Stochastic Models for Phylogenetic Trees on Higher-order Taxa

David Aldous and Maxim Krikun and Lea Popovic

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

July 4, 2006

Abstract

xxx draft in progress, intended for Journal of Mathematical Biology.
1 Introduction

This paper provides some mathematical details of part of a broader project we call “coherent modeling of macroevolution”. A future paper [7] addressed to less mathematically-focused biologists will provide more detailed background, motivation, relation with previous biological literature, and discuss the “bottom line for biology” of such mathematical models. Here we give only a very brief sketch of such matters in section 1.1 and then proceed to outline (section 1.2) the specific results of this paper.

1.1 Background

Stochastic models for time series of species numbers within a clade and for phylogenetic trees of extant species in a clade can be traced back to Yule [12]. Such models treat speciations and extinctions as random (in some way). In studying such models one is not asserting that real macroevolution was purely random; rather, one wishes to compare real data with the predictions of a random model to see what patterns require biological explanation (e.g. adaptive radiations [6]), or to make inference about unobservables (e.g. the time of origin of the primates [11]).

One aspect of this subject is where the data consists of time series or phylogenetic trees on some higher-level taxa (genera or families, say) instead of species. In the fossil record of the distant past it is difficult to resolve specimens to the species level, and the species-level data is liable to be incomplete, so that statistical analysis of time series (relying e.g. on the celebrated compendia of Sepkoski [10]) is in practice done at the level of genera or families. In discussing phylogenies within large extant groups such as birds or mammals, it is impractical to show all species, so one shows trees indicating how major subgroups are related. And the same holds for extinct groups (see e.g. the fascinating tree [9] on dinosaur genera). In looking at such data and asking the basic question – what patterns imply biological effect rather than being consistent with “just chance” – two extra difficulties arise. First, the classification into genera or families involves human judgement which is inevitably at least somewhat subjective. Second, while one could just take genera (say) as entities in themselves and apply species-level models directly to genera [5], this seems conceptually unsatisfactory: genera are sets of species and so, as part of “coherent modeling of macroevolution”, one would like genus-level models to be based upon underlying species-level models. While these difficulties are certainly mentioned in the biological literature, we have not seen any very thorough mathematical study. Our
purposes in this paper are to lay out a conceptual framework for studying such questions, and then to give mathematical analysis of the predictions of a particular probability model.

1.2 The two topics of this paper

The first topic concerns methodology of classification (without any involvement of probability models). Suppose we know the true phylogenetic tree on a clade of species, that is on some founder species \( s \) (typically extinct) and the set of all descendant species (extant and extinct) of \( s \). How might one assign species to genera? (From now on we write genus, genera for concreteness to indicate any higher level of the taxonomic hierarchy). Suppose we distinguish certain species as “new type” due to some characteristic judged biologically significant which persists in descendant species. Then it seems sensible to use these as a basis for classification – very roughly, a “new type” species is founder of a new genus. This set-up ignores various practical problems (one seldom has the complete tree on extinct species; which characteristics should one choose as significant?) but does lead us to a purely mathematical question:

Consider schemes which, given a phylogenetic tree in which some species are distinguished as “new type”, classify all species into genera. What desirable properties can such schemes have?

Section 2 gives our answer. One might hope there was some single “natural scheme”, but it turns out that different desiderata dictate different schemes. We pick out three schemes which we name fine, medium, coarse and describe their properties.

Studying to what extent actual taxonomic practice resembles one of these theoretical schemes would make an interesting project in statistical analysis, but this is not our purpose here. Instead, as our second topic we use these theoretical classification schemes to consider the questions:

- In what ways might phylogenetic trees on genera, or time series of genera, differ from those on species?
- In what ways might phylogenetic trees on the species of an \( n \)-species genus differ from phylogenetic trees on the species of an \( n \)-species clade?
- How does the choice of classification scheme for determining genera affect such differences?
We study these questions under a certain probability model for the underlying phylogenetic tree on species. This model, described in section 3.2 and studied in detail in [2], is intended to formalize the idea of “purely random” history subject to a given number $n$ of extant species. Section 4 investigates the statistical structure of phylogenies on genera obtained by combining the species-level model with the genera classification schemes, and this combination is the conceptual novelty of the paper.

xxx say something about our specific results in section 4.

2 Phylogenetic trees and genera classification schemes

2.1 Cladograms

For good reasons, both practical and theoretical, phylogenetic relationships are usually presented via a cladogram, a binary tree (cf. Figure 4) without time scale and without identifying branchpoints with explicit taxa. Our mathematical discussion of genera classification schemes would be easier if we started with a cladogram on species. But our goal is to see how phylogenies on genera emerge from some underlying process of macroevolution at the species level in which species originate and go extinct at particular times. This requires starting with a species-level model on phylogenetic trees as defined below, even though ultimately one may choose to express relationships between genera using cladograms.

2.2 Phylogenetic trees

Our basic assumptions about macroevolution in a clade of species are logically simple, although oversimplified in reality.

- Each species has a “time of origin” and either is extant or has a “time of extinction”;

- Each species (except the founder of the clade) originates as a “daughter” of some “parent” species in the clade, not simultaneously with any other daughter.

A phylogenetic tree records all this information (birth and extinction times; parent-daughter relationships). There are many different ways to draw such a tree. Figure 1 (top left) uses one convention; time increases downwards, a species is indicated by a vertical line from time of origin to time of extinction or the present time, and parent-daughter relationship is indicated by a horizontal line with the daughter on the right.
**Figure 1.** Illustration of our schemes for defining genera in terms of new types. Above left is a complete clade of 6 extant and 16 extinct species (abcd · · · uv), with two species \{i, s\} designated as new types and marked ⊙. In the fine scheme, this induces 8 genera (3 extant), whose tree is shown above right. The other schemes are shown below, with compressed time scale.
For later use in proofs we state some language for discussing this particular representation of phylogenetic trees. Given a parent-daughter pair, there is a branchpoint on the parent’s line, from which a right edge leads to the daughter and a continuing edge leads down to another branchpoint or the leaf representing extinction time of the parent species or the current time. See Figure 2.

**Figure 2.** Terminology for phylogenetic trees.

Any species determines a subclade consisting of itself and all its descendant species. Similarly a continuing edge determines a subclade (consisting of the species whose line contains the continuing edge, and later daughter species and their descendants).

### 2.3 Desirable properties for genera classification schemes

As mentioned earlier, the question we study in section 2 is:

> given a phylogenetic tree and a subset of its species designated as “new type”, how can one classify all species into genera?

We start by considering three desirable properties for classification schemes. For all our schemes we require the following weak formalization of the idea that “new type” species should initiate new genera.

**Property 1.** A genus cannot contain both a species $a$ which is a descendant of some “new type” species $s$ and also a species $b$ which is not a descendant of $s$.

Here “descendant” includes $s$ itself, so in particular a “new type” species and its parent must be in different genera.
Next note that if we required every genus to be a clade (monophyletic) then we could never have more than one genus, because otherwise some parent-daughter pair \{a, b\} would be in different genera and then the genus containing a is not a clade. We will consider a weaker property. Any two distinct species a, b have a most recent common ancestor \text{MRCA}(a, b), which is some species (maybe a or b). Given three distinct species \{a, b, c\}, say \((a, b)\) are more closely related than \((a, c)\) if \text{MRCA}(a, b) is a descendant of \text{MRCA}(a, c). Here again we allow \text{MRCA}(a, b) = \text{MRCA}(a, c).

xxx Previous version required \text{MRCA}(a, b) \neq \text{MRCA}(a, c) but this seems wrong; see next xxx comment.

**Property 2.** Given three distinct species \{a, b, c\}, with a and b in the same genus and c in a different genus, then \((a, b)\) are more closely related than \((a, c)\).

As another kind of desirable property, one would like to be able to draw a tree or cladogram on genera in some unique way, and the next property (for a classification scheme) provides one formalization of this idea.

**Property 3.** Choosing one representative species from each genus and drawing the cladogram on these species gives a cladogram which does not depend on choice of representative species.

### 2.4 Three genera classification schemes

The properties above are always satisfied by the trivial scheme in which each species is declared to be a separate genus. Roughly speaking, such properties become easier to satisfy if one uses more genera, so one should consider schemes which produce the smallest number of genera consistent with specified properties. More precisely, if \((G_i)\) and \((G'_j)\) are two different classifications of the same set of species into genera, say \((G'_j)\) is coarser than \((G_i)\) if each \(G'_j\) is the union of one or more of the \(G_i\). Our main result in section 2, Theorem 1, gives explicit constructions of the coarsest genera classification schemes satisfying various properties.

Observe that any way of attaching “marks” to some edges of the phylogenetic tree (in the representation of section 2.2) can be used to define genera, by declaring that species \{a, b\} are in the same genus if and only if the path in the tree from the leaf a to the leaf b contains no marked edge. Here are three ways one might attach marks to edges.

(a) At each parent-daughter branchpoint where the daughter is “new type”, mark the right edge (from parent to daughter).

(b) At each parent-daughter branchpoint where the daughter’s subclade contains some “new type” species, mark the right edge (from parent to...
daughter).

(c) At each parent-daughter branchpoint where the daughter’s subclade and the continuing edge subclade both contain some “new type” species, mark both the right edge and the continuing edge.

Now define three genera classification schemes as follows.

Coarse scheme: create marks according to rule (a).

Medium scheme: create marks according to rules (a) and (c).

Fine scheme: create marks according to rules (a) and (b).

In each case the marks define genera as above. Figure 9 provides a visual catalog of these rules.

Theorem 1  (i) The coarse scheme defines genera with Property 1, and is coarser than any other scheme satisfying Property 1.

(ii) The medium scheme defines genera with Properties 1 and 3, and is coarser than any other scheme satisfying Properties 1 and 3.

(iii) The fine scheme defines genera with Properties 1 and 2, and is coarser than any other scheme satisfying Properties 1 and 2.

2.5 Proof of Theorem 1

Showing that the schemes define genera with the stated properties is straightforward, as follows.

Case (i). The marks from rule (a) ensure Property 1.

Case (iii). Consider genera defined using marks from rules (a) and (b). Suppose \{a, b\} are in the same genus and c in a different genus. The path from a to c contains a marked edge, on some parent-daughter “right edge” \((s, s')\), and either \{a, b\} are in the subclade of s and c is in the subclade of \(s'\); or \{a, b\} are in the subclade of \(s'\) and c is in the subclade of s.

In either case MRCA\((a, c)\) is s or an ancestor of s, and MRCA\((a, b)\) is s or a descendant of s. So Property 2 holds.

xxx here we can’t say that MRCA\((a, b)\) is a strict descendant of MRCA\((a, c)\).

mkp: the difference with previous version is that there i treated MRCA as a vertex of a phylogenetic tree, and it’s just impossible to have MRCA\((a, b) = MRCA(a, c)\) on a binary tree.

mkp: to make this prove precise, we need to take for \(s, s'\) the first marked edge on the path from a to c, i.e. the closest to a (there may be more than one marked edge on this path); otherwise there is no guarantee a and b are in the subclade of s – they may as well be in the subclade of some distant ancestor of s. Below is an alternative proof by contradiction, which i find more clear:
**Case (iii*).** Consider genera defined using marks from rules (a) and (b). Suppose \{a, b\} are in the same genus and c in a different genus. We’ll prove by contradiction that \((a, b)\) are necessarily more closely related than \((a, c)\).

If \((a, b)\) are not more closely related than \((a, c)\), there exist a species \(s\) such that both \(a\) and \(c\) are in the subclade of \(s\), while \(b\) is not (it’s possible that \(a = s\) or \(c = s\)). But the path from \(a\) to \(c\) contains a marked edge, meaning that there is at least one new type species in the subclade of \(s\). According to rule (b), the edge between \(s\) and its parent \(s'\) is marked, and since \(b\) is not in the subclade of \(s\), it cannot be in the same genera with \(a\).

**Case (ii).** Let \{a, b, c, \ldots\} be representatives of the different genera, and let \(a'\) be in the same genus as \(a\). We need to show that the cladograms on \{a, b, c, \ldots\} and on \{a', b, c, \ldots\} are the same. Consider the cladogram on \{a, a', b, c, \ldots\}. In this cladogram consider the branchpoint above the leaf \(a\) and the branchpoint above the leaf \(a'\). If these branchpoints are the same or linked by an edge then the two cladograms are indeed identical. Otherwise we have a cladogram as in Figure 3.

![Figure 3. Cladogram arising in proof of case (ii).](image)

Edges in the cladogram can be identified with paths in the phylogenetic tree. Because \(a\) and \(a'\) are in the same genus, there is no mark on the path from \(a\) to \(a'\). Because \(b\) and \(c\) are in different genera, there is a mark on the cladogram edges to \(b\) and to \(c\). But then by examining rules (a) and (c) we see there must be a mark on the cladogram edge from MRCA(a, b) to MRCA(a, c), contradicting the assumption that \(a\) and \(a'\) are in the same genus. This verifies case (ii).

We now need to prove the “coarser than” assertions. In each case, it is enough to show that if \(G\) is a genus in some scheme satisfying the relevant properties, then it is part of a genus in the specified scheme (coarse, medium, fine). In other words, we need to show that if \(a\) and \(b\) are in the same genus in some scheme satisfying the relevant properties, then the path from the leaf of \(a\) to the leaf of \(b\) in the phylogenetic tree does not contain any marks of the relevant kind. We will argue by contradiction, supposing that some edge \((c, d)\) on the path *does* have a mark.
Case (i). Here $(c, d)$ is a parent-daughter edge and $d$ is a “new type”. One of $\{a, b\}$ – say $b$ – is in the subclade of $d$, and so $a$ is not in that subclade. But this violates Property 1.

Case (iii). Again $(c, d)$ is a parent-daughter edge, and we may assume $b$ is in the subclade of $d$, and so $a$ is not in that subclade. Also some species $f$ in the subclade of $d$ is a “new type” species. By Property 1 $f$ is in a different genus from $a$, then by Property 2 MRCA$(a, b)$ is a descendant of MRCA$(b, f)$. But this is impossible, because MRCA$(b, f)$ is in the subclade of $d$ whereas MRCA$(a, b)$ is not.

Case (ii). As in case (i) we cannot have $(c, d)$ being a parent-daughter edge and $d$ being a “new type”. An alternate case is that $(c, d)$ is a parent-daughter edge and some $g \neq d$ in the subclade of $d$ is a “new type”, and also there is some “new type” species $f$ in the subclade determined by the continuing edge at $c$. By Property 1, the three species $\{f, d, g\}$ are in different genera and their cladogram is as in Figure 4 (left side).

![Cladograms arising in the proof of Theorem 1(iii).](image_url)

As before, assume species $b$ (which might coincide with $d$ or $g$) is in the subclade of $d$ and species $a$ is not. Then species $b$ must attach to the cladogram somewhere to the lower right of $\alpha$, and species $a$ must attach to the cladogram on one of the other edges at $\alpha$. Whether or not the genus containing $\{a, b\}$ is one of the genera containing $f$ or $d$ or $g$, this violates Property 3.

The final case is where it is the continuing edge at $c$ that is in the path from $a$ to $b$. But in this case the same argument gives the Figure 4 (right side) cladogram; now $a$ must attach to the branch to the lower left of $\alpha$ while $b$ must attach to one of the other two branches from $\alpha$. Again Property 3 is violated.

### 2.6 Further results for the genera classification schemes

These further results are intended to elucidate properties of the genera classification schemes, but are not needed for our analysis of the probability model.
xxx so we could delete this section if paper seems too long.

Figure 1 illustrates the typical behavior of the schemes. If one knew the true phylogenetic tree then the coarse scheme is clearly unsatisfactory (it puts $g$ and $r$ into the same genus despite the fact that $g$ is more closely related to the $\{i\}$ genus than to $r$ while $r$ is more closely related to the $\{stuv\}$ genus than to $g$). But one can imagine settings where an unknown tree is in fact the Figure 1 tree but, based on fragmentary fossil data, one assigns the coarse genera. The other two schemes seem more reasonable when one does know the correct tree on species.

Recall that a genus is paraphyletic if it includes its MRCA. Proposition 2 will show that genera produced from the coarse scheme or the fine scheme are always paraphyletic, and genera produced by the medium scheme are paraphyletic except in one atypical case. From Theorem 1 it is clear that the coarse scheme is coarser than (or the same as) the medium scheme and the fine scheme. Proposition 2(iii) will show that, except for the same atypical case, the medium scheme is coarser than (or the same as) the fine scheme. Figure 5 illustrates what makes the case atypical: there must be some species with at least four daughter species.

![Diagram](image)

**Figure 5.** An atypical tree and its genera. There are two “new type” species, $\{c,d\}$. Note that the coarse/fine genus $\{abef\}$ is paraphyletic while the medium genus $\{ef\}$ is not.

**Proposition 2**  
(i) Genera in the coarse scheme or the fine scheme are always paraphyletic.

(ii) If a medium genus $G$ with MRCA $a$ is not paraphyletic, write $(a,b)$ for the last right edge for which some species in $G$ is in the subclade of $b$. Then subsequent to daughter $b$, species $a$ has at least two other daughters whose subclades contain “new type” species.
(iii) Let $G$ be a fine genus which is not a subset of (or equal to) some medium genus. Let $a$ be the MRCA of $G$, so $a \in G$ by part (i). Let $b$ be the first daughter of $a$ for which the subclade of $b$ contains some species in $G$. Then the conclusion of part (ii) holds for this pair $(a, b)$.

Proof. Consider a genus $G$ whose MRCA $a$ is not in $G$, and let $(a, b)$ be the edge specified in (ii). The path between the leaves of some two species of $G$ must go along the edge $(b, a)$ and upwards along the species line of $a$ from this branchpoint $\beta$. Because that path contains no marks, to have $a$ in a different genus there must be a mark on the upwards path from the leaf of $a$ to the branchpoint $\beta$. This cannot happen with the coarse or fine genera, where there are no marks on upwards edges. With medium genera, there cannot be a mark on the continuing edge at $\beta$ (because $(a, b)$ has no mark). So there must be a mark on some subsequent edge of the species $a$ line, which implies the stated conclusion in (ii). For part (iii), let $G$ and $a$ be as in the statement. Because $G$ is not a medium genus, there is some species $f$ in $G$ which is in a different medium genus from $a$; let $c$ be the first daughter of $a$ such that the subclade of $c$ contains such a species $f$. The path from the leaf of $a$ to the leaf of $f$ contains some mark of type (c). This mark must be somewhere between the leaf of $a$ and the branchpoint of edge $(a, c)$ (otherwise there would be a mark of type (b) on edge $(a, c)$). The argument in part (ii) can be repeated to obtain the stated conclusion with $c$ in place of $b$, which implies the conclusion for $b$.

Operations on trees. Here we briefly say how the genera classifications can change when the background structure (tree and distinguished “new type” species) is changed.

(a) Suppose we don’t change the tree, but declare another species to be “new type”. This can only increase the number of marks and so can only make the genera partition become less coarse. For the coarse scheme it will typically create exactly one new genus. For the other schemes it may create more than one extra genus. For instance, in Figure 1 the designation of $s$ as a “new type” has created the fine genera $\{stuv\}, \{opqr\}, \{mn\}, \{l\}$. For the medium scheme it typically creates either one or two extra genera.

(b) Suppose we add an extra species to the tree (as the daughter of some already present species) and this extra species in not “new type”. For the fine and coarse genera, and typically for the medium genera, the new species in put into the same genus as its parent. Using Figure 6 we can see an atypical case with medium genera. If a new daughter of $a$ has branchpoint between the branchpoints of $c$ and $d$, it forms a new genus by itself, while
if its branchpoint is between the branchpoints of $e$ and $f$ then it is put into the \{ef\} genus.

2.7 Further results for the genera classification schemes

These are results related to previous section which we do use later.

In the coarse and the fine schemes, each mark is on a right edge, and the corresponding daughter species is the MRCA of its genus. So by the paraphyletic property (Proposition 2) in the coarse or fine scheme the number of genera is exactly equal to the number of marks plus one, the “plus one” for the genus containing the founder of the clade. The case of medium genera is more complicated. The path upwards from the leaf representing a species will reach a first mark (or the founder of the clade – let us associate a “virtual mark” with the founding of the clade) which does not depend on choice of species in the genus, and which is different for different genera. Thus each genus can be identified with a different mark. For instance, in Figure 6 the genus \{ef\} is identified with the virtual mark whereas genus \{ab\} is associated with the mark on the continuing edge down from the branchpoint of $d$ to the branchpoint of $c$. However, not every mark has an associated genus. For instance, if $e$ and $f$ were absent from Figure 6, then the virtual mark would have no associated genus. It is easy to check the following condition.

**Lemma 3** For medium genera, a mark is associated with a genus unless the next downward following branchpoint is a “rule (c)” branchpoint, in which both parent and daughter subclade contain new type species.

Thus the number of medium genera equals the number of marks (including the virtual mark) minus the number of such marks satisfying the condition above.
3 Tree statistics and the probability model

A statistic of a phylogenetic tree or cladogram is just a number (or collection of numbers) intended to quantify some aspect of the tree. The goal of this paper is to study, under a probability model for “purely random macroevolution”, how statistics might change when one goes from species-level trees to trees on higher-order taxa, purely as a logical consequence of the process of classification without any special biological significance. For concreteness let us start by listing some statistics for trees. Then we state our model for species-level random macroevolution, and finally combine ingredients to derive models of genus-level trees.

3.1 Examples of statistics

(a) Number of extant taxa.
(b) The time series (number of taxa in existence, as a function of past time), which includes in particular
   the time of origin of the tree
   the total number of (extinct or extant) taxa
   The maximum number of taxa in existence at any one time.
(c) The time of MRCA of extant taxa
(d) The number of extinct taxa which are ancestral to some extant taxon.
(e) Statistics dealing with the shape of the cladogram on extant taxa (see section 5.2).

3.2 The species-level probability model

We want a probability model for the phylogenetic tree on a clade with $n$ extant species (for given $n$) which captures the intuitive idea of “purely random macroevolution”. Our choice is the model below, studied in detail in [2], where some arguments (not repeated here) in its favor are presented. In (b) the phrase “rate 1” means “with probability $dt$ in each time interval of length $dt$”. So the time unit in the model equals mean species lifetime.

(a) The clade originates with one species at a random time before present, whose prior distribution is uniform on $(0, \infty)$.
(b) As time runs forward, each species may become extinct or may speciate, each at rate 1.
(c) Condition on the number of species at the present time $t = 0$ being exactly equal to $n$. 

14
Figure 6. A realization of $c$–TREE$_{20}$, a complete clade on 20 extant species. The figure is drawn so that each species occupies a vertical line (from time of origin to time of extinction (or present)), different species evenly spaced left-to-right (so that each subclade is a consecutive series), using the convention: daughters are to right of parents, earlier daughters rightmost. On the left are time series: the outer line is total number of species, the inner line is number of ancestors of extant species. Marks $\star$ indicate “new type” species, used later to construct genera. In this realization there were a total number 142 of extinct species, with a maximum of 38 species at one time; $T_{mrca} = 9.05$ and $T_{origin} = 12.75$. 

15
The “posterior distribution” on the evolution of lineages given this conditioning is then a mathematically completely defined random tree on \( n \) extant species, which we write as \( c - \text{TREE}_n \) (here \( c \) is mnemonic for complete) \(^1\). See Figure 6 for a realization with \( n = 20 \).

This model is sufficiently simple that one can do many calculations (exact formulas for given \( n \), and \( n \to \infty \) asymptotic approximations). See [2] for results for the phylogenetic tree statistics mentioned in section 3.1. The induced cladogram on the \( n \) extant species has the same distribution (ERM, for equal rates Markov) as in simpler models such as the Yule process (speciations but no extinctions: see next section) or the Moran/coalescent models (number of species fixed at \( n \), with simultaneous extinction/speciation of two random species). Properties of this distribution are well understood (xxx refs).

One lesson from [2] is that in this model, many phylogenetic tree statistics are highly variable between realizations. For instance, the time since MRCA scales as \( nT \) where \( T \) is a random variable with mean 1 but with infinite variance.

3.3 The probability model for higher-order taxa

Start with the model above for the complete tree \( c - \text{TREE}_n \) on a clade with \( n \) extant species. Introduce a parameter \( 0 < \theta < 1 \), and suppose that each species (extinct or extant) independently has chance \( \theta \) to be a new type. Then any of the three schemes from section 2.4 can be used to define an induced tree on genera, which we shall call \( \text{GENERAT}_{n, \text{species}}^\theta \) or \( \text{GENERAT}_{n, \text{species}}^{\theta, \text{medium}} \) or \( \text{GENERAT}_{n, \text{species}}^{\theta, \text{coarse}} \). Figures 7 and 8 show realizations derived from the 162-species clade in Figure 6. Further realizations can be found on our web site [www.stat.berkeley.edu/users/aldous/Research/Phylo/index.html](http://www.stat.berkeley.edu/users/aldous/Research/Phylo/index.html).

Decreasing the parameter \( \theta \) will increase the average number of species per genus: alternatively, regard decreasing \( \theta \) as moving up the taxonomic hierarchy.

This framework for probability models of higher-order macroevolution is the conceptual novelty of this paper, so (before proceeding to mathematical calculations in the next section) let us add some discussion.

xxx anomalous case of null medium genera – does it happen in simulation? Theoretical probability?

\(^1\)In (a) we use an improper [total probability is infinite] prior distribution, but after conditioning the posterior distribution of \( c - \text{TREE}_n \) is proper [total probability is 1].
Figure 7. A realization of the tree $\text{GENERA}_{n, \text{species}}^{\theta, \text{fine}}$ on extant and extinct genera, with $n = 20$ extant species and $\theta = 0.04$. It was derived from the realization of $c - \text{TREE}_{20}$ in Figure 1, with the “new type” species there indicated by ⋆. In this realization, there were 7 such “new type” species, producing 25 genera, of which 5 were extant. Letters \{a, b, c, ..., m\} indicate which of these fine genera are combined to form medium genera in Figure 8.
Figure 8. The phylogenetic trees on medium genera (left) and coarse genera (right) corresponding to the fine genera in Figure 7. There are 13 medium and 8 coarse genera.

Ingredients of model We can view the model as having three ingredients:

- the probability model (section 3.2) for phylogenetic trees on species
- the idea of using “new type” species to define genera, and the probability model above for new type species
- the particular classification schemes for defining genera.

Obviously one could choose to vary details of the first and third ingredients.

Previous models Yule [12] proposed the basic model for speciations without extinctions. Initially there is one species; each species has daughter species at rate 1. Though this species model is familiar nowadays, the main point of Yule’s work is invariably overlooked. He superimposed a model of genera by supposing that, from within each existing genus, a new species of new genus arises at some constant stochastic rate $\lambda$. This leads to a
one-parameter family of long-tailed distributions for number of species per genus (see [1] for brief description). Yule’s model perhaps foreshadows “hierarchical selection above the species level” [4]; in contrast, our model for higher-level taxa does not assume separate genus-level biological effect, but rather combines species-level novelty with conventions about how systematists construct genera. In other words, our model is intended to capture the “neutral” idea that a subclade is defined by a heritable character but that this character has no “selective advantage”, i.e. that the species in the subclade have unchanged speciation and extinction rates. This “neutral” idea was studied in [5], but there the partition of species into genera was based only on the size of subclades.

Nuances of the model  We claim that the model is sufficiently flexible that any question one might ask about genera-level macroevolution statistics can be formulated within the model. Our initial description assumed a clade with a specified number $n$ of species (and hence a random number of genera), but it’s usually more natural to work in one of the two following settings.

Model for the phylogenetic tree on $g$ extant genera. For a given number $g$, we start with the species-level model from section 3.2 with the improper prior; instead of conditioning on $n$ extant species we define genera as above, and condition on $g$ extant genera. This gives a model for random phylogenetic trees which we call e.g. $\text{GENERA}^\theta_{g,\text{fine}}$ where the superscript records the value of the “probability of new type” parameter $\theta$ and which of the classification schemes is used. Now all the kind of statistics for phylogenetic trees listed in section 3.1 can in principle be studied within this model.

Model for the phylogenetic tree of species within a genus. The conceptual point here is that genera are often not clades (some subtree forming a different genus may be absent) so that the statistical properties of the tree on species in a $m$-species genus will not coincide with those for a tree on species in a $m$-species clade. More subtly, even if a genus is a clade then the fact that some rule is used to select which clades are genera will alter statistical properties. Our model for the tree on species in a typical extinct genus, $\text{GENUS}^\theta_{g,\text{fine,extinct}}$ say, is as the $n \to \infty$ limit of a randomly chosen genus within the $n$-species model. Fortunately this limit interpretation constitutes a mathematical simplification (see “proof strategy” in section 4.2). Similarly, our model $\text{GENUS}^\theta_{g,\text{fine,extant}}$ for the tree on species in a typical extant genus is as the $n \to \infty$ limit of a randomly chosen extant genus within the $n$-species model.

Several higher levels. Finally, note that conceptually it would be simple to model simultaneously two or more higher levels such as \{genus, family\} by using two probabilities $\theta_{\text{family}} < \theta_{\text{genus}}$. See section 4.3.
4 Mathematical results for the stochastic model

We have now implicitly raised $2 \times 2 \times 3 \times K$ mathematical problems; where $K$ is the number of interesting statistics of phylogenetic trees, and where we have $2 \times 2 \times 3 = 12$ probability models arising from different combinations of

- tree on genera, or tree of species within a genus
- extant or extinct clades
- coarse, medium or fine genera.

In the remainder of the paper we present some solutions, emphasising those problems for which we can find reasonably explicit analytic solutions for different genera schemes. Of course one can answer any numerical question via simulation, and in the sequel [7] we identify the most interesting features of the model for biology and study them via simulation where necessary. Tables and graphs illustrating some of the formulas below will also be given in [7].

We first describe the central ingredient of our arguments and then treat the mathematically simplest case, of the tree of species in a typical extinct genus.

4.1 A/Z analysis

For mathematical analysis it is convenient to work with “lineage segments”. That is, in Figure 2 the species represented by the vertical line on the left has three lineage segments, determined by the two cuts at the branchpoints where the two daughter species branched off. Thus we can re-draw a phylogenetic tree as in Figure 9, where a lineage segment either ends with extinction of the species (or the current time), or else splits into two new lineage segments, the left one representing the continuation of the parent species and the right one representing the daughter species. If the daughter is “new type” we represent this as a black circle at the branchpoint; otherwise, a white circle. Each lineage segment is declared either “type A” if some descendant species is new type, or “type Z” if not. So A and Z are mnemonics for “any” and “zero”. Here “descendant” does not include the species containing the lineage segment itself. The point of this representation is that our marks which define the different genera schemes depend only on the A/Z classification of the subclades and whether the daughter is new type. Figure 9 catalogs the rules for creating marks for the three genera schemes.
4.2 Tree of species in an extinct genus

In this section we will derive the following formulas.

**Theorem 4** For the tree of species in an extinct genus in our probability model:

(a) The mean number (μ, say) of species in the genus is

\[
\begin{align*}
\mu & = \frac{\theta - 1}{(1 - \theta)} \quad \text{(coarse)} \\
\mu & = \frac{\theta - 1/2}{(1 - \theta)} \quad \text{(fine)} \\
\mu & = \frac{\theta - 1}{(3 - \theta^{1/2})^{-1}(1 + \theta^{1/2})} \quad \text{(medium)}.
\end{align*}
\]

(b) The generating function \( G(z) = E z^G \) of the number \( G \) of species in the genus is

\[
\begin{align*}
\frac{1}{2} + \frac{1 - \sqrt{(2 - \theta)^2 - 4(1 - \theta)z}}{2(1 - \theta)} & \quad \text{(coarse)} \\
\frac{z}{\sqrt{\theta}} \left( \frac{1}{1 - \sqrt{\theta} + \sqrt{1 - z(1 - \theta)} - \frac{1 - \theta}{1 + \sqrt{1 - z(1 - \theta)}}} \right) & \quad \text{(fine)} \\
\frac{1}{3 - \sqrt{\theta}} \left( 1 - \sqrt{\theta} + \frac{(1 + \sqrt{\theta})z - (1 - \sqrt{\theta})}{1 + \sqrt{1 - z(1 - \theta)}} \right) & \quad \text{(medium)}
\end{align*}
\]

(c) The distribution function \( P(L \leq t) \) for the lifetime \( L \) of the genus is

\[
\begin{align*}
\frac{e^{\theta t} - 1}{e^{\theta t} - (1 - \theta)} & \quad \text{(coarse)} \\
\theta^{-1/2} \frac{2(1 - 2\sqrt{\theta})^{-1}(1 - \sqrt{\theta} - \sqrt{\theta}e^{(2\sqrt{\theta} - 1)t}) - (\theta^{-1/2} - \sqrt{\theta})(e^{2\sqrt{\theta}t} - 1)}{(1 + \sqrt{\theta})e^{2\sqrt{\theta}t} - (1 - \sqrt{\theta})} & \quad \text{(fine)}
\end{align*}
\]

for medium scheme see eq. (26) on page 33.
Proof strategy. The continuous time critical binary branching process (CBP) starts at time 0 with one individual. Thereafter, each species is liable to become extinct (rate 1) or to speciate (rate 1). This is of course the underlying stochastic model for species from section 3.2. An intuitively easy result ([2] Proposition 4) states that, in the \( n \to \infty \) limit of our \( n \)-species model, the process consisting of a randomly chosen species (\( \sigma \), say) and its descendants, with time measured from the origin of species \( \sigma \), is the CBP process, run until extinction. So by applying A/Z analysis to the CBP process (\( T, \) say) we can assign genera to \( T \) and then study the genus containing \( \sigma \). This is our proof strategy. But be aware that the notion of “typical extinct genus” does not correspond exactly to “genus in \( T \) containing \( \sigma \)”. To make an exact correspondence, recall (section 2.7) that each genus can be identified with a mark in the underlying \( n \)-species tree; so we need to condition on \( T \) starting with a lineage segment which contains such a mark in the underlying tree (Figure 9). In practice this is not difficult to do, using the following lemma.

**Lemma 5** In the CBP, the probabilities \((p_A, p_Z)\) that the initial lineage segment are (type A, type Z) equal

\[
p_Z = \frac{1}{1+\sqrt{\theta}}, \quad p_A = 1 - p_Z = \frac{\sqrt{\theta}}{1+\sqrt{\theta}}.
\]

The generating function for the number \( N \) of species in the CBP is

\[
Ez^N = 1 - \sqrt{1-z}.
\]

**Proof.** Formula (10) is classical (xxx refs?). Formula (9) is the solution of the equation

\[
p_Z = \frac{1}{2} + \frac{1}{2}p_Z^2(1-\theta)
\]

which arises as follows. The “1/2” terms are the probabilities of extinction first (making the lineage be type Z) or speciation. In the latter case the only way to get type Z is if both lineages are type Z and also the daughter species is not new type.

**Derivation of formulas (a) for mean number \( \mu \) of species** In the coarse and the fine genera schemes, the fact that distinct genera correspond to distinct marks implies

\[
\mu^{-1} = P(\text{typical parent-daughter edge has mark})
\]

in the \( n \to \infty \) limit. In the coarse case, marks occur only when daughter is new type, so this probability equals \( \theta \), giving (1). The fine scheme result (2) follows from
Lemma 6 In the $n \to \infty$ limit model, or in $T$, the chance that a typical parent-daughter edge has a fine mark equals $\sqrt{\theta}$.

Proof. The edge does not have a fine mark if and only if the daughter is not new type and the lineage starting with daughter is type Z. So the probability in question equals

$$1 - (1 - \theta)p_Z = 1 - (1 - \theta)/(1 + \sqrt{\theta}) = 1 - (1 - \sqrt{\theta}).$$

Now consider the medium scheme. Call a lineage “type B” if the first event is speciation (instead of extinction) and if then both lineages are type A (so a type B lineage is also a type A lineage, but not necessarily conversely). Using Lemma 3 and the fact that branchpoints occur at the same rate as daughter species, the analog of (11) for medium genera is

$$\mu^{-1} = E(\text{number of medium marks at a typical branchpoint whose lineages are not type B}).$$

We now need four calculations.

$$P(\text{left lineage has mark}) = \sqrt{\theta} p_A$$

$$P(\text{left lineage has mark, is type B}) = \sqrt{\theta}(1/2)p_A\sqrt{\theta}$$

$$P(\text{right lineage has mark}) = \theta + (1 - \theta)p_A^2$$

$$P(\text{right lineage has mark, is type B}) = \theta(1/2)p_A\sqrt{\theta} + (1 - \theta)p_A^2(1/2)p_A\sqrt{\theta}.$$

Let’s argue only the final one. If the daughter is new type (chance $\theta$) we need the daughter lineage to speciate (chance 1/2) and the left sublineage to be type A (chance $p_A$) and the right sublineage to be type A (chance $\sqrt{\theta}$ by Lemma 6). If instead the daughter is not new type (chance 1 - $\theta$) then we need the original continuing lineage to be type A (chance $p_A$) and we need the daughter’s lineage to be type B (chance $1/2p_A\sqrt{\theta}$ by argument above) which implies the original right lineage has a mark.

Using these four formulas we get

$$\mu^{-1} = \sqrt{\theta} p_A - \sqrt{\theta}(1/2)p_A\sqrt{\theta} + (\theta + (1 - \theta)p_A^2) - (\theta(1/2)p_A\sqrt{\theta} + (1 - \theta)p_A^2(1/2)p_A\sqrt{\theta}).$$

Using (9) and some manipulations we find

$$\mu^{-1} = \frac{\theta(3 - \sqrt{\theta})}{1 + \sqrt{\theta}}$$

leading to (3).
Derivation of formulas (b) for generating functions of number of species per genus First consider the coarse scheme. Recall that for coarse genera, the MRCA of the different genera are exactly the different “new type” species. So a coarse genus consists of its “new type” founder and its descendant species, with the modification that any “new type” descendant species are discarded (and so don’t have descendants). Because the relative chances of a species to first (become extinct; have daughter species which is not “new type”) are \((1; 1-\theta)\), it is clear that the species in a coarse genus behave as a Galton-Watson process whose offspring distribution \(D\) is shifted geometric \((p = 1/(2-\theta))\);

\[
P(D = d) = \frac{1}{2-\theta} \left(\frac{1-\theta}{2-\theta}\right)^d, \quad d \geq 0.
\] (12)

By classical theory, the probability generating function \(g(z) = E(z^G)\) of the total size \(G\) of the Galton-Watson process is determined by the probability generating function \(f_D(z) = E(z^D)\) as the unique positive solution of the equation \(g(z) = zf_D(g(z))\). When the offspring distribution is shifted geometric \((p)\) we have \(f_D(z) = p/(1 - (1-p)z)\), and hence \(g(z) = (1 - \sqrt{1 - 4p(1-p)z})/2(1-p)\). Setting \(p = 1/(2-\theta)\) gives (4).

We now consider fine genera. For a species \(s\) write \(B_s\) for the event “neither \(s\) nor any descendant is new type”. Writing \(\sigma\) for the daughters of \(s\), define

\[
D = \sum_\sigma 1_{B_\sigma}, \\
D' = \sum_\sigma 1_{B_{c\sigma}}, \\
N_\sigma = \text{number of species in subclade of } \sigma, \\
M = 1 + \sum_\sigma N_\sigma 1_{B_\sigma}.
\]

Take \(s\) to be a random species in the \(n \to \infty\) limit of the \(n\)-species model, so that the subclade of \(s\) is distributed as the CBP. Then \(s\) is the MRCA of its fine genus if and only if event \(B_s^c\) occurs, in which case the size of the genus is \(M\). So the generating function of “number of species per typical fine genus” is

\[
E(z^M | B_s^c) = \frac{Ez^M 1_{B_s^c}}{P(B_s^c)} = \frac{Ez^M - Ez^M 1_{B_s}}{\sqrt{\theta}}
\] (13)

because \(P(B_s^c) = \sqrt{\theta}\) by Lemma 6.
From the definition of $M$,
\[
E(z^M) = z \sum_{d \geq 0} P(D = d) H^d(z)
\]
where $H(z)$ is the generating function of size of a clade for which event $B$ occurs, that is
\[
H(z) = E(z^M | B_s).
\]
Event $B_s$ occurs if and only if $s$ is not new type, and $D' = 0$, so
\[
E(z^M 1_{B_s}) = (1 - \theta) z \sum_{d \geq 0} P(D = d, D' = 0) H^d(z).
\]
Throughout the lifetime of $s$, the chances of the next event being

(s goes extinct; daughter $\sigma$ and event $B_{\sigma}$; daughter $\sigma$ and event $B_{\sigma}^c$)

are
\[
\left( \frac{1}{2}; \frac{1 - \sqrt{\theta}}{2}; \frac{\sqrt{\theta}}{2} \right)
\]
which easily implies
\[
P(D = d) = (1 - \rho)^d \rho, \ d \geq 0; \quad \rho = 1/(2 - \sqrt{\theta})
\]
\[
P(D = d, D' = 0) = \frac{1}{2} \left[ \frac{1}{2} (1 - \sqrt{\theta}) \right]^d.
\]
Next, because conditioning on $B_s$ is conditioning on no species to be new type,
\[
P(M = n | B_s) = \frac{P(M = n) (1 - \theta)^n}{P(B_s)} = \frac{P(M = n) (1 - \theta)^n}{1 - \sqrt{\theta}}
\]
by Lemma 6 again. Thus, in terms of the unconditioned generating function $G(z) = E z^M = 1 - \sqrt{1 - z}$ at (10) the conditioned generating function is
\[
H(z) = E(z^M | B_s) = \frac{G((1 - \theta) z)}{1 - \sqrt{\theta}} = \frac{1 - \sqrt{1 - z(1 - \theta)}}{1 - \sqrt{\theta}}.
\]
We can now calculate
\[
E(z^M) = z \sum_{d \geq 0} (1 - \rho)^d \rho H^d(z) = \frac{\rho z}{1 - (1 - \rho) H(z)}
\]
\[
= \frac{z}{1 - \sqrt{\theta} + \sqrt{1 - z(1 - \theta)}}
\]
\[
E(z^M 1_{B_s}) = (1 - \theta) z \sum_{d \geq 0} \frac{1}{2} \left[ \frac{1}{2} (1 - \sqrt{\theta}) \right]^d H^d(z)
\]
\[
= \frac{(1 - \theta) z \frac{1}{2}}{1 - \frac{1}{4}(1 - \sqrt{\theta}) H(z)} = \frac{(1 - \theta) z}{1 + \sqrt{1 - z(1 - \theta)}}.
\]
Inserting into (13) gives the desired formula (5).

We now consider the medium scheme. Consider the initial lineage in the CBP. Define $M(z)$ to be the generating function of the number of species in the CBP such that there is no medium mark on the path between this initial lineage and the species label at its extinction time. Define $M_A(z), M_Z(z)$ similarly but conditioning on the initial lineage being type $A$ or type $Z$. So

$$M(z) = p_A M_A(z) + p_Z M_Z(z)$$

for $p_A, p_Z$ at (9); also

$$M_Z(z) = H(z) = \frac{1 - \sqrt{1 - z(1 - \theta)}}{1 - \sqrt{\theta}}$$ (15)

for $H(z)$ at (14) because each is the generating function for number of species in the CBP conditioned on no new type species.

We first show how the desired generating function $G(z)$ of number of species in a typical extinct medium genus is related to the quantities above. Consider medium marks in the $n \to \infty$ limit of the $n$-species model; recall (Lemma 3) that we can associate such marks with medium genera but that such genera may be empty. There are three possible categories of medium mark which might be associated with a branchpoint; below we state their probabilities and the generating function (g.f.) of the number of species below the mark which contribute to the number of species in the medium genus associated with the mark.

- Mark on parent-daughter edge where daughter is new type: chance = $\theta$, g.f. = $M(z)$ because subclade of daughter is distributed as CBP.
- Daughter is not new type but parent-daughter edge has mark: chance = $(1 - \theta)p_A^2$, g.f. = $M_A(z)$ because initial lineage of daughter must be type A.
- Mark on continuing lineage of parent; chance = $p_A \sqrt{\theta}$ using Lemma 6, g.f. = $M_A(z)$ because continuing lineage is type A.

Adding these contributions gives

$$G(z) = \theta M(z) + \left((1 - \theta)p_A^2 + p_A \sqrt{\theta}\right) M_A(z)$$

whose interpretation is as the $n \to \infty$ limit of $n^{-1} \sum_i z^{g_i}$ for the sizes of medium genera associated with marks $i$ in the $n$-species model. To get the desired $G(z)$ we need to discard the null genera and normalize to a probability distribution, so

$$G(z) = \frac{G(z) - G(0)}{G(1) - G(0)}.$$
The formula for $\bar{G}$ reduces to

$$\bar{G}(z) = \frac{\theta}{1+\sqrt{\theta}}(M_Z(z) + 2M_A(z))$$

and then, because $M_Z(0) = 0$,

$$G(z) = \frac{M_Z(z) + 2(M_A(z) - M_A(0))}{1 + 2(1 - M_A(0))}. \quad (16)$$

To get an equation for $M_A(z)$, consider the initial lineage of a CBP. In order for this to be type A, one of the following three possibilities must occur at the first branchpoint; we state their (unconditional) chances and their contribution to $M_A(z)$ if they occur.

Parent-daughter edge has no fine mark; and either continuing lineage is type A, daughter lineage is type Z, either continuing lineage is type Z, daughter lineage is type A. Chance $2(1 - \theta)p_Ap_Z$; contribution to g.f. = $M_A(z)M_Z(z)$ because both lineages contribute.

Parent-daughter edge has a fine mark; continuing lineage is type A. Chance $p_A\sqrt{\theta}$; contribution to g.f. = 1 because this is the case where the genus is null.

Parent-daughter edge has a fine mark; continuing lineage is type Z. Chance $p_Z\theta$; contribution to g.f. = $M_Z(z)$ because only the continuing lineage contributes.

The unconditional probabilities $(2(1 - \theta)p_Ap_Z, p_A\sqrt{\theta}, p_Z\theta)$ of these three cases become (after normalization to sum to 1) the conditional probabilities given original lineage is type A: $(1 - \sqrt{\theta}, \frac{1}{2}\sqrt{\theta}, \frac{1}{2}\sqrt{\theta})$. Thus

$$M_A(z) = (1 - \sqrt{\theta})M_A(z)M_Z(z) + \frac{1}{2}\sqrt{\theta} + \frac{1}{2}\sqrt{\theta}M_Z(z)$$

whose solution is

$$M_A(z) = \frac{\sqrt{\theta}}{2} \cdot \frac{1 + M_Z(z)}{1 - (1 - \sqrt{\theta})M_Z(z)}$$

$$= \frac{\sqrt{\theta}}{2} \cdot \frac{1 + M_Z(z)}{1 + \sqrt{1 - z}(1 - \theta)}$$

using (15) in the denominator. Inserting into (16) gives

$$G(z) = (M_Z(z) + 2M_A(z) - \sqrt{\theta})/(3 - \sqrt{\theta})$$

$$= \frac{1}{3 - \sqrt{\theta}} \left(1 - \sqrt{\theta} + \frac{(1 + \sqrt{\theta})z - (1 - \sqrt{\theta})}{1 + \sqrt{1 - z}(1 - \theta)}\right)$$

which is the desired expression (6).
Derivation of formulas (c) for distribution of genus lifetime

As before we consider the CBP. Write \( L \) for the lifetime of the genus (in some scheme) containing the founding species, and write \( \text{type}(\iota) \) for the type (A or Z) of the initial lineage \( \iota \). Write

\[
q_Z(t) = P(L \leq t, \text{type}(\iota) = Z) \\
q_A(t) = P(L \leq t, \text{type}(\iota) = A).
\]

We shall argue that these distribution functions satisfy the differential equations

\[
q'_Z(t) + 2q_Z(t) = (1 - \theta)q_Z^2(t) + 1, \quad (17)
\]

\[
(q'_Z + q'_A) + 2(q_Z + q_A) = R(q_A, q_Z, p_A) + 1 \quad (18)
\]

with initial conditions

\[ q_Z(0) = q_A(0) = 0, \]

where \( p_A = \sqrt{\theta}/(1 + \sqrt{\theta}) \) (9) and where \( R \) is a function depending on genera scheme, given in (21,22,??) below. These equations are derived using the \textit{backwards equations} method for branching processes (xxx refs). That is, in initial time \( dt \) we have (to first order in \( dt \))

- chance \( 1 \cdot dt \) of extinction
- chance \( 1 \cdot dt \) of a branchpoint
- chance \( 1 - 2 \cdot dt \) of neither.

To derive (17) note that the “extinction” possibility implies the lineage is type Z. so

\[
q_Z(t + dt) = 1 \cdot dt + \Xi_t 1 \cdot dt + (1 - 2 \cdot dt)q_Z(t) \quad (19)
\]

where

\[
\Xi_t = P(L \leq t, \text{type}(\iota) = Z|\text{branchpoint at time 0+}).
\]

This rearranges to

\[
q'_Z(t) + 2q_Z(t) = 1 + \Xi_t. \quad (20)
\]

In order to have \( L \leq t \) and \( \text{type}(\iota) = Z \) after a branchpoint at time 0+, we need the daughter to not be new type, and we need both subsequent lineages to be type Z, so \( \Xi_t = (1 - \theta)q_Z^2(t) \), giving (17).

To derive (18), repeat the argument for (19) to get

\[
q'_Z(t) + q'_A(t) + 2(q_Z(t) + q_A(t)) = 1 + R_t
\]

\[
R_t = P(L \leq t|\text{branchpoint at time 0+}).
\]
We now consider each scheme in turn, to get formulas for \( R_t \). In each case we condition on a branchpoint at time \( 0^+ \), and write \( L_l, L_r \) for the “lifetime of genus” quantity \( L \) applied to the lineages \( \iota_l, \iota_r \) after the branchpoint.

*For coarse genera*, in order to have \( L \leq t \) we need

- either: daughter is new type, \( L_l \leq t \);
- or: daughter is not new type, \( L_l \leq t, L_r \leq t \).

So

\[
R_t = \theta(q_Z(t) + q_A(t)) + (1 - \theta)(q_A(t) + q_Z(t))^2 := R_{\text{coarse}}(q_A(t), q_Z(t)). \tag{21}
\]

*For fine genera*, in order to have \( L \leq t \) we need

- either: daughter is new type, \( L_l \leq t \);
- or: daughter is not new type, \( L_l \leq t, \text{type}(\iota_r) = A \);
- or: daughter is not new type, \( L_l \leq t, \text{type}(\iota_r) = Z, L_r \leq t \).

So

\[
R_t = \theta(q_A(t) + q_Z(t)) + (1 - \theta)(q_A(t) + q_Z(t))(p_A + q_Z(t)) := R_{\text{fine}}(q_A(t), q_Z(t), p_A). \tag{22}
\]

*For medium genera*. The case of medium genera is more complicated and will be treated below separately.

xxx Note \( q_Z(t) \) is just the distribution of lifetime of a certain subcritical binary branching process, so presumably (23) is some known result.

Solving these equations we get the following results:

\[
q_Z(t) = \frac{e^{2\sqrt{\theta}t} - 1}{(1 + \sqrt{\theta})e^{2\sqrt{\theta}t} - (1 - \sqrt{\theta})} \tag{23}
\]

\[
\left. q_Z(t) + q_A(t) \right|_{\text{coarse}} = \frac{e^{\theta t} - 1}{e^{\theta t} - (1 - \theta)} \tag{24}
\]

\[
\left. q_A(t) + q_Z(t) \right|_{\text{fine}} = 1 - \frac{2\sqrt{\theta}}{1 - 2\sqrt{\theta}} \cdot \frac{(1 - \sqrt{\theta}) - \sqrt{\theta}e^{(2\sqrt{\theta}-1)t}}{(1 + \sqrt{\theta})e^{2\sqrt{\theta}t} - (1 - \sqrt{\theta})}. \tag{25}
\]

xxx in the fine case we get \( 0/0 \) when \( 2\sqrt{\theta} = 1 \); can we rewrite formula to avoid this?

We now need to translate these formulas \( P(L \leq t) = q_Z(t) + q_A(t) \) for the distribution function of \( L \) (the size of genus in a CBP containing the founder species \( \sigma \)) into formulas for the distribution function of \( L \) (the size of a typical extinct genus), the relation being that \( L \) has the conditional distribution of \( L \) given that \( \sigma \) (regarded as a species sampled from a \( n \to \infty \) limit clade) is founder of a genus in that clade. For coarse genera, we are just conditioning on \( \sigma \) being a new type species, which has no effect on the
distribution of the CBP, so formula (24) immediately becomes formula (7) for $L$.

For fine genera, the marking rule implies

\[ \mathcal{L} \text{ has the conditional distribution of } L \text{ given that either } \sigma \text{ or some descendant of } \sigma \text{ is new type.} \]

There is chance $\theta$ for $\sigma$ to be new type, and chance $(1 - \theta)p_A$ that $\sigma$ is not new type but some descendant is new type. So

\[
P(\mathcal{L} \leq t) = \frac{\theta}{\theta + (1 - \theta)p_A} (q_A(t) + q_Z(t)) + \frac{(1 - \theta)p_A}{\theta + (1 - \theta)p_A} \frac{q_A(t)}{p_A}
\]

\[
= \frac{q_A(t) + q_Z(t) - (1 - \theta)q_Z(t)}{\sqrt{\theta}}
\]

using the fact $\theta + (1 - \theta)p_A = \sqrt{\theta}$. Now (23,25) give formula (8).
lifetime for medium genera  We identify each genus with the oldest species in this genus (in fine and coarse schemes this is essentially the same as identifying a genus with a mark on a tree, but not in medium scheme), the birth time of this species giving the ”starting point”, which we use to measure genera lifetime.

Consider a species $a$ that originates from it’s parent $b$ at the branchpoint $\beta$.

For $a$ to be the oldest species in it’s genera there are three alternatives:
1) there is a medium mark on the lineage of $a$; in this case we say that the right subtree below $\beta$ has type $B$
2) there is a medium mark on the parent-daughter edge between $a$ and $b$, but no marks on the lineage of $a$
3) there are no marks on the lineage of $a$ and no medium mark on $a - b$ parent-daughter edge, but $a$ is still the oldest in it’s genera.

It’s clear that in first two cases above no species older than $a$ can be in the same genus (since by definition the two species are in the same genus only if the path in the tree between the corresponding leafs contains no marked edge). Consider in detail the third case.

Since $a$ and $b$ are in different genera, but there is no medium mark on the path from $a$-leaf to branchpoint $\beta$, there is necessarily a mark between $\beta$ and $b$-leaf. That is, the left tree below $\beta$ has type $B$.

Note that this means that $a$ necessarily has type $Z$; otherwise $\beta$ would be a branchpoint of type $A + A$ and both edges below it would have medium marks.

If the branchpoint right above $\beta$ is the starting point of another daughter of $b$ (this has probability $1/2$), this daughter must be either new-type either type $A$ (probability $(1 - \theta)p_A + \theta = \sqrt{\theta}$); otherwise it would be in the same genera with $a$ and $a$ would not be the oldest in it’s genera.

If the branchpoint $\beta'$ right above $\beta$ is the starting point of $b$ itself (probability $1/2$), then (denote by $b'$ the parent of $b$) for $a$ and $b'$ to be in different genera, either the parent-daughter edge $b - b'$ must have a medium mark, either the segment of the lineage of $b'$ between $\beta'$ and $b'$-leaf must have a medium mark. But in the latter case, both subtrees below $\beta'$ are of type $A$, so we there is a mark on parent-daughter edge $b - b'$ anyway (probability $\theta + (1 - \theta)p_A = \sqrt{\theta}$).

The probability for the initial lineage to be of type $B$ satisfies the equation

$$p_B = \frac{1}{2}p_A\sqrt{\theta} + \frac{1}{2}(1 - \sqrt{\theta})p_B,$$
so the first alternative has probability

\[ p_B = \frac{\theta}{(1 + \sqrt{\theta})^2}, \]

the second alternative has probability

\[ \theta(1 - p_B) + (1 - \theta)(p_A - p_B)p_A = \frac{\theta(\sqrt{\theta} + 2)}{(1 + \sqrt{\theta})^2}, \]

and the third alternative has probability

\[ p_Z p_B \left( \frac{1}{2} \sqrt{\theta} + \frac{1}{2} \sqrt{\theta} \right) = \frac{\theta}{(1 + \sqrt{\theta})^2}. \]

Together this gives probability for a randomly chosen species to be the oldest in its genera equal to

\[ \frac{\theta(3 - \sqrt{\theta})}{1 + \sqrt{\theta}}, \]

which agrees with (3).

Now let’s return to the lifetime distribution. As before we write \( L \) for the lifetime of the genus containing the founding species, and write \( \text{type}(\iota) \) for the type of the initial lineage \( \iota \). We’ll need three types: \( Z, B \) (which is subset of \( A \)) and \( \bar{A} \) (which stands for ”\( A \) but not \( B \)”). Write

\[
\begin{align*}
q_Z(t) &= P(L \leq t, \text{type}(\iota) = Z) \\
q_{\bar{A}}(t) &= P(L \leq t, \text{type}(\iota) = \bar{A}) \\
q_B(t) &= P(L \leq t, \text{type}(\iota) = B).
\end{align*}
\]

\( q_Z(t) \) was calculated earlier. \( q_B(t) \) is defined by

\[ q_B'(t) + 2q_B(t) = q_B(t) + \theta q_{\bar{A}}(t) + (1 - \theta)p_A q_{\bar{A}}(t), \quad q_B(0) = 0, \]

\( q_{\bar{A}} \) is defined by

\[ q_{\bar{A}}'(t) + 2q_{\bar{A}}(t) = \theta q_Z(t) + 2(1 - \theta)q_Z(t)q_{\bar{A}}(t) + (1 - \theta)q_Z(t)q_{\bar{B}}(t), \quad q_{\bar{A}}(0) = 0, \]

where \( q_{\bar{B}}(t) \) is defined by

\[ q_{\bar{B}}'(t) + 2q_{\bar{B}}(t) = (1 - \theta)q_B(t)q_Z(t) + \sqrt{\theta}p_A, \quad q_{\bar{B}}(0) = 1/2p_A \sqrt{\theta}. \]

Here \( q_{\bar{B}} \) is the probability that the lineage \( \iota \) has type \( B \) and the oldest leaf, reachable from the initial point of the CBP (that is such that there is no
medium marks on the path to this leaf) is not older than \( t \). Since it’s possible that the first branchpoint in the CBP is of type \( A + A \), and no leaves are reachable from the origin, we need a non-zero initial condition for \( q_B \).

Solving the differential equations above, we find

\[
q_B(t) = \frac{\theta(e^{2t\sqrt{\theta}} - 1) + \theta^{3/2}e^{t\sqrt{\theta} - t}}{(1 + \sqrt{\theta})e^{2t\sqrt{\theta}} + \theta - 1},
\]

\[
q_A(t) = \frac{XXX}{(1 + \sqrt{\theta})e^{2t\sqrt{\theta}} + \sqrt{\theta} - 1}^{2}(1 + \sqrt{\theta})
\]

\[
XXX \ = \ \left( \theta + \sqrt{\theta} \right) e^{3t\sqrt{\theta}} + \theta \left( 2\theta - 2t\sqrt{\theta} - 4t - \sqrt{\theta} + 2t\theta - 1 \right) e^{2t\sqrt{\theta}}
\]

\[
+ \theta^{3/2} - \sqrt{\theta} - \theta \left( \theta e^{2t\sqrt{\theta}} + e^{2t\sqrt{\theta}} \sqrt{\theta} + \theta - \sqrt{\theta} \right) e^{t(\sqrt{\theta} - 1)}
\]

and

\[
q_B(t) = \sqrt{\theta} \int_0^t q_A(u)e^{u-t} du.
\]

Finally, the overall genera lifetime distribution is a weighted sum of distributions above, more exactly

\[
q_{\text{medium}}(t) = \frac{1 + \sqrt{\theta}}{\theta(3 - \sqrt{\theta})} \left( q_B(t) + \theta q_Z(t) + \sqrt{\theta} q_A(t) + \sqrt{\theta}(1 - \theta)q_B(t)q_Z(t) \right).
\]

(26)
4.3 Tree of extinct genera

We now consider aspects of the trees on genera, illustrated in Figures 7 and 8. For the first result, each genus has some number (maybe zero) of “direct offspring” genera. For instance, in Figure 7 (fine genera) genus $b$ has two offspring, the first genus $c$ and the genus $k$. For Figure 7 the numbers of fine genera with (0; 1; 2) offspring genera are (4; 18; 3). Let us write “offspring tree” for the tree recording this relationship. So the offspring tree carries less information than the complete tree (e.g. the lifetime of a genus is not included) but more information than the induced cladogram (e.g. it includes the identity of MRCA genera).

**Proposition 7** (a) The offspring tree of a typical extinct fine genus and its descendant genera is distributed as the Galton-Watson branching process with offspring distribution $\xi$ such that

\[
P(\xi = 0) = p_A
\]

\[
P(\xi = i) = (1 - p_A)^2 p_A^{i-1}, \quad i \geq 1
\]

where $p_A = \sqrt{\theta} / (1 + \sqrt{\theta})$.

(b) For coarse genera, the corresponding offspring tree is distributed as the Galton-Watson branching process whose offspring distribution $\xi$ has generating function

\[
Ez^\xi = \frac{1 - \frac{1}{2}\theta z - \sqrt{\frac{1}{4}\theta^2 z^2 - \theta z + \theta}}{1 - \theta}.
\]

(27)

In both cases $E\xi = 1$, so that (as expected) the “critical” property of the species-level model is preserved at the genus level. Note also that in Figure 7, where $p_A = 1/6$, the data on offspring frequency matches well the distribution (a) of $\xi$, even though Figure 7 refers to the extant setting.

**Proof.** (a) Consider a species $\sigma$ as the founder of CBP and as a sampled species from a large clade conditioned on the edge (parent($\sigma$), $\sigma$) having a fine mark. Consider the fine genus $g_0$ founded by $\sigma$. The number of offspring genera is exactly the number $\xi$ of daughter species $\sigma_i$ of $\sigma$ such that the edge $(\sigma, \sigma_i)$ has a fine mark. It easily follows that the offspring tree under consideration in Proposition 7 is the Galton-Watson process with offspring distribution $\xi$. Recall the $A/Z$ analysis from section 4.1. If the initial lineage of $\sigma$ is type $Z$ then $\xi = 0$. If it is type $A$ (probability $q$, say) then at each marked edge $(\sigma, \sigma_i)$ there is some probability ($r$, say) that the continuing
lineage of $\sigma$ is type A. So the distribution of $\xi$ has the form

$$P(\xi = 0) = 1 - q$$
$$P(\xi = i) = qr^{i-1}(1 - r), \ i \geq 1.$$ 

To calculate $r$, note that conditioning on a parent-daughter edge having a fine mark (which forces the lineage above the split to be type A) does not affect probabilities for the type of the continuing parental-species lineage, so $r = p_A$ in (9). To calculate $q$, in the setting of the founder $\sigma$ of CBP,

$$q = P(\text{lineage is type A} | \text{lineage is type A, or \sigma is new type}).$$

The conditioning event has chance $\sqrt{\theta}$ by Lemma 6. So

$$q = p_A/\sqrt{\theta} = p_Z = 1 - p_A$$

giving the distribution in (a).

For coarse genera, each genus is founded by a new type species, so clearly the offspring tree we seek is the Galton-Watson process with offspring distribution $\xi$ described as follows.

Start CBP with a species $\sigma$ which is not new type, but disallow descendants of any new type species; let $\xi$ be the number of new type species.

Because $\sigma$ may (become extinct; have new type daughter; have not new type daughter) with chances $(1/2; \theta/2; (1 - \theta)/2)$, the generating function $\Phi(z) := E z^\xi$ satisfies the equation

$$\Phi = \frac{1}{2} \left( 1 + \theta z \Phi + (1 - \theta) \Phi^2 \right)$$

whose solution is stated in (27).

xxx explicit formula for distribution of $\xi$ in (b)?

xxx we should be able to do the medium case by considering a two-type process; I haven’t tried.

**Interpretation** There are several ways to interpret Proposition 7 as a statement about “typical trees on extinct genera”. First, we could consider the tree on all genera in a large clade; given that a subtree has $g$ genera, its distribution (for the tree of offspring) is the Galton-Watson process in Proposition 7 conditioned on having exactly $g$ genera.

Another interpretation uses the genus/family model mentioned at the end of section 3.3. Set $\theta_{\text{family}} < \theta = \theta_{\text{genus}}$ and suppose that each species
has chance $\theta$ to be new genus type or new family type, and chance $\theta_{\text{family}}$ to be new family type. Now we can consider “the tree on genera in a typical family” in the way analogous to “the tree on species in a typical genus” previously studied.

**Proposition 8** In the coarse scheme, the offspring tree on genera in a typical extinct family is distributed as the Galton-Watson branching process whose offspring distribution $\xi'$ has generating function

$$\Phi'(z) = \Phi(z + \theta'(1 - z))$$

where $\theta' = \theta_{\text{family}}/\theta$ and where $\Phi(z)$ is the generating function (27).

**Proof.** The process of all descendant species of a typical “new family type” species $\sigma$ is just CBP. So as in Proposition 7(b), the offspring tree of genera (where a genus may or may not be in a new family) is just the Galton-Watson process with offspring distribution $\xi$ at (27). We want the subprocess containing only genera in the same family as $\sigma$. Since a new genus type has chance $\theta'$ to be a new family type, the subprocess is just the Galton-Watson process with offspring distribution $\xi'$ described by

the conditional distribution of $\xi'$ given $\xi$ is Binomial($\xi, 1 - \theta'$)

and (28) follows.

xxx it seems likely that the fine case of Proposition 8 can be done, building on the analysis of “number of species per fine genus” earlier.
4.4 Tree of species in an extant genus

**Proposition 9** In the coarse scheme, the number $\tilde{N}$ of extant species in the genus of a typical extant species has Geometric($\theta$) distribution

$$P(\tilde{N} = n) = \theta (1 - \theta)^{n-1}, \quad n \geq 1. \quad (29)$$

So the number $N$ of extant species in a typical extant genus has the inverse-size-biased distribution

$$P(N = n) = \frac{(1 - \theta)^n}{c_{\theta} n}, \quad n \geq 1 \quad (30)$$

where $c_{\theta} = \sum_{n \geq 1} (1 - \theta)^n / n = -\log \theta$. So

$$EN = \frac{\theta - 1}{\theta \log \theta}.$$ 

xxx Maxim has a generating function argument for (29) in his section 3.3. As a check I outline a different argument below for (30). Please try to get further results in this “extant genus” setting. For instance, “time since origin of an extant clade” must have density function proportional to $t \rightarrow P(X_t \geq 1)$ in (31) below. What about time since MRCA? What about the fine scheme? Lea’s notes (June 2005) on local limits may be helpful here.

**Outline alternate proof of Proposition 9.** Consider CBP where the initial species $\sigma$ is new type. Let $X_t$ be the number of species alive at time $t > 0$ which are in the same coarse genus as $\sigma$. Then

$$P(N = \cdot) \propto \int_0^\infty P(X_t = \cdot) \, dt \quad (31)$$

because in the underlying infinitely-large clade, new type species arose at constant rate in the past. Now $(X_t)$ is the birth-and-death process with transition rates

$q(x, x + 1) = (1 - \theta)x$

$q(x, x - 1) = x$

By Markov chain theory the “mean occupation time” in (31) is proportional to the stationary distribution $\pi(\cdot)$ of the process $(X_t)$ (where we insert some arbitrary transition rate $0 \rightarrow 1$). But the stationary distribution satisfies

$$\pi(x + 1)/\pi(x) = q(x, x + 1)/q(x + 1, x) = (1 - \theta)x/(x + 1)$$

whose solution is (30).
References

[1] D.J. Aldous. Stochastic models and descriptive statistics for phylo-


5 Misc

5.1 Notes to Lea and Maxim on this version

What’s new in this version?

• A technical modification in “more closely related than” definition (section 2.3)

• A little more about the genera schemes (sections 2.6 and 2.7)

• Formulas (3) and (6) for mean and PGF of number of species per medium extinct genus.

• Maxim’s formulas (24,??,25) for lifetime of extinct genera need some minor modifications (in fine and medium cases) to have the correct interpretation; I’ve done this for the fine case in (8).

• Propositions 7 and 8 and “rough notes” in section 5.2.

What would we like to do to finish this paper?

• Ideally we would like 3 theorems parallel to Theorem 4, to cover all 4 possible combinations of
tree of species in a genus / tree on genera
extinct/extant.

Propositions 7 and 8 make a small start on one of these theorems, and Proposition 9 is a very small start on another.

• More specifically, we should be able to do the following:

• distribution of lifetime of extinct medium genera (see note at end of section 4.2)

• analogs of Propositions 7 and 8 and 9 for different schemes

• some calculations with the “tree shape” statistics (section 5.2).

Other comments

• It would be good to find some standard monograph on branching processes for citing standard results/techniques.

• Let’s concentrate on theorems and formulas in this paper, and save graphs/tables for the “biology” paper.
• Similarly in section 4 I wanted to focus on stating and deriving explicit formulas rather than on more abstract “structural math” as in Maxim’s recoloring lemma.

• I had a student check Maxim’s formulas (23 - 25). Maxim, add comment on how you got solutions (by hand or computer?)

• I haven’t used Maxim’s A/Z rate table (copied at end) but one might be able to simplify proofs using it.

• Some material from previous drafts is not included here: usually because it doesn’t lead to explicit formulas. Please think if other material can be made to yield explicit formulas.
5.2 Three tree shape statistics

A wide variety of statistics designed to quantify aspects of the “shape” of trees have been proposed and studied. See [3, 8] for recent references. We will consider only three simple statistics. These are defined for cladograms on \( n \) leaves. Our convention (cf. Figure 3) is that a cladogram has a degree-1 root and so has exactly \( n - 1 \) branchpoints. In the definitions below, we write \( i \) for a typical leaf and \( \beta \) for a typical branchpoint.

Define the depth \( d(i) \) of leaf \( i \) to be the number of edges on the path from the root to \( i \). A branchpoint \( \beta \) defines two subtrees; write \( s_1(\beta), s_2(\beta) \) for the number of leaves in these subtrees. If \( s_1(\beta) = s_2(\beta) = 1 \) then \( \beta \) induces a cherry.

Definitions. For a \( n \)-leaf cladogram define

\[
D := \frac{1}{n} \sum_i d(i) \quad \text{average depth} \tag{32}
\]

\[
I := \frac{1}{n-1} \sum_{\beta} |s_1(\beta) - s_2(\beta)| \quad \text{Colless’ index} \tag{33}
\]

\[
C := \frac{2}{n} \text{ (number of cherries)} \quad \text{proportion of cherries.} \tag{34}
\]

Intuitively, an increase in “imbalance” is associated with an increase in \( D \) and \( I \), and a decrease in \( C \). The normalizations (in \( n \)) are rather arbitrary. It is easy to see that \( D \) cannot be smaller than order \( \log n \).

In the context of a probability model for \( n \)-leaf cladograms we get random variables \( D_n, I_n, C_n \) and we shall consider only the \( n \to \infty \) asymptotics of \( ED_n, EI_n, EC_n \). We can make a crude classification of models as

- balanced if \( ED_n/\log n \) is bounded
- unbalanced if \( ED_n/\log n \to \infty \).

In [3] it is observed that \( |D - I| = O(\log n) \). So for unbalanced models, the first-order behavior of \( EI_n \) is the same as for \( ED_n \).

xxx quote results for Yule (ERM) and for uniform models.

**Rough notes on** \( EC_n \). A cherry in the cladogram is a pair \((s_0, s_1)\) in the offspring tree of section 4.3 such that \( s_1 \) is the youngest daughter of \( s_0 \), and \( s_1 \) has no children. For a Galton-Watson b.p. with offspring distribution \( \xi^* \) which is critical \((E\xi^* = 1)\) we have:
conditional on total population = \( n \), then in the \( n \to \infty \) limit,
\[
\frac{1}{n} \text{(number of (} s_0, s_1 \text{) pairs )} \to P(\xi^* > 0)P(\xi^* = 0). \tag{35}
\]

If instead the Galton-Watson b.p. has subcritical offspring \( \xi \), then (35) holds for the “tilted” distribution \( \xi^* \) defined by

\[
P(\xi^* = i) = P(\xi = i)z^i/\Phi(\xi(z))
\]
where \( z \) is chosen to make \( E\xi^* = 1 \), that is \( z \) solves

\[
z\Phi'(z)/\Phi(z) = 1.
\]

Now in the setting of section 4.2 (tree of species in an extinct genus) in the coarse case the offspring distribution \( \xi \) is given by (12). The tilted distribution is \( P(\xi^* = i) = (1/2)^{i+1}, \ i \geq 0 \) which (surprisingly?) does not depend on \( \theta \). So, if \( C_n \) is our “cherry statistic” for “tree of species in extinct coarse genus, conditioned on \( n \) species”, we get

\[
\lim_n EC_n = 2P(\xi^* > 0)P(\xi^* = 0) = 1/2
\]
for the coarse case.

xxx we surely can do the fine case; what about the medium case?

Also, we can consider, as in Proposition 7, the tree of genera conditioned on having \( n \) genera. So in the fine case, Proposition 7(a) gives

\[
\lim_n EC_n = 2P(\xi > 0)P(\xi = 0) = \frac{2\sqrt{\theta}}{(1 + \sqrt{\theta})^2}.
\]

In the coarse case, Proposition 7(b) gives

\[
P(\xi = 0) = \Phi(0) = \frac{1-\sqrt{\theta}}{1-\theta} = \frac{1}{1+\sqrt{\theta}}.
\]
This (surprisingly?) gives exactly the same conclusion

\[
\lim_n EC_n = 2P(\xi > 0)P(\xi = 0) = \frac{2\sqrt{\theta}}{(1 + \sqrt{\theta})^2}
\]
as in the fine case.
5.3 Branching processes

Instead of sampling a CBP with mutation points and then imposing genera structure on the whole tree, we consider a two-type (types $A$ and $Z$) branching process that allows to sample the resulting tree directly, i.e. the distribution of this process trajectories coincides with the distribution of CBP trees with genera structure on them.

The shape of phylogenetic tree is described by critical binary branching process

$$
\begin{array}{c|c}
\text{parent} \rightarrow \text{offspring} & \text{rate} \\
X \rightarrow X, X & 1 \\
X \rightarrow \emptyset & 1 \\
\end{array}
$$

In order to capture the information on mutation branching points, let’s add particles of type $N$, which do nothing but indicate that the branching point it was emitted from is a mutation branching point.

$$
\begin{array}{c|c}
\text{parent} \rightarrow \text{offspring} & \text{rate} \\
X \rightarrow X, X & \theta \\
X \rightarrow X, X & (1 - \theta) \\
X \rightarrow \emptyset & 1 \\
N \rightarrow \emptyset & 2 \\
\end{array}
$$

(The last transition $N \rightarrow \emptyset$ has rate 2 for convenience reason only).

The branching process (36) is a.s. extinctive. For each finite realisation of (36), considered as a planar tree re-label the edges, changing edge $X$ to $Z$ if there is no $N$-edges below it, and to $A$ otherwise. We get some distribution on three-type ($A, Z, N$) trees. It turns out that this distribution is described by the following branching process

$$
\begin{array}{c|c|c|c|c|c|c}
\text{par.} \rightarrow \text{offspr.} & \text{rate} & \text{par.} \rightarrow \text{offspr.} & \text{rate} \\
A \rightarrow A, A & (1 - \theta)\theta p_{pA}A \cdot n_A & A \rightarrow A, A, N & \theta p_{pA}A \cdot n_A \\
A \rightarrow A, Z & (1 - \theta)\theta p_{pZ}A \cdot n_A & A \rightarrow A, Z, N & \theta p_{pZ}A \cdot n_A \\
A \rightarrow Z, A & (1 - \theta)\theta p_{pZ}A \cdot n_A & A \rightarrow Z, A, N & \theta p_{pZ}A \cdot n_A \\
Z \rightarrow Z, Z & (1 - \theta)\theta p_{pZ}Z \cdot n_Z & A \rightarrow Z, Z, N & \theta p_{pZ}Z \cdot n_Z \\
Z \rightarrow \emptyset & 1 \cdot n_Z & N \rightarrow \emptyset & 2 \\
\end{array}
$$

(37)

where

$$
p_Z = \frac{1}{1 + \sqrt{\theta}}, \quad p_A = 1 - p_Z = \frac{\sqrt{\theta}}{1 + \sqrt{\theta}}.
$$

43
and $n_A, n_Z$ are normalisation terms

$$n_Z = 1 + \sqrt{\theta}, \quad n_A = \frac{1 + \sqrt{\theta}}{\sqrt{\theta}}.$$ 

We prove this using (the recoloring lemma, in Maxim’s earlier version).