree, one should estimate (cf. nonlinear regression) the median size of the smaller daughter clade as a function of the size of the parent clade, and use this function as a descriptor of balance/imbalance in the given tree. Section 4.4 describes why this should be preferable to simply using some numerical summary statistic. Carrying through this program turns out to be less easy than anticipated, because I have not located any data-set satisfying all the desired criteria, so what is proposed here may be regarded as a program for describing future large trees. Figures 3 – 5, described below in more detail, illustrate three data-sets. It is convenient to make a log-log plot and to ignore small parent clades. Rather then estimating medians, as proposed above, the scatter diagrams show lines giving the approximate median value of the size of the smaller daughter clade predicted by the beta-splitting model, for several values of β, in particular the values for the Markov ($\beta = 0$) and PDA ($\beta = -1.5$) models. In other words, if the model were true then the scatter diagram for a typical tree would have about half the points above the line and half below the line, throughout the range.

Figure 3 shows the scatter diagram for the tree of 475 species of seed plants given by Chase et al. (1993). The figure clearly shows that the tree is less balanced than predicted by the Markov model and more balanced than predicted by the PDA model.

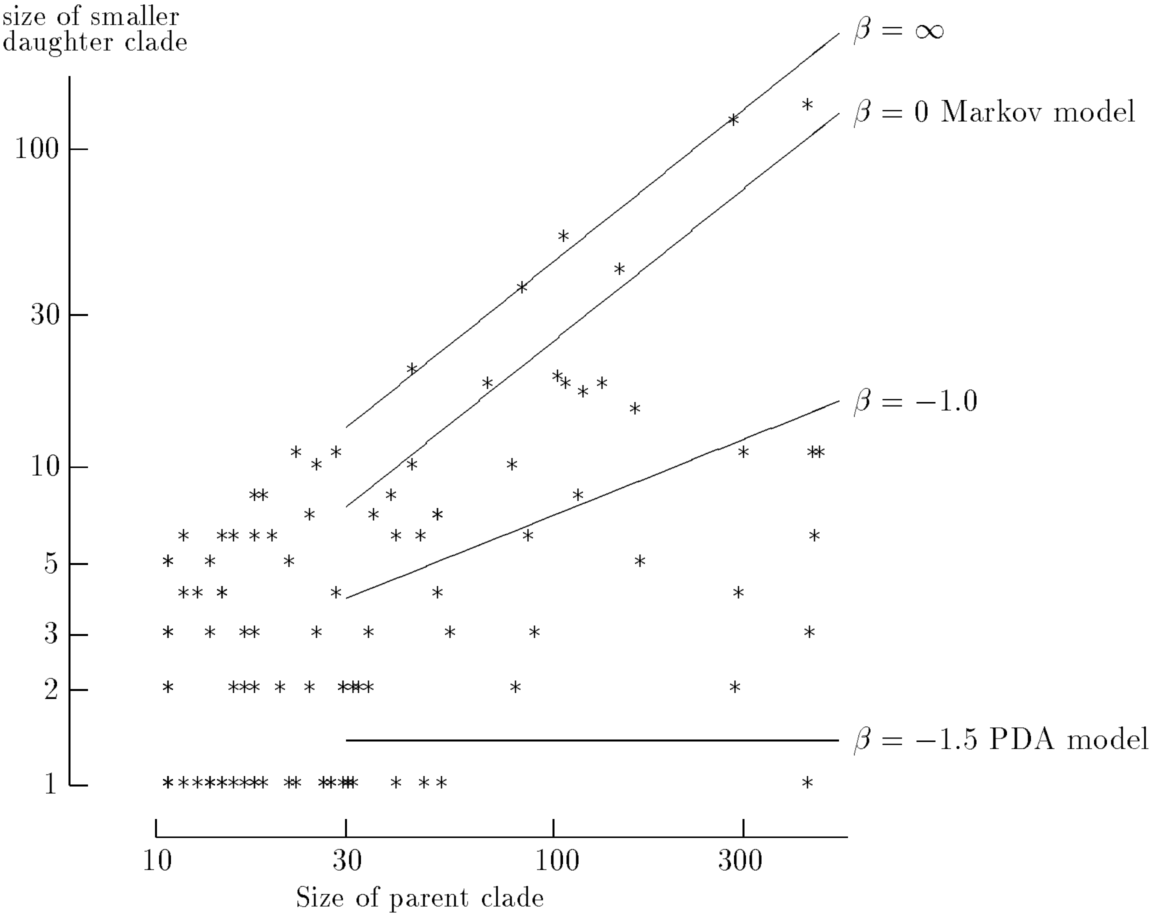

FIG. 3. *Splits in the tree of Chase et al (1993), and approximate median lines for the beta-splitting model. Note the log-log scale.*

As Chase et al (1993) write, their chosen species "include all major lineages" but have "an uneven taxonomic distribution", and this seems typical of published large trees. So one cannot rule out the possibility of selection bias influencing tree balance, as well as the other sources of possible bias discussed in Mooers and Heard (1997). Ideally we would like to study trees which are complete (include all extant taxa), large ($>$ 400 taxa, say), fully resolved (only binary splits), and where the taxa are species (avoiding the possible subjectivity of grouping species into higher taxa). Such a study would show whether the pattern of figure 3 persists. Unfortunately we can find no such trees published in usable form. The largest complete trees we can find have 200 - 300 species and invariably have non-binary splits. For instance, figure 4 shows the scatter diagram for a tree of 268 species of Myodochini given by Harrington (1980), in which the non-binary splits were arbitrarily resolved to binary splits marked by $\circ$.

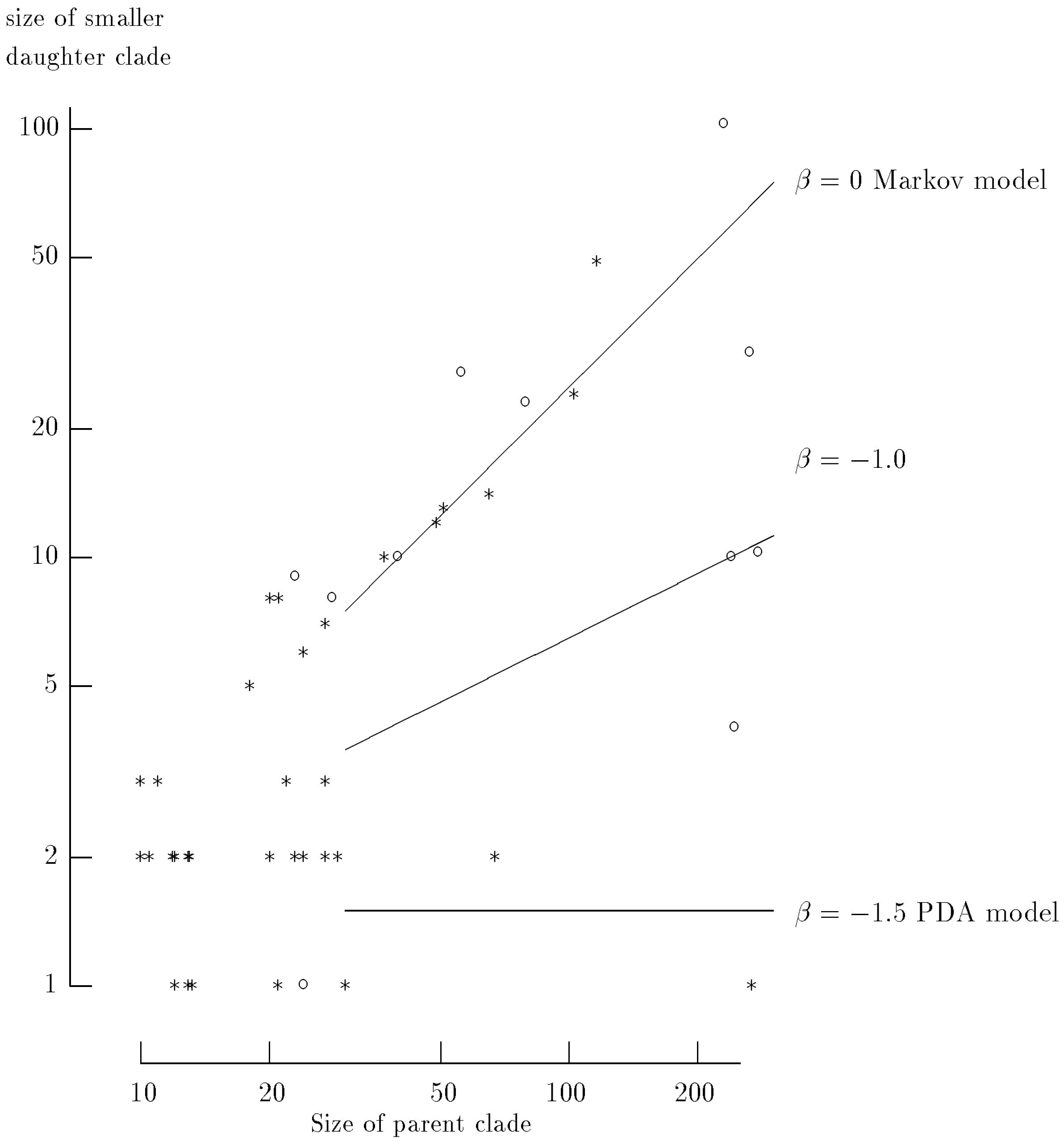

FIG. 4. *Splits in the tree of Harrington (1980).*

As well as the arbitrary binary resolutions, the comparatively few splits of large clades make the correct median line hard to determine – this is typical of trees in this size-range – though again the diagram shows the tree is less balanced than the Markov model predicts and more balanced than the PDA model predicts. To take a different approach, but one with an even greater risk of selection bias, figure 5 shows the scatter diagram corresponding to a list of 30 basal splits of monophyletic lineages given by Guyer and Slowinski (1993, table 1).

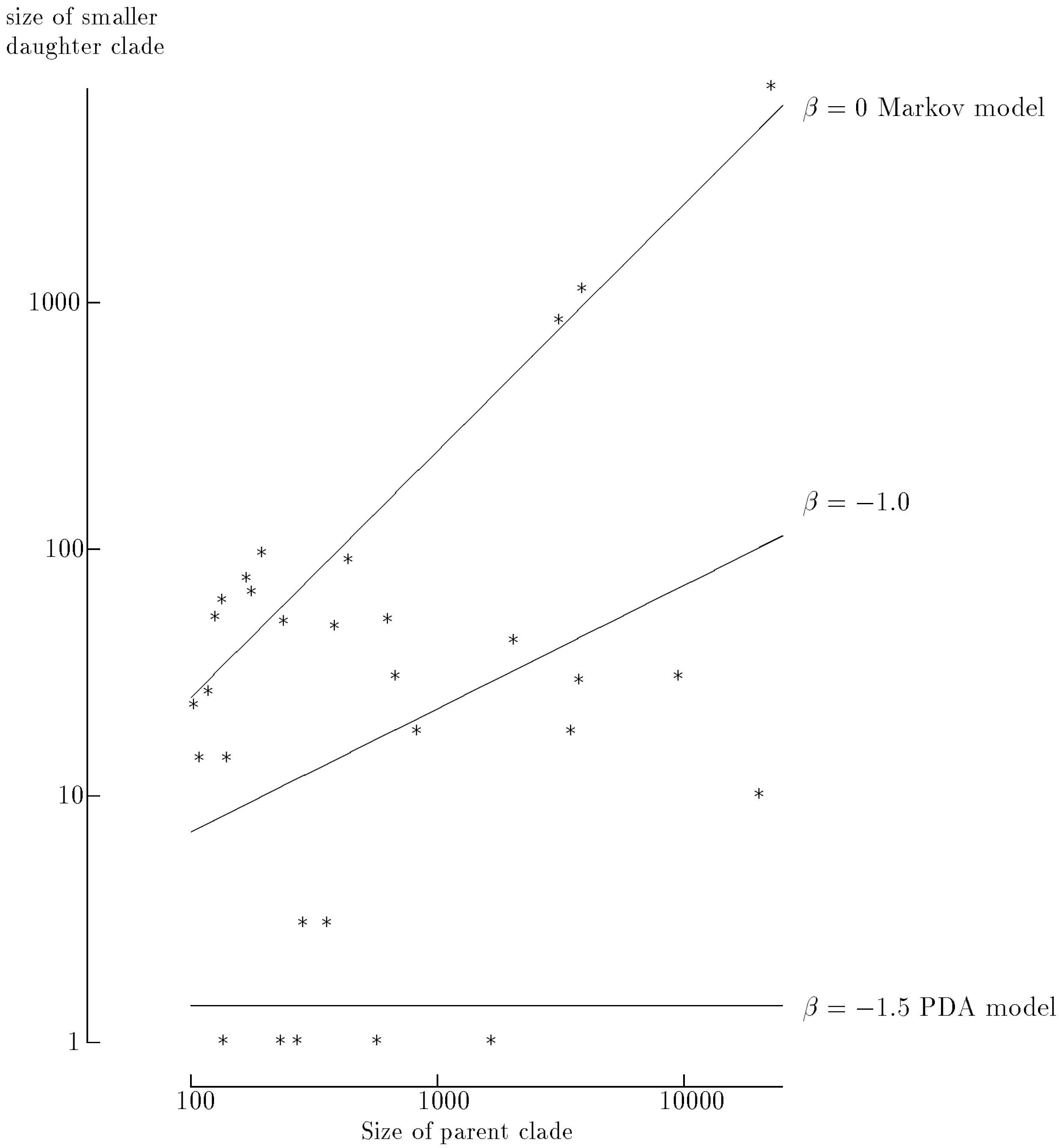

FIG. 5. *The 30 splits in Guyer and Slowinski (1991, table 1).*

These three data-sets present a clear picture: the trees are more balanced than predicted by the PDA model and less balanced than predicted by the Markov model. This is not a new observation: rather, it supports the current view of biologists.

> Estimated phylogenetic trees tend to be more unbalanced than expected under the Markov model, and this tendency is independent of methodological details of the trees' estimation. *Heard (1996).*

4.2 Technical asides

It is easy to raise technical issues in the program above; let me just address two such issues, and refer to Mooers and Heard (1997) for more extensive discussion.

Selection bias. In choosing a group of species to study, a biologist will typically choose a group believed *a priori* to be a clade qualitatively separate from others. In other words, if the true evolutionary tree has a split of a size-45 clade into daughter clades of sizes 15 and 30, then a biologist is likely to choose one of these daughter clades to study, and the study will not record that particular split. Thus in aggregate studies of small or medium trees, selection bias may affect the observed pattern of splits near the root. Looking at individual large trees, without supposing they are necessarily representative, largely avoids this issue.

Systematic bias caused by reconstruction algorithms. No-one believes that published large trees are completely accurate. It has been argued (see Mooers and Heard (1997) p. 39-40) that systematic bias in tree shape is created. Roughly, the argument is that algorithms start by implicitly assuming all possible trees are equally likely. If the correct tree were a typical tree under the Markov model, then a sequence of reconstructed trees based on more and more data would presumably interpolate between a typical PDA tree and the correct typical Markov tree; so if there is not enough data to identify the correct tree then the reconstructed tree will tend to err on the side of a PDA tree, i.e. to be too unbalanced. While such reconstruction error is a legitimate concern, tree imbalance also seen in classical trees based on morphological data constructed "by hand" without computer algorithms. In this setting one can argue that people are biased toward creating too-balanced trees, as providing tidier classifications.

4.3 Studies of small and medium trees

There have been two studies of small trees. Guyer and Slowinski (1991) take clades of size $n = 5$ and Savage (1983) takes $4 \leq n \leq 7$. These two studies reached opposite conclusions on which of the Markov model and the PDA model gives the better fit. So it is hard to interpret their conclusions.

4.4 Balance statistics

Of course, a more classical way to think about tree imbalance is to invent a single numerical summary statistic. Many such statistics have been discussed in the biological literature by Kirkpatrick and Slatkin (1993), Rogers (1996) and Shao and Sokal (1990), and are reviewed in Mooers and Heard (1997) p. 34-35. A common choice of balance statistic is

$$I_c := \binom{n-1}{2}^{-1} \sum_v |s_1(v) - s_2(v)|$$

where $s_1(v)$ and $s_2(v)$ are the sizes of the two daughter clades at a branchpoint v. Though intuitively reasonable, justifying this particular choice via any theoretical statistical criterion seems problematical.

However, our proposal for describing imbalance via median regression seems preferable in two ways. First, such graphical methods are more informative, and more in the modern statistical spirit. Second is the issue (*calibration*) of comparing trees with different numbers n of taxa. A summary statistic allows us to say that one size-n_1 tree is (by a certain criterion) more balanced or less balanced than another size-n_1 tree, but it doesn't allow us to compare directly two trees of different sizes (n_1 and n_2, say). One could standardize statistics with respect to some stochastic model, but that begs the question of which stochastic model to use. In contrast, the scatter diagram – regression method permits a direct comparison of different-sized trees, by superimposing the two scatter diagrams.

4.5 Alternative types of data

This essay focuses on phylogenetic trees of extant species, for the simple reason that most relevant data currently being published are of this form. But note that there are two other possible types of data. First, a tree drawn as in figures 1 and 2 is asserting only a certain combinatorial structure – recall that more precisely, such trees are called *cladograms*. One can envisage a phylogenetic tree having also an explicit time scale, and hence asserting how long ago different lineages diverged. Some reconstruction algorithms seek to construct such "trees with time scale". Working with such data would make a genuine difference, in that one of our conceptual points (that biologically different stories may lead to the same mathematical model; section 5.1) would no longer be true. One can also argue that imbalance in cladograms which reflects short edges in the phylogenetic tree is artificial, and would be correctly downplayed by measures of imbalance based on trees with time scale. On the other hand the accuracy of estimates of divergence times is often questionable. See e.g. Paradis (1998) for references to statistical analysis of such data.

A second approach to studying macroevolution is described in Steven Jay Gould's popular writings (e.g. Gould (1977) page 132; Gould (1989) page 303): use fossil data to estimate numbers of different species (or higher taxa) in existence as a function of time, during some geological era. Fitting birth-and-death type stochastic processes to such data (a natural extension of Yule's idea) has been studied by several authors (Gould et al. 1977, Stanley et al. 1981, Stoyan et al. 1983) and one can regard this as a complementary approach to the study of statistical aspects of macroevolution.

5 Stochastic modeling of macroevolution since Yule

5.1 Two biological pictures

As described in section 1, Yule needed a two-part model to study distribution of species per genus. Ironically, it is simpler to use his idea to model phylogenetic trees, because the Yule process itself, run until there are n species, defines a random n-species phylogenetic tree (this is the *Markov model*). The Yule process can be regarded as the simplest model for what biologists call *adaptive radiation.* In brief, the idea is that the first species with some novel useful feature, a *key innovation* (the first species of bird, for instance) will comparatively rapidly speciate into different species which all preserve the distinguishing innovating feature but which otherwise adapt to different ecological niches.

In a sense adaptive radiation emphasizes selection and fitness. Alternatively one can imagine macroevolution as *neutral* (to borrow the language of population genetics), and model extinctions and speciations as pure chance. In a simple model, take the number n of species in a group to be fixed, and at each step choose at random one species to become extinct and another to speciate (in population genetics, this is the *Moran model* (Ewens 1979, sec. 3.3)). Imagining this process to have continued from the distant past until now, there is a phylogenetic tree on the n extant species, and we can call this model of a random phylogenetic tree the *coalescent model.* This model has a simple "backwards" description. Regard the n species as n "lines of descent". Pick uniformly at random a pair of lines, and merge them into one, giving now $n-1$ different lines of descent. Repeat until there is only 1 line of descent.

At this point it seems that one could discriminate between these two biological possibilities by seeing which model fits the data better. But it turns out that the two models are mathematically identical: that is, they give the same probability distribution on cladograms! (Here the fact that cladograms have only combinatorial structure, not a time scale, becomes relevant.) In fact this equivalence holds much more generally. Specify arbitrarily a sequence $n_0, n_1, \ldots, n_k$ of positive integers changing by ± 1 at each step, representing number of species in successive time periods. Define a stochastic model for speciation and extinction within these constraints as follows.

> (*) A step $n_{j+1} = n_j - 1$ indicates extinction of a uniform random species; a step $n_{j+1} = n_j + 1$ indicates speciation from a uniform random species.

The Markov model is the special case $(n_0, n_1, \ldots, n_k) = (1, 2, \ldots, n)$, and the coalescent model is equivalent to the special case $(n_0, n_1, \ldots, n_k) = (n, n-1, n, n-1, \ldots, n)$. It is easy to see that in the general case the following is true. Take the phylogenetic tree on the current n_k species and assume there is a single common ancestor in the first time period. Then the tree (regarded as a cladogram) has the same distribution as the coalescent tree, and hence as the Markov tree.

While this observation is not new (see e.g. Slowinski and Guyer (1989)), its implications have, in my view, been insufficiently emphasized in the biological literature. One is studying what tree shape says about macroevolution; macroevolution is driven by speciations and extinctions; so one might take for granted that tree shape has some relation to overall rates of speciation and extinction. But it doesn't. That is, within model (*) (which by analogy with population genetics (Ewens 1979 sec 3.4) one might call the *exchangeable* model) the shape of the cladogram on extant species is unaffected by (and hence tells us nothing about) past overall rates of speciation and extinction. While these rates can be studied using other types of data (section 4.5), the equivalence of models makes the observed tree imbalance even more puzzling; what does it signify?

5.2 More elaborate stochastic models of phylogenetic trees

To a modern probabilist, the natural "general model" of macroevolution would be as a branching Markov process (which have been studied in various contexts, e.g. Asmussen and Hering 1983). Suppose
(i) each species has some "type" x;

(ii) a species of type x becomes extinct at some (stochastic) rate $\alpha(x)$;
(iii) a species of type x gives rise to daughter species of types y at (stochastic) rate $\beta(x,y)$;
(iv) a species may change its type according to some specified Markov process.

With so many parameters one could presumably fit any tree; so what special cases are more sensible? Here's one conceptual approach. The general view amongst evolutionary biologists is that, except for mass extinctions and their aftermath, overall numbers of species do not tend to increase or decrease exponentially fast. So in the language of branching processes, models should typically be close to *critical*, in the sense that the overall mean number of daughter species before extinction equals 1. Conceptually, one can ask whether macroevolution is dominated by chance (with only occasional adaptive radiations) or by selection (ongoing replacement of less fit by more fit species). Mathematically, the former possibility could be modeled as the *general neutral model*, the branching Markov process in which for each type x the extinction rate $\alpha(x)$ equals the total speciation rate $\sum_y \beta(x,y)$ (or its continuous-type analog). Such a process is critical, but contrasts with the *exchangeable* model of section 5.1 by allowing the rate to vary between species, which is biologically reasonable (e.g. varying by size of organism). It is not hard to guess heuristically (e.g. by considering an extreme case in which extinction-speciation rates are either very small or very large) that such variation should tend to increase tree imbalance; formalizing that idea in some generality is an interesting mathematical challenge. The biology literature has focussed more on generalizing the Yule model to models with varying speciation rates but without extinction. Heard (1996; see also papers cited therein) uses Monte Carlo simulation to study several such models, and concludes

> As suspected, variation in speciation ... rates among lineages within a clade does tend to produce unbalanced phylogenetic trees, reminiscent of those seen in the systematic literature ... [but] it is clear that estimated trees from the literature correspond to very high, perhaps implausibly high, levels of rate variation.

To mention one of several further models described in Mooers and Heard (1997), Rogers (1996) considers a variation in which new species cannot speciate during an initial time period, and observes this alone may lead to increased imbalance. de Queiroz (1998) critically reviews statistical attempts to detect adaptive radiation by studying sizes of sister clades.

5.3 Biology underlying stochastic models

The mathematical models in sections 5.1 and 5.2 treat speciation and extinction as intrinsically random, rather than being derived from some underlying biological mechanism which one might model instead. The generally accepted *allopatric* theory of speciation asserts that new species typically arise from small, geographically isolated, subpopulations of existing species. But what constitutes "geographical isolation", and what causes it, will vary greatly from one group of species to another, making it difficult to formulate any general biology-motivated mathematical model of speciation (see Mooers and Heard (1997) p. 49-50 for references to models tied to specific kinds of organisms). Various modes of allopatric speciation suggest unbalanced trees. For instance, climatic change may cause the range of a species to shrink, leaving subpopulation in isolated pockets, each having a chance to speciate and all such new species being direct offshoots from the original species. In contrast, the predominant cause of extinctions is more controversial (Raup 1991) and so extinction is seldom modeled except as being intrinsically random. Attempts to devise models in which speciation and extinction are consequences of some explicit underlying mechanism tend to be complex and arbitrary: see for instance Aldous (1995b).

6 Summary

The observed imbalance of published phylogenetic trees remains a puzzle. While it is easy to propose biological mechanisms which might give rise to imbalance, biologists have not reached any consensus on a dominant effect. Many non-mathematical biologists would be dismissive of the whole project, arguing that macroevolutionary history, like human history, is a mosaic of singular events not amenable to mathematical modeling. On the other hand one can look forward to the day when there is largely-accepted phylogenetic tree for most of the millions of extant species, and (in contrast to human history) this is a definite data-set: a priori refusal to subject it to statistical scrutiny seems perverse.

In view of the ever-increasing number of published phylogenetic trees and the size of the literature relating to tree balance (Mooers and Heard (1997) cite almost 150 papers), it is surprising that there has not been any careful large-scale study of tree balance since the early 1990s. We advocate such a study, emphasising descriptive statistics rather than stochastic models. The notion of median regression described in section 4.1 seems one promising way of describing balance in more detail. We would like to see the current summary sentence

> actual trees are more unbalanced than predicted by the Markov model

replaced by a more positive statistical description of empirical tree shape. Such a description would provide a baseline useful for several purposes, e.g. distinguishing particular phylogenetic trees whose shape is different from typical, or assessing qualitative fit of stochastic models.

Biologists' concern with formal goodness of fit tests seems misplaced – surely all these models are better regarded as crude caricatures instead of precise hypotheses. As described in section 5.1, one can distinguish between a conceptual view of macroevolution dominated by adaptive radiation and a view of macroevolution as dominated by chance. Recent stochastic models based on variations of the Yule branching process are in a sense implicitly assuming the adaptive radiation viewpoint; in seeking to distinguish between these views, it seems desirable to study in parallel models assuming overall equilibrium between speciations and extinctions. More speculatively, one might hope to devise models which enable one to quantify the amount of selection during macroevolutionary history required to produce observed tree imbalance.

Acknowledgements. This paper grew from an invited talk at the June 1998 DIMACS Symposium on Estimating Large Scale Phylogenies; I thank the organizers (Tandy Warnow, Junhyong Kim, Ken Rice) of that Symposium. I also thank Susan Holmes for ongoing discussions, and Stephen Heard for helpful correspondence. Referees' comments improved the exposition.

References

ALDOUS, D.J. (1995a). Probability distributions on cladograms. In *Random Discrete Structures* (D.J. Aldous and R. Pemantle, eds.) 1 – 18. volume 76 of *IMA Volumes in Mathematics and its Applications*. Springer-Verlag, Berlin. Available via homepage http://www.stat.berkeley.edu/users/aldous.

ALDOUS, D.J. (1995b). Darwin's log: a toy model of speciation and extinction. *J. Appl. Probab.* **32** 279–295.

ASMUSSEN, S. and HERING, H. (1983). *Branching Processes.* Birkhauser.

ATHREYA, K.B. and NEY, P. (1972). *Branching Processes.* Springer, Berlin.

BININDA-EMONDS, O.R.P. and RUSSELL, A.P. (1996). A morphological perspective on the phylogenetic relationships of the extant phocid seals (Mammalia: Carnivora: Phocidae). *Bonner Zoologische Monographien* **41** 1–256.

BREIMAN, L. (1994). The 1991 census adjustment: undercount or bad data? *Statistical Science* **9** 458–475.

CHASE, M.W. et al (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. *Ann. MO Botanical Garden* **80** 528–580.

DE QUEIROZ, A. (1998) Interpreting sister-group tests of key innovation hypotheses. *Systematic Biology* **47** 710–718.

ELDREDGE, N. and CRACRAFT, J. (1980). *Phylogenetic Patterns and the Evolutionary Process.* Columbia University Press, New York.

EWENS, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87–112.

EWENS, W.J. (1979). *Mathematical Population Genetics.* Springer-Verlag, Berlin.

EWENS, W.J. (1990). Population genetics theory – the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory* (S.

Lessard, ed.) 177–227. Kluwer Academic, Dordrecht.

GOULD, S. J. (1977). *Ever Since Darwin: Reflections in Natural History.* W. W. Norton, New York.

GOULD, S. J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History.* W. W. Norton, New York.

GOULD, S. J. , RAUP, D.M., SEPKOSKI, J.J., SCHOPF, T.J.M. and SIMBERLOFF, D.S. (1977). The shape of evolution: a comparison of real and random clades. *Paleobiology* **3** 23–40.

GUYER, C. and SLOWINSKI, J.B. (1991). Comparisons between observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution* **45** 340–350.

GUYER, C. and SLOWINSKI, J.B. (1993). Adaptive radiation and the topology of large phylogenies. *Evolution* **47** 253–263.

HARRINGTON, B. J. (1980). A generic level revision and cladistic analysis of the Myodochini of the world (Hemiptera, Lygaeidae, Rhyparochrominae). *Bull. Amer. Museum Natural History* **167** 45–166.

HARRIS, T.E. (1963). *The Theory of Branching Processes.* Springer-Verlag, New York.

HEARD, S.B. (1996). Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution* **50** 2141–2148.

HOLMES, S. P. (1998). Phylogenies: An overview. In *Statistics in Genetics* (B. Halloran and S. Geisser, eds.) 81–118. Volume 112 of *IMA Volumes in Mathematics and its applications.* Springer-Verlag, New York.

JAGERS, P. (1975). *Branching Processes with Biological Applications.* Wiley.

KARLIN, S. and TAYLOR, H.M. (1975). *A First Course in Stochastic Processes.* Academic Press, second edition.

KINGMAN, J.F.C. (1980). *Mathematics of Genetic Diversity.* S.I.A.M., Philadelphia PA.

KINGMAN, J.F.C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248.

KIRKPATRICK, M. and SLATKIN, M. (1993). Searching for evolutionary pattern in the shape of a phylogenetic tree. *Evolution* **47** 1171–1181.

KOTZ, S. and JOHNSON, N.L. (1989). Yule distributions. In *Encyclopedia of Statistical Sciences* **9** 191. Wiley, New York.

LAWLER, G. F. (1995). *Introduction to Stochastic Processes.* Chapman and Hall.

MADDOX, J. (1998). *What Remains to be Discovered.* The Free Press, New York.

MOOERS, A.O. and HEARD, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quarterly Rev. Biology* **72** 31–54.

PARADIS, E. (1998). Detecting shifts in diversification rates without fossils. *American Naturalist* **152** 176–187.

RAUP, D.M. (1991). Extinction: Bad Genes or Bad Luck? W.W. Norton, New York.

ROGERS, J.S. (1996). Central moments and probability distributions of three measures of phylogenetic tree balance. *Systematic Biology* **45** 99–110.

ROSS, S. (1983). *Stochastic Processes.* Wiley.

SAVAGE, H.M. (1983). The shape of evolution: Systematic tree topology. *Biological J. Linnean Soc.* **20** 225–244.

SHAO, K. and SOKAL, R.R. (1990). Tree balance. *Systematic Zoology* **39** 266–276.

SLOWINSKI, J.B. and GUYER, C. (1989). Testing the stochasticity of patterns of organisimal diversity: an improved null model. *American Naturalist* **134** 907–921.

STANLEY, S.M., SIGNOR, P.W. III, LIDGARD, S. and KARR, A.F. (1981). Natural clades differ from "random" clades: simulations and analyses. *Paleobiology* **7** 115–127.

STOYAN, D., STOYAN, H. and FIKSEL, T. (1983). Modelling the evolution of the number of genera in animal groups (Yule's problem revisited). *Biometrical J. (Journal of Mathematical Methods in Biosciences)* **25** 443–

451.

TAVARE, S. (1984). Line-of-descent and genealogical processes and their applications in population genetics. *Theoret. Population Biol.* **26** 119–164.

YULE, G.U. (1924). A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. *Philos. Trans. Roy. Soc. London Ser. B* **213** 21–87.

A Appendix

A.1 The beta-splitting model

Instead of devising more elaborate models based on biological considerations, as in section 5.2, one can instead ask a purely mathematical question.

> Is there a convenient one-parameter family of probability distributions on n-taxon trees which interpolates between the completely unbalanced and the completely balanced cases and which includes the PDA model and the Markov model as special cases?

This question was studied in Aldous (1995a), where the *beta-splitting* family was introduced and shown to have the desired properties, except for being less "natural" (either mathematically or biologically) than one would like. The definition is reviewed in the next section. Varying the parameter β covers almost the entire balance/imbalance spectrum. The essential information is presented in table 3, where as in section 4.1 "median split" indicates approximate median size of the smaller daughter clade in a split of a m-taxa parent clade, for large m.

TABLE 3
Aspects of the beta-splitting model.

β	description	median split
-2	completely unbalanced	1
-1.5	PDA model	1.5
-1	unnamed	$\sqrt{m}$
0	Markov model	$m/4$
∞	an almost completely balanced model	$m/2$

Figures 2 – 4 indicate that the $\beta \approx -1$ model gives a better fit to these data-sets than either the Markov or the PDA model.

As analogies to this model, both Yule's distribution (1) for number of species per genus, and the Ewens sampling formula for random allele frequencies are one-parameter families where the parameter has a natural biological interpretation (in (1) ρ relates to the ratio of speciation rate to rate of appearance of new genera; in the Ewens sampling formula the parameter relates to the total number of new alleles arising by mutation in each generation). The beta-splitting family is biologically artificial, in that it does not correspond to any simple description of speciation and extinction forwards in time, and so the parameter β has no *a priori* biological interpretation.

A.2 More about the beta-splitting model

Mathematical motivation for the beta-splitting model and further mathematical properties are given in Aldous (1995a) and will not be repeated here.

We start with a rather general mathematical construction of a random tree on n taxa. Place n "particles" labeled $\{1, 2, \ldots, n\}$ on the unit interval at uniform random positions. Split the interval at a random point chosen from some probability density f. Rescale subintervals to unit length, and repeat recursively on subintervals for as long as subintervals contain at least 2 particles. Associate a tree in the natural way, illustrated in figure 4.

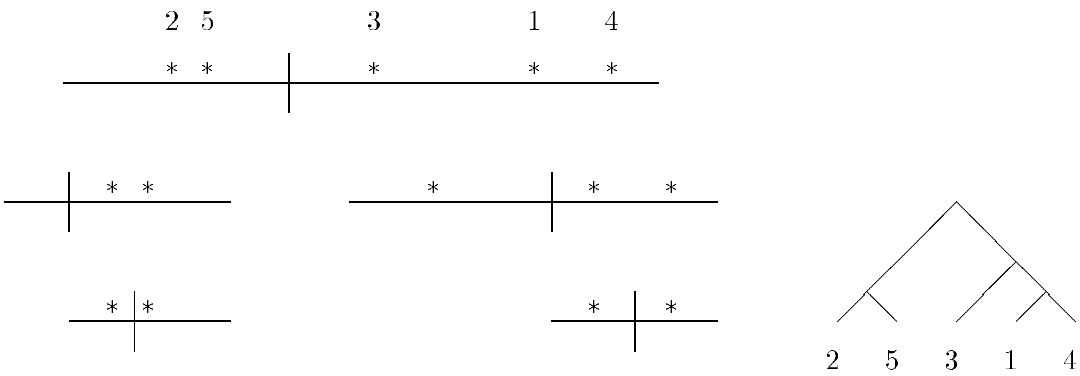

FIG. 6. *A tree construction.*

The chance that the left branch at the root has size i equals

$$a_n^{-1} \binom{n}{i} \int_0^1 x^i (1-x)^{n-i} \, f(x)dx, \quad 1 \leq i \leq n-1 \qquad (2)$$

for normalizing constant

$$a_n = \int_0^1 (1 - x^n - (1-x)^n) f(x)dx$$

The *beta-splitting family*, with parameter $-2 \leq \beta \leq \infty$, is the specialization where we split the unit interval with "density"

$$f(x) \propto x^{\beta}(1-x)^{\beta}.$$

Though this is only a probability density for $\beta > -1$, the definition (2) makes sense for $\beta > -2$, and the limit case $\beta = -2$ is the completely unbalanced tree.

Implicit in (2) is a formula for the distribution of the size I_m of the smaller daughter group in a split of a size-m group. Because the integrals in (2) involve the Gamma (factorial) function, this formula simplifies when β is an integer or an integer $+\frac{1}{2}$. Explicitly (Aldous 1995a sec. 4.1) for $1 \leq i < m/2$

$$\begin{aligned} P(I_m = i) &= \frac{2}{m-1} && \text{(Markov model)} \\ &= \binom{m}{i}\frac{c_i c_{m-i}}{c_m} && \text{(PDA model)} \\ &= \frac{m}{i(m-i)h_{m-1}} && (\beta = -1) \end{aligned}$$

where $c_n = (2n-3)(2n-5)\cdot\cdot 3 \cdot 1$ and $h_{m-1} = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{m-1}$, and where for even m and $i = m/2$ the probabilities are halved. From these exact formulas one deduces the asymptotic formulas for median split given in table 3. Figures 2 – 4 may mislead by suggesting that the median lines are straight. Of course the exact medians are integer-valued – that's why we didn't extend the median lines below size 30 parent clades. However, for the β-values in the figures the lines are indeed asymptotically straight, as suggested by the formulas for median split given in table 3.