Mathematical Biology

CrossMark

Topology and inference for Yule trees with multiple states

Lea Popovic[1] · Mariolys Rivas[1]

Abstract We introduce two models for random trees with multiple states motivated by studies of trait dependence in the evolution of species. Our discrete time model, the *multiple state ERM tree*, is a generalization of Markov propagation models on a random tree generated by a binary search or 'equal rates Markov' mechanism. Our continuous time model, the *multiple state Yule tree*, is a generalization of the tree generated by a pure birth or Yule process to the tree generated by multi-type branching processes. We study state dependent topological properties of these two random tree models. We derive asymptotic results that allow one to infer model parameters from data on states at the leaves and at branch-points that are one step away from the leaves.

Keywords Ancestral tree · Multi-type branching process · Yule tree · Binary search tree · Tree topology · Parameter reconstruction

Mathematics Subject Classification 60J80 · 60G55 · 60G57 · 60J10 · 60J85 · 60J27 · 92D10 · 92D25

1 Introduction

During the past decade there has been considerable activity in studying the effect trait differences may have on the rates of speciation and extinction in the evolution of species (Fitzjohn (2012)) gives an excellent presentation of these recent developments). The possibility that diversification may be trait dependent implies that these rates should not be inferred using standard trait independent methods. New likelihood

---

✉ Lea Popovic
 lea.popovic@concordia.ca; lpopovic@mathstat.concordia.ca

[1] Department of Mathematics and Statistics, Concordia University, Montreal, QC H3G 1M8, Canada

methods that make better use of phylogenetic information were recently developed: "BiSSE" for binary state speciation and extinction (Maddison et al. 2007); "QuaSSE" for quantitative traits (Fitzjohn 2010); "GeoSSE" for geographic character traits (Goldberg et al. 2011), "CLASSE" for punctuated modes of character change (Goldberg and Igic 2012); and were used to make new conclusions about a number of different clades (see a recent survey by NG and Smith 2014).

Inferring the evolutionary process poses in general a non-trivial reconstruction problem, as neither the rates nor the ancestral states in the phylogeny of present day species are known. The underlying ancestral trees are typically assumed to be known and reconstructed from aligned DNA sequence data. Predicting ancestral states is then typically done with one of a number of heuristic methods based on the principle of either: counting, maximum parsimony, or maximum likelihood. In such studies a Markov chain of state changes is assumed to propagate down from the root along the given tree. Many interesting theoretical results exist on the ability to reconstruct ancestral states along the tree and the state at the root from the states observed at the leaves (see the survey of Mossel and Steel 2005 and the paper of Gascuel and Steel 2014 for some recent developments). The focus so far was on reconstructing hidden states along the underlying tree, rather than parameters of the Markov chain which propagates them. In phylogenetics the underlying tree is assumed to be either a random discrete binary tree, or a random Yule tree generated by a neutral pure birth process. The shape of such a tree has the distribution of 'equal rates Markov' (ERM), and the tree resulting after propagating states can be called a *multiple state ERM tree*. We extend the model of propagating states down the tree to also include correlations of states between edges with the same branch-point. As information only on the leaves is insufficient for reconstruction, we will also use correlated substructures of the tree. This is the main reason why we focus on counts of *cherries* and *pendants* of different types (c.f. next page), as these objects will prove to be sufficient in identifying parameters of the underlying Markov chain which propagated the multiple states on the ERM tree.

Inference for evolutionary processes whose birth and death rates are trait dependent adds an additional layer of mathematical difficulty. If we have a trait with finitely many variants (or a continuum of variants is discretized into finitely many bins) the full tree evolves according to a multi-type branching process, in which rates of speciation to different offspring 'types' (representing different states) and the rate of extinction is specific to the state of that lineage. In such a branching process the shape of the tree and its edge lengths are inseparable from the distribution of states on the lineages. Both the ratio of speciation to extinction rates for each state, as well as the transitions from a certain state to another play an important role in how the ancestral states are distributed along the tree. The ancestral tree of this branching process, obtained by pruning away the extinct lineages, turns out to be a random tree we call *multiple state Yule tree*. In such a tree the chance of a lineage splitting is state dependent. Consequently, this leaves a signature in terms of which splits are more frequent than others and is reflected in the proportion of different types of splits at the tips of the tree. Fortunately again, long term limit of counts of the different types of *cherries* and *pendants* prove (under some mild assumptions) to be sufficient tools for uncovering the underlying process which generated the multiple state Yule tree. We will use this information in the reconstruction of model parameters.

Very few practically applicable results have been obtained from theoretical analyses of ancestral trees of multi-type branching processes. Deriving an exact distribution for the ancestral tree is unsurprisingly challenging, as determining the likelihood of any split requires the knowledge of the parental 'type' (state), and hence also all ancestral states on that lineage. Popovic and Rivas (2014) developed a coalescent point-process approach to generating ancestral trees using the tips in an infinite (quasi-stationary) multi-type Galton-Watson branching processes. This construction relied on a horizontal exploration of the tips which was developed by Aldous and Popovic (2005) and extended by Lambert and Popovic (2013) (the standard vertical coalescent construction is not possible for branching processes with state-dependent offspring distributions). It can be used for simulating and computing likelihood of ancestral trees, but calculating its statistical features is not easy, except in some very special cases.

In this paper, we focus on analyzing newly introduced a priori models on possible ancestral tree shapes. Our *multiple state ERM tree* is a discrete time model that is an extension of Markov propagation models on a random tree generated by a mechanism which picks a random leaf to extend on. Our *multiple state Yule tree* is a continuous time model and is a generalization of the tree generated by a pure birth process to one with multiple states and state transitions. In order to investigate the topological features of these two new models we analyze the number of different types of cherries and different types of pendants in the tree: *cherries* are pairs of leaves each of which is adjacent to a common ancestor, and *pendants* are edges of leaves whose immediate ancestor is not a single edge away from another leaf (in other words, pendants are edges to all the leaves that are not in a cherry). We use the random recursive mechanisms for generating splits in the trees to obtain exact results for finite sized trees, as well as asymptotic results as the trees grow in size. The distribution of the number of pendants and cherries in the tree reflects the model parameters and can be used to infer them, justifying the applicability of our main results.

For the multiple state ERM random trees, we focus on the numbers of cherries of the different types. We identify the means and variances of these random variables (Propositions 3 and 5), and we also derive asymptotic results as the number of leaves in the tree grows (Theorems 6 and 9). We use the limiting fraction of different cherries to infer the probabilities in the model (Corollary 11). Examples of particular models for multiple state ERM trees are discussed in Sect. 2.2. For the multi-type branching process we first identify the process obtained by pruning away the extinct lineages (Proposition 13) as a specific version of a multiple state Yule tree with state transitions. Using the distribution of states at the leaves (Lemma 16) we find the distribution of different cherries and pendants (Propositions 19 and 21) in a general multiple state Yule tree. We also derive their asymptotics in the long time-scale limit (Theorem 23) and provide the way in which the original speciation and extinction rates can be inferred from these topological features (Corollary 25).

2 Multiple state ERM trees

Consider a (single state) random tree constructed recursively, from a single node leaf, by picking at each step one leaf uniformly at random and creating a branch-point by

attaching two new leaves to it. The distribution of this tree is called 'equal rates Markov' (ERM) (first investigated by Harding 1971) and has a long list of mathematical results associated to it (Aldous 1996, 2001). Trees with this distribution can be generated in a number of different ways, forwards in time—by using a (pure birth) Yule process stopped the first time it reaches a prescribed number of leaves and ignoring the random lengths of its branches, or backwards in time—starting from a prescribed number of leaves using a neutral (coalescent) Moran process. Trees with this distribution have been used in numerous studies as a null model in investigating patterns in tree shapes (Mooers and Heard 1997). In terms of its statistical features the number of cherries C_n for a tree with n leaves is known (McKenzie and Steel 2000) to have the following properties:

$$\mathbb{E}[C_n] = \frac{n}{3} \ \text{ for } \ n \geq 3, \quad \mathbb{V}[C_n] = \frac{2n}{45} \ \text{ for } \ n \geq 5;$$

where $\mathbb{V}[C_n] = \text{Var}[C_n]$, and the distribution satisfies a central limit theorem:

$$\frac{C_n - n/3}{\sqrt{2n/45}} \Rightarrow N(0, 1).$$

These results were shown using an extended Pólya urn process by Smythe (1996) and Janson (2004).

We consider a multiple state version of this random tree, where each node (branch-points and leaves) has a state $k \in \mathcal{K}$ associated with it. The shape of the tree is constructed in the same way as in the single state process, with each leaf having the same chance, regardless of its state, of being picked at random to create the next branch-point with two new leaves attached to it. The states of the two leaves being attached, however, depend on the state of the leaf that they are being attached to. For each state i, $j_1, j_2 \in \mathcal{K}$ the probabilities $q_i^{j_1,j_2}$ determine the chance that a leaf of state i has states j_1, j_2 attached to it. Since we do not distinguish between different embeddings of the tree in the plane, we can w.l.o.g. assume $j_1 \leq j_2$. We call this tree a *multiple state ERM tree*. The random tree with multiple states is distributed as a Markov field (with propagation matrix $\{q_i^{j_1,j_2}\}_{i, j_1 \leq j_2 \in \mathcal{K}}$) on an ERM tree. Note that, for each i, $\sum_{j_1 \leq j_2} q_i^{j_1,j_2} = 1$. In order to avoid trivial cases that generate only single state trees we will assume throughout that $q_i^{ii} \neq 1, \forall i$.

For the sake of simplicity we consider $\mathcal{K} = \{1, 2\}$. There are $k^2(k + 1)/2 = 6$ different types of cherries $\{111, 112, 122, 211, 212, 222\}$ and $k^2 = 4$ different types of pendants $\{11, 12, 21, 22\}$, as illustrated in Fig. 1. Figure 2 illustrates cherries in an example of a tree with $n = 5$ leaves.

2.1 Moments of the number of different types of cherries

For a tree with n leaves we let $N_1(n)$ denote the number of state 1 leaves, $N_2(n) = n - N_1(n)$ the number of state 2 leaves, and $C_i^{j_1 j_2}(n)$ the number of cherries of type $i j_1 j_2$. Their means are relatively straightforward to calculate.
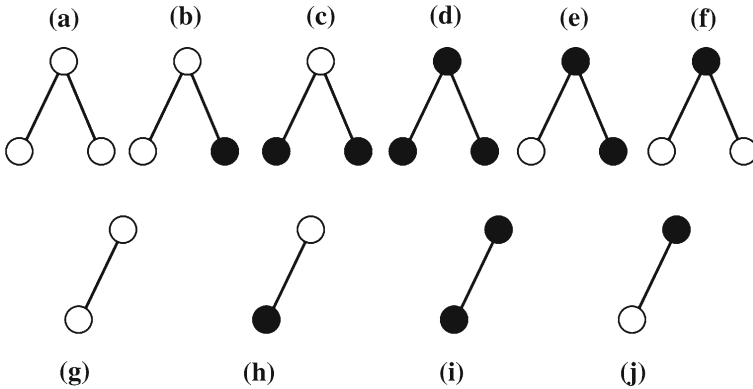
Fig. 1 Type 1 is denoted by a *blank circle*, and type 2 by a *full circle*; different types of cherries: **a** type 111, **b** type 112, **c** type 122, **d** type 222, **e** type 212 and **f** type 211; and different types of pendants: **g** type 11, **h** type 12, **i** type 22 and **j** type 21

Fig. 2 Tree with $N_1(5) = 3$, $N_2(5) = 2$, one cherry of type 211, one cherry of type 212, and one pendant of type 22, where type 1 is denoted by a *blank circle*, and type 2 by a *full circle*
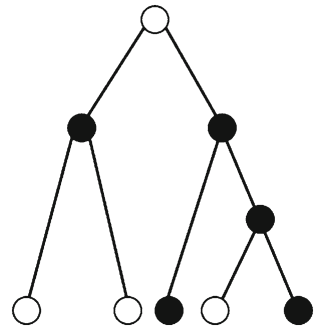
Lemma 1 *Assume the probabilities* $\{q_i^{j_1 j_2}\}_{i, j_1 \leq j_2 \in \{1,2\}}$ *satisfy* ($\star$): $c_1 - c_2 \notin \{-2, 2\}$ *for* $c_1 := 2q_1^{11} + q_1^{12}$ *and* $c_2 := 2q_2^{11} + q_2^{12}$*. Then,* $\forall n \geq 3$,

$$v_1(n) := \mathbb{E}[N_1(n)] = \frac{c_2 n}{2 - c_1 + c_2}$$
$$- \frac{(2c_2 - (2 - c_1 + c_2)v_1(2))\Gamma(n - 1 + c_1 - c_2)}{(2 - c_1 + c_2)\Gamma(c_1 - c_2 + 2)\Gamma(n)},$$

where $\Gamma(n)$ *is the gamma function, and* $v_1(2) = \begin{cases} c_1, & \text{if } N_1(1) = 1 \text{ (initial leaf state is 1)} \\ c_2, & \text{if } N_2(1) = 1 \text{ (initial leaf state is 2)}. \end{cases}$
Analogous formula holds for $v_2(n) := \mathbb{E}[N_2(n)]$ *in which:* c_1 *is replaced by* $c_1' := 2q_2^{22} + q_2^{12}$ ($= 2 - c_2$)*,* c_2 *is replaced by* $c_2' := 2q_1^{22} + q_1^{12}$ ($= 2 - c_1$) *(notice* $c_1' - c_2' = c_1 - c_2$ *remains the same), and* $v_1(2)$ *is replaced by*
$v_2(2) = \begin{cases} c_1', & \text{if } N_2(1) = 1 \text{ (initial leaf state is 2)} \\ c_2', & \text{if } N_1(1) = 1 \text{ (initial leaf state is 1)}. \end{cases}$

Proof The result follows from a straightforward recursion, for any $2 \leq n_1 \leq n$, we have

$$
\mathbb{P}[N_1(n) = n_1] = \left(\frac{n_1 q_1^{12}}{n-1} + \frac{(n-n_1-1)q_2^{22}}{n-1} \right) \mathbb{P}[N_1(n-1) = n_1]
$$
$$
+ \left(\frac{(n_1-1)q_1^{11}}{n-1} + \frac{(n-n_1)q_2^{12}}{n-1} \right) \mathbb{P}[N_1(n-1) = n_1 - 1]
$$
$$
+ \left(\frac{(n-n_1+1)q_2^{11}}{n-1} \right) \mathbb{P}[N_1(n-1) = n_1 - 2]
$$
$$
+ \left(\frac{(n_1+1)q_1^{22}}{n-1} \right) \mathbb{P}[N_1(n-1) = n_1 + 1].
$$

This yields a recurrence relation for $G_n(x) = \sum_{n_1 \geq 0} \mathbb{P}[N_1(n) = n_1] x^{n_1}$, which when differentiated and evaluated at $x = 1$ results in the recurrence relation for $v_1(n)$

$$
v_1(n+1) = \left(q_2^{12} + 2q_2^{11} \right) + \frac{1}{n} \left(n + q_1^{12} + 2q_1^{11} - q_2^{12} - 2q_2^{11} - 1 \right) v_1(n) \quad (1)
$$

and solving it we obtain the claimed result.

Remark 2 The condition $c_1 - c_2 \neq 2$ rules out trivial cases generating single state trees $\{q_1^{11} = 1, q_2^{22} = 1\}$ of only state 1 or state 2 (depending on initial state). The condition $c_1 - c_2 \neq -2$ rules out the unusual special case of completely alternating states $\{q_2^{22} = 1, q_1^{11} = 1\}$. However, a number of interesting cases are covered by our results, as shown at the end of this section.

Proposition 3 *Under the same conditions* (⋆) *as in Lemma* 1, $\forall n \geq 3$, *for*

$$
\mu_1^{11}(n) := \mathbb{E}\left[C_1^{11}(n) \right], \quad \mu_1^{12}(n) := \mathbb{E}\left[C_1^{12}(n) \right], \quad \mu_1^{22}(n) := \mathbb{E}\left[C_1^{22}(n) \right]
$$

we have

$$
\mu_1^{11}(n) = \frac{3(2 - c_1 + c_2)(2\mu_1^{11}(3) - q_1^{11}v_1(2)) + n(n-1)(n-2)q_1^{11}c_2}{3(2 - c_1 + c_2)(n-1)(n-2)} - q_1^{11}C(n)
$$
$$
\mu_1^{12}(n) = \frac{3(2 - c_1 + c_2)(2\mu_1^{12}(3) - q_1^{12}v_1(2)) + n(n-1)(n-2)q_1^{12}c_2}{3(2 - c_1 + c_2)(n-1)(n-2)} - q_1^{12}C(n)
$$
$$
\mu_1^{22}(n) = \frac{3(2 - c_1 + c_2)(2\mu_1^{22}(3) - q_1^{22}v_1(2)) + n(n-1)(n-2)q_1^{22}c_2}{3(2 - c_1 + c_2)(n-1)(n-2)} - q_1^{22}C(n)
$$

where $v_1(2)$, c_1, c_2 *are as in Lemma* 1, *the constants* $C(n)$ *are*

$$
C(n) := \frac{(2c_2 - (2 - c_1 + c_2)v_1(2)) \, \Gamma(n - 1 + c_1 - c_2)}{(2 - c_1 + c_2)\Gamma(c_1 - c_2 + 2)\Gamma(n)},
$$

and the initial values are

$$\mu_1^{11}(3) = \begin{cases} \left(q_1^{11}\right)^2 + q_1^{11}q_1^{12}/2 & \text{if } N_1(1) = 1 \\ q_1^{11}q_2^{11} + q_1^{11}q_2^{12}/2 & \text{if } N_2(1) = 1, \end{cases}$$

$$\mu_1^{12}(3) = \begin{cases} \left(q_1^{12}\right)^2/2 + q_1^{11}q_1^{12} & \text{if } N_1(1) = 1 \\ q_2^{11}q_1^{12} + q_2^{12}q_1^{12}/2 & \text{if } N_2(1) = 1 \end{cases}$$

$$\mu_1^{22}(3) = \begin{cases} q_1^{11}q_1^{22} + q_1^{12}q_1^{22}/2 & \text{if } N_1(1) = 1 \\ q_2^{12}q_1^{22}/2 + q_2^{11}q_1^{22} & \text{if } N_2(1) = 1. \end{cases}$$

Analogous formulae hold for

$$\mu_2^{11}(n) := \mathbb{E}\left[C_2^{11}(n)\right], \quad \mu_2^{12}(n) := \mathbb{E}\left[C_2^{12}(n)\right], \quad \mu_2^{22}(n) := \mathbb{E}\left[C_2^{22}(n)\right]$$

in which: probabilities $q_1^{j_1 j_2}$ are replaced by $q_2^{j_1 j_2}$, $v_1(2)$ is replaced by $v_2(2) = 2 - v_1(2)$, c_1 and c_2 are replaced by c_1' and c_2' respectively (as in Lemma 1), the constants $C(n)$ remain the same if the initial state is interchanged, and $\mu_1^{j_1 j_2}(3)$ are replaced by $\mu_2^{j_1 j_2}(3)$ obtained by fully interchanging states in the formulae for $\mu_1^{j_1 j_2}(3)$.

Proof Since at each step new leaves are attached in pairs, there is no need to keep track of the number of different pendants. It suffices to keep track of the number of different states of leaves and only of the cherries of the specific type we are trying to calculate. Let

$$f_n^{i j_1 j_2}(n_1, k) := \mathbb{P}\left[N_1(n) = n_1, C_i^{j_1 j_2}(n) = k\right]$$

and

$$F_n^{i j_1 j_2}(x, y) = \sum_{n_1 \geq 0, k \geq 0} f_n^{i j_1 j_2}(n_1, k) x^{n_1} y^k.$$

Using recursion arguments, for any $3 \leq n_1 \leq n, k \geq 1$, we have

$$\begin{aligned} f_n^{111}(n_1, k) = & \left(\frac{2k q_1^{11}}{n-1} + \frac{(n-n_1)q_2^{12}}{n-1}\right) f_{n-1}^{111}(n_1 - 1, k) \\ & + \left(\frac{(n_1 - 2k)q_1^{12}}{n-1} + \frac{(n-n_1-1)q_2^{22}}{n-1}\right) f_{n-1}^{111}(n_1, k) \\ & + \left(\frac{(n-n_1+1)q_2^{11}}{n-1}\right) f_{n-1}^{111}(n_1 - 2, k) \\ & + \left(\frac{(n_1 + 1 - 2k)q_1^{22}}{n-1}\right) f_{n-1}^{111}(n_1 + 1, k) \end{aligned}$$

$$+ \left(\frac{(n_1 - 1 - 2(k-1))q_1^{11}}{n-1} \right) f_{n-1}^{111}(n_1 - 1, k - 1)$$

$$+ \left(\frac{2(k+1)q_1^{22}}{n-1} \right) f_{n-1}^{111}(n_1 + 1, k + 1)$$

$$+ \left(\frac{2(k+1)q_1^{12}}{n-1} \right) f_{n-1}^{111}(n_1, k + 1),$$

for cherries of type 111,

$$f_n^{112}(n_1, k) = \left(\frac{kq_1^{12}}{n-1} + \frac{(n - n_1 - k - 1)q_2^{22}}{n-1} \right) f_{n-1}^{112}(n_1, k)$$

$$+ \left(\frac{(n_1 - k - 1)q_1^{11}}{n-1} + \frac{(n - n_1 - k)q_2^{12}}{n-1} \right) f_{n-1}^{112}(n_1 - 1, k)$$

$$+ \left(\frac{(n - n_1 - k + 1)q_2^{11}}{n-1} \right) f_{n-1}^{112}(n_1 - 2, k)$$

$$+ \left(\frac{(k+1)q_2^{11}}{n-1} \right) f_{n-1}^{112}(n_1 - 2, k + 1)$$

$$+ \left(\frac{(k+1)q_2^{12}}{n-1} + \frac{(k+1)q_1^{11}}{n-1} \right) f_{n-1}^{112}(n_1 - 1, k + 1)$$

$$+ \left(\frac{(k+1)q_2^{22}}{n-1} \right) f_{n-1}^{112}(n_1, k + 1)$$

$$+ \left(\frac{(n_1 - k + 1)q_1^{12}}{n-1} \right) f_{n-1}^{112}(n_1, k - 1)$$

$$+ \left(\frac{(n_1 - k + 1)q_1^{22}}{n-1} \right) f_{n-1}^{112}(n_1 + 1, k)$$

$$+ \left(\frac{(k+1)q_1^{22}}{n-1} \right) f_{n-1}^{112}(n_1 + 1, k + 1)$$

for cherries of type 112, and

$$f_n^{122}(n_1, k) = \left(\frac{(n - n_1 - 2k)q_2^{12}}{n-1} + \frac{(n_1 - 1)q_1^{11}}{n-1} \right) f_{n-1}^{122}(n_1 - 1, k)$$

$$+ \left(\frac{(n - n_1 - 2k - 1)q_2^{22}}{n-1} + \frac{n_1 q_1^{12}}{n-1} \right) f_{n-1}^{122}(n_1, k)$$

$$+ \left(\frac{2(k+1)q_2^{11}}{n-1} \right) f_{n-1}^{122}(n_1 - 2, k + 1)$$

$$+ \left(\frac{2(k+1)q_2^{12}}{n-1} \right) f_{n-1}^{122}(n_1 - 1, k + 1)$$

$$+ \left(\frac{2(k+1)q_2^{22}}{n-1} \right) f_{n-1}^{122}(n_1, k + 1)$$

$$+ \left(\frac{(n_1+1)q_1^{22}}{n-1} \right) f_{n-1}^{122}(n_1 + 1, k - 1)$$

$$+ \left(\frac{(n - n_1 - 2k + 1)q_2^{11}}{n-1} \right) f_{n-1}^{122}(n_1 - 2, k).$$

for cherries of type 122. Each of these equations yields a recurrence relation for the corresponding joint generating function $F_n^{ij_1 j_2}(x, y)$ by summing over n_1 and k. Differentiating and evaluating them at $x = y = 1$ then provides recurrences for the means $\mu_i^{j_1, j_2}(n)$

$$\mu_1^{11}(n+1) = \frac{n-2}{n}\mu_1^{11}(n) + \frac{q_1^{11}}{n}\nu_1(n)$$

$$\mu_1^{12}(n+1) = \frac{n-2}{n}\mu_1^{12}(n) + \frac{q_1^{12}}{n}\nu_1(n) \qquad (2)$$

$$\mu_1^{22}(n+1) = \frac{n-2}{n}\mu_1^{22}(n) + \frac{q_1^{22}}{n}\nu_1(n)$$

solving which, with the expression for $\nu_1(n)$ from Lemma 1, gives the claimed formulae.

Remark 4 Simple algebra shows that the mean numbers of all cherries $\sum_{i, j_1 \leq j_2} \mu_1^{j_1 j_2}$ (n) add up to $n/3$, corresponding to the known mean number of cherries in a single-state ERM tree.

Proposition 5 *Assume that $c_1 - c_2 \notin \{-2, -1, 0, 1, 3/2, 2\}$ for c_1, c_2 as in Lemma 1. Then, $\forall n \geq 5$, for*

$$\sigma_1^{11}(n) := \mathbb{V}\left[C_1^{11}(n) \right], \quad \sigma_1^{12}(n) := \mathbb{V}\left[C_1^{12}(n) \right], \quad \sigma_1^{22}(n) := \mathbb{V}\left[C_1^{22}(n) \right]$$

we have

$$\sigma_1^{11}(n), \sigma_1^{12}(n), \sigma_1^{22}(n) \sim O(n) + O(n^{c_1 - c_2 - 1}) + O(n^{2(c_1 - c_2 - 1)}).$$

The same asymptotics hold for

$$\sigma_2^{11}(n) := \mathbb{V}\left[C_2^{11}(n) \right], \quad \sigma_2^{12}(n) := \mathbb{V}\left[C_2^{12}(n) \right], \quad \sigma_2^{22}(n) := \mathbb{V}\left[C_2^{22}(n) \right]$$

as the exponents $c_1' - c_2' = c_1 - c_2$ are the same in these cases.

Proof Using recurrence relations for the generating functions $F_n^{111}(x, y)$, $F_n^{112}(x, y)$, $F_n^{122}(x, y)$ from the proof of Proposition 3, taking second derivatives in x, y and evaluating them at $x = y = 1$ yields recurrence equations for the variances for the number of cherries for each of the types 111, 112 and 113, respectively. For $j_1 \leq j_2 \in \{1, 2\}$ let

$$R_{yy}^{1j_1j_2}(n) := \frac{\partial^2 F_n^{1j_1j_2}(x, y)}{\partial y^2}\bigg|_{x=1,y=1}, \quad R_{xy}^{1j_1j_2}(n) := \frac{\partial^2 F_n^{1j_1j_2}(x, y)}{\partial x \partial y}\bigg|_{x=1,y=1}.$$

From equations for $F_n^{1j_1j_2}(x, y)$ we obtain the recurrence relation for $R_{yy}^{1j_1j_2}(n)$, $R_{xy}^{1j_1j_2}(n)$ as

$$R_{yy}^{1j_1j_2}(n)$$
$$= \frac{\Gamma(n-4)}{\Gamma(n)}\left(\sum_{n_1=1}^{n-1}\frac{2q_1^{j_1j_2}(R_{xy}^{1j_1j_1}(n_1) - 2\mu_1^{j_1j_2}(n_1))\Gamma(n_1+1) + 24R_{yy}^{1j_1j_2}(5)}{n_1\Gamma(n_1-3)}\right)$$

The equations for $R_{xy}^{1j_1j_2}(n)$ satisfy the recurrence relations, for any $n \geq 4$

$$R_{xy}^{111}(n+1) = \frac{1}{n}\Big((2-2c_1+c_2n)\mu_1^{11}(n) + (n-3+c_1-c_2)R_{xy}^{111}(n)$$
$$+2q_1^{11}v_1(n) + q_1^{11}R(n)\Big)$$

$$R_{xy}^{112}(n+1) = \frac{1}{n}\Big((1-c_1+c_2n-c_2)\mu_1^{12}(n) + (n-3+c_1-c_2)R_{xy}^{112}(n)$$
$$+2q_1^{12}v_1(n) + q_1^{11}R(n)\Big)$$

$$R_{xy}^{122}(n+1) = \frac{1}{n}\Big((c_2n-2c_2)\mu_1^{22}(n) + (n-3+c_1-c_2)R_{xy}^{112}(n) + q_1^{22}R(n)\Big)$$

Recall that $G_n(x) = \sum_{n_1 \geq 0}\mathbb{P}[N_1(n) = n_1]x^{n_1}$. Thus, $R(n) = \frac{\partial^2 G_n}{\partial x^2}|_{x=1,y=1}$ is the second moment for the number of leaves N_1 of state 1 (see Lemma 1) and satisfies the recurrence, for $n \geq 3$

$$R(n) = \frac{1}{n}\Big(2nq_2^{11}+(2nc_2+2q_1^{11}-6q_2^{11} - 2q_2^{12})v_1(n) + (n-2+2c_1-2c_2)R(n)\Big)$$

From this we have,

$$\sigma_i^{j_1j_2}(n) = R_{yy}^{ij_1j_2}(n) + \mu_i^{j_1j_2}(n) - (\mu_i^{j_2j_2}(n))^2 \tag{3}$$

Explicitly solving these equations for the variances requires much more complicated calculations than for the means. Using Maple yields formulae which are quite long and cluttered. However, expanding these formulae with respect to n we obtain the asymptotic results above.

2.2 Special cases of multiple state ERM models

To illustrate how multiple state ERM trees for different speciation models give different cherry distributions, we consider particular cases corresponding to specific values of $\{q_i^{j_1 j_2}\}_{i, j_1 \leq j_2 \in \{1,2\}}$:

(a) The *'single state'* model is the trivial one in which the only state in the tree is the initial one: $q_1^{11} = q_2^{22} = 1$ $(c_1 - c_2 = 2)$;

(b) The *'alternating state'* model is one in which one state can only attach to itself leaves of the other state: $q_1^{22} = q_2^{11} = 1$ $(c_1 - c_2 = -2)$

(c) In the *'neutral to state'* model the branch-point state does not determine the probabilities of leaf states: for each $j_1 \leq j_2$: $q_i^{j_1 j_2}$ is independent of whether $i = 1, 2$ $(c_1 - c_2 = 0, c_1 + c_1' = 2)$;

(d) In the *'only mixed state'* model each state has only mixed states attached to it: $q_i^{12} = 1$ for both $i = 1, 2$ $(c_1 - c_2 = 0)$;

(e) The *'asymmetric change in state'* represents a model where one state can be randomly gained from the other but once gained can no longer be lost: $\{q_1^{11} = 1, q_2^{11} = q_2^{22} = (1 - q_2^{12})/2\}$ or $\{q_2^{22} = 1, q_1^{11} = q_1^{22} = (1 - q_1^{12})/2\}$ $(c_1 - c_2 = 1)$;

Single state

$q_1^{11} = q_2^{22} = 1$ $(c_1 - c_2 = 2)$

$\mu_1^{11}(n) = \dfrac{n}{3}$ all other $\mu_i^{j_1 j_2} = 0$, if $N_1(1) = 1$

$\mu_2^{22}(n) = \dfrac{n}{3}$ all other $\mu_i^{j_1 j_2} = 0$, if $N_2(1) = 1$

Alternating state

$q_1^{22} = q_2^{11} = 1$ $(c_1 - c_2 = -2)$

$\mu_1^{22}(n) = \mu_2^{11}(n) = \dfrac{n}{6}, n \geq 4$ all other $\mu_i^{j_1 j_2} = 0$

Neutral to state

$q_1^{11} = q_2^{11}, q_1^{12} = q_2^{12}, q_1^{22} = q_2^{22}$

$\mu_1^{11}(n) = \dfrac{n q_1^{11} c_1}{6}, \mu_2^{11}(n) = \dfrac{n q_2^{11} c_1'}{6},$

$(c_1 - c_2 = 0)$

$\mu_1^{12}(n) = \dfrac{n q_1^{12} c_1}{6}, \mu_2^{12}(n) = \dfrac{n q_2^{12} c_1'}{6},$

$\mu_1^{22}(n) = \dfrac{n q_1^{22} c_1}{6}, \mu_2^{22}(n) = \dfrac{n q_2^{22} c_1'}{6}$

Only mixed state

$q_1^{12} = q_2^{12} = 1$ $(c_1 - c_2 = 0)$

$\mu_1^{12}(n) = \mu_2^{12}(n) = \dfrac{n}{6}$, all other $\mu_i^{j_1 j_2} = 0$

Asymmetric change in state

$q_1^{11} = 1, q_2^{11} = q_2^{22}$ $(c_1 - c_2 = 1)$

$\mu_1^{11}(n) = \dfrac{n}{3}$, all other $\mu_i^{j_1 j_2} = 0$,

if $N_1(1) = 1$

$\mu_1^{11}(n) = \dfrac{n}{3} - \dfrac{1}{2}, \mu_2^{11} = \mu_2^{22} = \dfrac{1}{4}(1 - q_2^{12}),$

$\mu_2^{12} = q_2^{12}$, if $N_2(1) = 1$

For cases (a), (b) we could not use Proposition 3 and we calculated the means directly from the recurrence relations (1) and (2). In case (b), the mean number of cherries for $n = 3$ is $\mu_1^{22}(3) = 0$, $\mu_2^{11}(3) = 1$ if $N_1(1) = 1$; $\mu_1^{22}(3) = 1$, $\mu_2^{11}(3) = 0$ if $N_2(1) = 1$, but from $n \geq 4$ is equal to $n/6$ regardless of the initial state. The value of $c_1 - c_2 = c_1' - c_2' \in [-2, 2]$ reflects the extent to which leaves attach to leaves of

the same state; the higher it is, the more likely a leaf is to attach to a leaf of the same state, the two extreme cases are (a) and (b). The sum of means for all different types of cherries is $n/3$, as found by McKenzie and Steel (2000) for a single state ERM tree.

We can also calculate exact values for the variances of the numbers of cherries from (3) instead of relying only on asymptotics as in Proposition 5. The sum of variances for all different types of cherries coincides with the variance $2n/45$ of a single state ERM tree (McKenzie and Steel 2000) only in extreme cases (a), (b) when covariances of different cherry types are zero.

Single type

$q_1^{11} = q_2^{22} = 1 \ (c_1 - c_2 = 2)$

$\sigma_1^{11}(n) = \dfrac{2n}{45}$, all other $\sigma_i^{j_1 j_2}(n) = 0$, if $N_1(1) = 1$

$\sigma_2^{22}(n) = \dfrac{2n}{45}$, all other $\sigma_i^{j_1 j_2}(n) = 0$, if $N_2(1) = 1$

Alternating type

$q_1^{22} = q_2^{11} = 1 \ (c_1 - c_2 = -2)$

$\sigma_1^{22}(n) = \sigma_2^{11}(n) = \dfrac{2n}{45}, n \geq 4$ all other $\sigma_i^{j_1 j_2} = 0$

Neutral to type

$q_1^{11} = q_2^{11}, \ q_1^{12} = q_2^{12}, \ q_1^{22} = q_2^{22}$

$\sigma_1^{11}(n) = \dfrac{nq_1^{11}\left(6(q_1^{11})^2 + 15c_1 - 8q_1^{11}c_1^2\right)}{90},$

$(c_1 - c_2 = 0)$

$\sigma_1^{12}(n) = \dfrac{nq_1^{12}\left(6q_1^{11}q_1^{12} + 15c_1 - 8q_1^{12}c_1^2\right)}{90},$

$\sigma_1^{22}(n) = \dfrac{nq_1^{22}\left(6q_1^{11}q_1^{22} + 15c_1 - 8q_1^{22}c_1^2\right)}{90},$

Only mixed type

$q_1^{12} = q_2^{12} = 1 \ (c_1 - c_2 = 0)$

$\sigma_1^{12}(n) = \sigma_2^{12}(n) = \dfrac{7n}{90},$ all other $\sigma_i^{j_1 j_2}(n) = 0$

Asymmetric change

$q_1^{11} = 1, q_2^{11} = q_2^{22} \ (c_1 - c_2 = 1)$

$\sigma_1^{11}(n) = \dfrac{2n}{45}$, all other $\sigma_i^{j_1 j_2} = 0,$
if $N_1(1) = 1$

$\sigma_1^{11}(n) = \dfrac{2n}{45} + o(n), \ \sigma_2^{11}(n) = \sigma_2^{22}(n) = \frac{1}{16}\left(1 - q_2^{11}\right)^3,$

$\sigma_2^{12}(n) = \frac{1}{4}\left(q_2^{11}\right)^2\left(1 - q_2^{11}\right),$ if $N_2(1) = 1$

2.3 Asymptotic results for the number of cherries and pendants

To consider the full structure (with correlations) of all the cherries in a multiple state ERM, we also need to keep track of the number of different pendants $L_i^j(n)$ of type ij in a tree with n leaves. Let $X(n)$ be a single vector representing different types of cherries and pendants

$$X(n) = \left(C_1^{11}(n), C_1^{12}(n), C_1^{22}(n), C_2^{22}(n),\right.$$

$$\left. C_2^{12}(n), C_2^{11}(n), L_1^1(n), L_1^2(n), L_2^2(n), L_2^1(n)\right)$$

Its asymptotic behaviour as $n \to \infty$ can be characterized in terms of a strong law.

Theorem 6 *Assume the probabilities $\{q_i^{j_1 j_2}\}_{i,\, j_1 \leq j_2 \in \{1,2\}}$ are such that to every cherry it is possible to eventually attach every other cherry* (∗). *Then, as $n \to \infty$*

$$\frac{X_n}{n} \xrightarrow{\text{a.s}} v_1 := \frac{1}{3(2 - c_1 + c_2)} \begin{bmatrix} q_1^{11} c_2 \\ q_1^{12} c_2 \\ q_1^{22} c_2 \\ q_2^{11}(2 - c_1) \\ q_2^{12}(2 - c_1) \\ q_2^{22}(2 - c_1) \\ (c_1 c_2)/2 \\ (2 - c_1)c_2/2 \\ (2 - c_1)(2 - c_2)/2 \\ (2 - c_1)c_2/2 \end{bmatrix}.$$

where $c_1 := 2q_1^{11} + q_1^{12}$, $c_2 := 2q_2^{11} + q_2^{12}$ are as in Lemma 1.

Remark 7 The condition (∗) is a form of *irreducibility* of the cherry state space. It can be relaxed for multiple state tree models in which certain types of cherries are not at all appearing in the tree. We get the same strong law results on a state space (the vector X, the matrix A) that is restricted to the set of cherries that can appear in the tree, on which the condition (∗) holds.

Proof The proof relies on a Pólya urn representation of the different types of cherries and pendants: an *extended Pólya urn process* $(X(n))_{n \geq 0}$ is a Markov chain on $\mathbb{Z}_+^d$ where the coordinates of the random vector $X(n) = (X_1(n), \ldots, X_d(n))$ represent the number of balls of type $i \in \{1, \ldots, d\}$ in an urn at step n. The process starts at $X(0)$ and at each step balls of different types are added or removed from it. Each ball type has associated to it a positive weight $a_i \geq 0$, $i \in \{1, \ldots, d\}$ and a random vector $\xi_i = (\xi_{i1}, \ldots, \xi_{il})$ taking values in $\mathbb{Z}_+^d$, such that: $\xi_{ij} \geq 0$, $\forall j \neq i$ and $\xi_{ii} \geq -1$, $\forall i$ as well as $\mathbb{E}(\xi_{ij}^2) < \infty$.

The weights and random vectors together characterize the distribution of the transition matrix for the Markov chain:

(i) at each step a ball is randomly selected from the urn with the probability of selecting a ball of type i proportional to its weight a_i, that is the probability of drawing a ball of type i at time $n \geq 1$ is $a_i X_i(n-1)/\sum_j a_j X_j(n-1)$;

(ii) if a ball of type i was selected, then the number of balls of different types to be added to the urn is drawn according to the distribution ξ_{ij}, $j = 1, \ldots, d$. The condition $\xi_{ii} \geq -1$ means the selected ball that is removed from the urn may or may not be replaced on that step. It is useful to assume the urn never becomes empty, $|X(n)| > 0$, $\forall n \geq 0$. Let $a = (a_1, \ldots, a_d)$. The generating matrix of a Pólya urn is defined as $A := (a_j \mathbb{E}(\xi_{ji}))_{i,j=1}^d$, whose eigenvalues in decreasing

order of real parts are denoted by $\lambda_1 > \text{Re}(\lambda_2) \geq \text{Re}(\lambda_3) \ldots$ (Perron-Frobenius implies that λ_1 is real valued). The urn is called irreducible if, for any i, j, given the urn starts with a single ball of type i it is eventually possible to add a ball of type j to the urn.

A complete treatment of extended Pólya urns is given by Janson (2004). We state here only the results that are key for our proof. Assume the urn is such that: (a) it is irreducible; (b) $\lambda_1 > 0$, (c) λ_1 and λ_2 are simple eigenvalues with left and right eigenvectors $\boldsymbol{u}_1, \boldsymbol{v}_1$ and $\boldsymbol{u}_2, \boldsymbol{v}_2$ satisfying $\boldsymbol{u}_1 \cdot \boldsymbol{v}_1 = \boldsymbol{u}_2 \cdot \boldsymbol{v}_2 = 1$ and $\boldsymbol{a} \cdot \boldsymbol{v}_1 = 1$; (d) $\text{Re}(\lambda_2) > \text{Re}(\lambda_3)$. The last condition implies that the set of eigenvectors λ satisfying $\text{Re}(\lambda) > \lambda_1/2$ consists either only of λ_2 or it is empty. Under these assumptions Theorem 3.21 in the paper by Janson (2004) insures that, in the limit as $n \to \infty$,

$$\frac{\boldsymbol{X}_n}{n} \xrightarrow{\text{a.s}} \boldsymbol{v}_1.$$

The process of constructing a multiple state ERM tree can be viewed as a Pólya urn process: the balls of different types are all the different types of cherries and different types of pendants. For $\mathcal{K} = \{1, 2\}$ we have $d = 10$, as shown in Fig. 1. The ball types corresponding to any of the cherries have a weight $a_i = 2$, and those corresponding to pendants have a weight $a_i = 1$, as it is twice as likely to choose a cherry than a pendant when a leaf is picked uniformly at random. Careful consideration of the multiple state ERM construction rules (a cherry being selected means that a new pair of leaves is to be added to a randomly chose one of its leaves) shows that the generating matrix of this Pólya urn process is:

$$A := \begin{bmatrix}
-2\left(q_1^{12}+q_1^{22}\right) & q_1^{11} & 0 & 0 & q_1^{11} & 2q_1^{11} & q_1^{11} & 0 & 0 & q_1^{11} \\
2q_1^{12} & -\left(2-q_1^{12}\right) & 0 & 0 & q_1^{12} & 2q_1^{12} & q_1^{12} & 0 & 0 & q_1^{12} \\
2q_1^{22} & q_1^{22} & -2 & 0 & q_1^{22} & 2q_1^{22} & q_1^{22} & 0 & 0 & q_1^{22} \\
0 & q_2^{22} & 2q_2^{22} & -2\left(q_2^{11}+q_2^{12}\right) & q_2^{22} & 0 & 0 & q_2^{22} & q_2^{22} & 0 \\
0 & q_2^{12} & 2q_2^{12} & 2q_2^{12} & -\left(2-q_2^{12}\right) & 0 & 0 & q_2^{12} & q_2^{12} & 0 \\
0 & q_2^{11} & 2q_2^{11} & 2q_2^{11} & q_2^{11} & -2 & 0 & q_2^{11} & q_2^{11} & 0 \\
2 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 1 & 2 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & -1
\end{bmatrix}$$

whose eigenvalues can be shown to be: $\lambda_1 = 1$, $\lambda_2 = c_1 - c_2 - 1 = 2q_1^{11} + q_1^{12} - 2q_2^{11} - q_2^{12} - 1$, $\lambda_3 = \lambda_4 = -1$ and $\lambda_5 = \cdots = \lambda_{10} = -2$. The normalized right and left eigenvectors of the largest real eigenvalue can be calculated in terms of the parameters for the probabilities to be

$$
\boldsymbol{v}_1 = \frac{1}{3(2-c_1+c_2)}
\begin{bmatrix}
q_1^{11}c_2 \\
q_1^{12}c_2 \\
q_1^{22}c_2 \\
q_2^{22}(2-c_1) \\
q_2^{12}(2-c_1) \\
q_2^{11}(2-c_1) \\
c_1c_2/2 \\
(2-c_1)c_2/2 \\
(2-c_1)(2-c_2)/2 \\
(2-c_1)c_2/2
\end{bmatrix}, \quad
\boldsymbol{u}_1 =
\begin{bmatrix}
2 \\
2 \\
2 \\
2 \\
2 \\
2 \\
1 \\
1 \\
1 \\
1
\end{bmatrix}
$$

If we assume that $q_1^{11}, q_2^{22} \neq 1$, this excludes the case when the generated ERM tree is of a single type only, the urn process is irreducible, and also $\lambda_2 < 1$ is a simple eigenvalue. As all the assumptions are satisfied, applying the Theorem for Pólya urns we obtain the claimed results.

Remark 8 This agrees with our earlier result from Proposition 3 as the means of the number of cherries obtained earlier in fact satisfy $\mathbb{E}[X_i(n)]/n \to v_{1i}$, for $i = 1, \ldots, 6$ as $n \to \infty$; e.g. compare the strong law result to to the means calculated for finite n in our special cases of multiple state ERM models (except in the case (e) in which type 111 is a sink for the process). In particular, we could have used the framework of the extended Pólya urn scheme to calculate recursive equations for the means and variances of the numbers of different types of cherries in the process—unfortunately, this approach does not yield a simpler proof of results from Propositions 3 and 5.

Central limit law for its (normalized) asymptotic distribution holds as well.

Theorem 9 *Assume* $\{q_i^{j_1 j_2}\}_{i, j_1 \leq j_2 \in \{1,2\}}$ *satisfy* $q_1^{11}, q_2^{22} \neq 1$, *and* $c_1 - c_2 \neq 0$. *Then,*
(i) If $c_1 - c_2 = 3/2$, *as* $n \to \infty$,

$$
\frac{X_n - n\boldsymbol{v}_1}{n\ln(n)} \overset{d}{\Rightarrow} N(0, \boldsymbol{\Sigma}),
$$

with

$$
\boldsymbol{\Sigma} = C
\begin{bmatrix}
(q_1^{11})^2 & q_1^{11}q_1^{12} & q_1^{11}q_1^{22} & -q_1^{11}q_2^{22} & -q_1^{11}q_2^{12} & -q_1^{11}q_2^{11} & * & * & * & * \\
\checkmark & (q_1^{12})^2 & q_1^{12}q_1^{22} & -q_1^{12}q_2^{22} & -q_1^{12}q_2^{12} & -q_1^{12}q_2^{11} & * & * & * & * \\
\checkmark & \checkmark & (q_1^{22})^2 & -q_1^{22}q_2^{22} & -q_1^{22}q_2^{12} & -q_1^{12}q_2^{11} & * & * & * & * \\
\checkmark & \checkmark & \checkmark & (q_2^{22})^2 & q_2^{22}q_2^{12} & q_2^{22}q_2^{11} & * & * & * & * \\
\checkmark & \checkmark & \checkmark & \checkmark & (q_2^{12})^2 & q_2^{12}q_2^{11} & * & * & * & * \\
\checkmark & \checkmark & \checkmark & \checkmark & \checkmark & (q_2^{11})^2 & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & * \\
* & * & * & * & * & * & * & * & * & *
\end{bmatrix}
$$

where the constant C is given by

$$C := -8 \left(9 + 12(q_1^{11})^2 + 2q_2^{11}q_1^{12} + 4q_1^{11}q_2^{11} \right.$$
$$\left. + 4(q_1^{12})^2 + 14q_1^{11}q_1^{12} - 4q_2^{11} - 12q_1^{12} - 21q_1^{11} \right) /25;$$

the explicit expressions for entries marked by $*$ are omitted as they represent the covariances between cherries and the pendants, and pendants between themselves; and expressions for the entries marked by a $\checkmark$ follow from the symmetry of the covariance matrix.

(ii) If $c_1 - c_2 < 3/2$, as $n \to \infty$, then

$$\frac{X_n - n\boldsymbol{v}_1}{\sqrt{n}} \overset{d}{\Rightarrow} N(0, \boldsymbol{\Sigma}').$$

where $\boldsymbol{\Sigma}'$ can be obtained explicitly only in some special cases.

Proof The proof again relies on the corresponding result for our specific Pólya urn described in the proof of Theorem 6: if we assume all the conditions there plus $\mathrm{Re}(\lambda_2) \leq \lambda_1/2$, Theorems 3.22 and 3.23 in the paper by Janson (2004) insure that, as $n \to \infty$:

(i) if $\mathrm{Re}(\lambda_2) = \lambda_1/2$, then

$$\frac{X_n - n\lambda_1\boldsymbol{v}_1}{n\ln(n)} \overset{d}{\Rightarrow} N(0, \boldsymbol{\Sigma}),$$

where the covariance matrix is given by $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{T})\boldsymbol{\Sigma}_{II}(\boldsymbol{I} - \boldsymbol{T}^\mathsf{T})$, with $\boldsymbol{T} := \lambda_2^{-1}\lambda_1\boldsymbol{v}_1\boldsymbol{a}^\mathsf{T}\boldsymbol{v}_2\boldsymbol{u}_2^\mathsf{T}$, $\boldsymbol{\Sigma}_{II} := \boldsymbol{v}_2\boldsymbol{u}_2^\mathsf{T}\boldsymbol{B}(\boldsymbol{v}_2\boldsymbol{u}_2^\mathsf{T})$, and $\boldsymbol{B} := \sum_{i=1}^{l}\boldsymbol{v}_{1i}a_i\mathbb{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T})$;

(ii) if $\mathrm{Re}(\lambda_2) < \lambda_1/2$, then

$$\frac{X_n - n\lambda_1\boldsymbol{v}_1}{\sqrt{n}} \overset{d}{\Rightarrow} N(0, \boldsymbol{\Sigma}').$$

where the covariance matrix is given by $\boldsymbol{\Sigma}' := \int_0^\infty \psi(s, \boldsymbol{A})\boldsymbol{B}\psi(s, \boldsymbol{A})^\mathsf{T}e^{-\lambda_1 s}\lambda_1 ds - \lambda_1^2\boldsymbol{v}_1\boldsymbol{v}_1^\mathsf{T}$, with $\boldsymbol{B}$ as above and $\psi(s, \boldsymbol{A}) := e^{s\boldsymbol{A}} - \lambda_1\boldsymbol{v}_1\boldsymbol{a}^\mathsf{T}\int_0^s e^{t\boldsymbol{A}}dt$.

The two options on the eigenvalues correspond to: (i) $c_1 - c_2 = 3/2$, and (ii) $c_1 - c_2 < 3/2$, respectively. To explicitly calculate the covariance matrix $\boldsymbol{\Sigma}$ in (i) we need to find the normalized right and left eigenvectors corresponding to the second largest eigenvalue λ_2, which are given in term of the parameters for the probabilities as

$$v_2 = \frac{c_2}{(2 - c_1 + c_2)(c_2 - c_1 - 1)} \begin{bmatrix} q_1^{11} \\ q_1^{12} \\ q_1^{22} \\ q_2^{22} \\ q_2^{12} \\ q_2^{11} \\ c_1/(c_1 - c_2) \\ (2 - c_1)/(c_1 - c_2) \\ -(2 - c_1)/(c_1 - c_2) \\ c_2/(c_1 - c_2) \end{bmatrix},$$

$$u_2 = \frac{1}{c_2} \begin{bmatrix} -2(2 - c_1) \\ c_1 + c_2 - 2 \\ 2 \\ 2 \\ c_1 + c_2 - 2 \\ -2(2 - c_1) \\ c_1 - 2 \\ 1 \\ 1 \\ c_1 - 2 \end{bmatrix}.$$

Computing the matrix $\mathbf{B} = (c_1 - c_2 - 2)^{-1}[b_{i,j}]_{1 \le i, j \le 10}$ gives lengthy expression for its entries

$$b_{1,1} = -\frac{q_1^{11}}{3}\left(10q_2^{11} - 8q_1^{11}q_2^{11} + 5q_2^{12} - 4q_1^{11}q_2^{12}\right), \quad b_{1,2} = q_1^{11}\left(2q_2^{11} + q_2^{12}\right)q_1^{12},$$

$$b_{1,3} = -\frac{2q_1^{11}}{3}\left(2q_2^{11} + q_2^{12}\right)\left(-1 + q_1^{11} + q_1^{12}\right),$$

$$b_{1,4} = -\frac{q_1^{11}}{6}\left(2q_2^{11} + q_2^{12}\right)\left(-4 + 2q_1^{11} - q_1^{12}\right),$$

$$b_{1,5} = -\frac{q_1^{11}}{3}\left(2q_2^{11} + q_2^{12}\right)q_1^{12}, \quad b_{16} = 0,$$

$$b_{1,7} = -\frac{q_1^{11}}{3}\left(-2 + 2q_1^{11} + q_1^{12}\right)q_2^{12}, \quad b_{1,8} = -\frac{2q_1^{11}}{3}\left(-2 + 2q_1^{11} + q_1^{12}\right)q_2^{11},$$

$$b_{1,9} = \frac{q_1^{11}}{6}\left(-2 + 2q_1^{11} + q_1^{12}\right)\left(2q_2^{11} - q_2^{12}\right), \quad b_{1,10} = \frac{q_1^{11}}{3}\left(-2 + 2q_1^{11} + q_1^{12}\right)q_2^{12}$$

$$b_{2,1} = q_1^{12}\left(2q_2^{11} + q_2^{12}\right)q_1^{11}, \quad b_{2,2} = -\frac{q_1^{12}}{3}\left(10q_2^{11} + 5q_2^{12} - 4q_1^{12}q_2^{11} - 2q_1^{12}q_2^{12}\right)$$

$$b_{2,3} = -\frac{q_1^{12}}{3}\left(2q_2^{11} + q_2^{12}\right)\left(-1 + q_1^{11} + q_1^{12}\right),$$

$$b_{2,4} = -\frac{q_1^{12}}{3}\left(2q_2^{11} + q_2^{12}\right)\left(2q_1^{11} - 2 - q_1^{12}\right),$$

$$\ldots$$

$$b_{10,7} = -\frac{q_2^{12}}{6}\left(-2 + 2q_1^{11} + q_1^{12}\right)\left(2q_2^{11} + 3q_2^{12}\right),$$

$$b_{10,8} = -\frac{q_2^{11}}{6}\left(-2 + 2q_1^{11} + q_1^{12}\right)\left(-2 + 2q_2^{11} + 3q_2^{12}\right)$$

$$b_{10,9} = 0, \; b_{10,10} = -\frac{1}{2}\left(-2 + 2q_1^{22} + q_1^{12}\right)\left(q_2^{12} - 2 + 2q_2^{11}\right)$$

Further lengthy and cumbersome linear algebra (computed using Maple) provides the given entries for the variances and covariances of different types of cherries in $\boldsymbol{\Sigma}$ as claimed.

Calculating the matrix $\boldsymbol{\Sigma}'$ in (ii) is even more involved, due to its integral expressions, and can not be made to simplify other than in some very special cases. $\qquad \square$

Remark 10 The results above are consistent with our calculations of asymptotics for the variances of the number of cherries in Proposition 5 (cf. Remark 8). When $c_1 - c_2 < 3/2$ implies $c_1 - c_2 - 1 < 1/2$ and $2(c_1 - c_2 - 1) < 1$, and the individual variances are $O(n)$. When $c_1 - c_2 = 3/2$ the additional factor $\ln n$ comes from covariances in numbers of different types of cherries.

The asymptotic strong law allows us to approximate unknown parameters for the probabilities for ERM trees with a large number of leaves using counts of different types of cherries on the tree.

Corollary 11 *If the proportion of different types of cherries in a multiple state ERM tree is given by $\boldsymbol{x}_n = (X_1(n)/n, \ldots, X_6(n)/n)$ and the number of leaves n in the tree is large, one can approximately recover the parameters for the probabilities of the model to be*

$$q_1^{11} = \frac{x_1}{x_1 + x_2 + x_3}, \; q_1^{12} = \frac{x_2}{x_1 + x_2 + x_3}, \; q_1^{22} = \frac{x_3}{x_1 + x_2 + x_3},$$

$$q_2^{22} = \frac{x_4}{x_4 + x_5 + x_6}, \; q_1^{12} = \frac{x_5}{x_4 + x_5 + x_6}, \; q_1^{22} = \frac{x_6}{x_4 + x_5 + x_6},$$

as long as the total number of cherries with branch-point of type 1 and of type 2 are non-zero.

This result is completely intuitive from a law of large numbers perspective: the parameters for the probabilities for having a branch-point of type ij_1j_2 are given by the limiting fraction of cherries of type ij_1j_2. Our results on the variability of the number of cherries allows one to make a more precise statement about the error one is making using such an approximation when the number of leaves is finite. Note that the asymptotic limits are independent of the initial state of the process, which is due to the ergodicity of the underlying Markov chain. Although from a practical point of view this is certainly an advantage: no knowledge of the initial state is needed to recover the parameters in the model, these results do not say anything about the ability to reconstruct the state at the root. We next briefly address this question.

In the standard Markov propagation model on trees the probabilities for the types of two leaves attaching to the same branch-point are independent. These are given by a stochastic transition matrix $\boldsymbol{S} = [s_{ij}]_{i,j\in\{1,\ldots,k\}}$ where s_{ij} is the probability

that a leaf of type j will attach to a type i. In our notation this gives probabilities $q_i^{j_1 j_2} = 2 s_{i j_1} s_{i j_2}$, $j_1 < j_2$ and $q_i^{jj} = s_{ij}^2$. Reconstruction of types for Markov propagation models has been extensively studied (see the survey of Mossel and Steel 2005). We only illustrate how the information on cherries can be used as a proxy to determine whether robust reconstruction is possible or not. Without going into all the details we recall that 'reconstruction problem is solvable' if there exist two different types which when used at the root of the tree propagate asymptotically different distributions (measured by total variation) on the leaves of the tree. This roughly means that the leaf types contain a non-vanishing amount of information on the type of the root of the tree as the number of leaves $n \to \infty$. A key result (Mossel and Steel 2005) then states that on a binary tree the reconstruction problem is solvable when $\lambda_2 > 1/\sqrt{2}$, where λ_2 is the second largest eigenvalue of the propagation matrix S. When $k = 2$ this condition becomes $|s_{11} + s_{22} - 1| > 1/\sqrt{2}$ which, using Corollary 11, is equivalent to

$$\left| \sqrt{v_{1,1}/(v_{1,1} + v_{1,2} + v_{1,3})} + \sqrt{v_{1,4}/(v_{1,4} + v_{1,5} + v_{1,6})} - 1 \right| > 1/\sqrt{2}.$$

3 From discrete to continuous multiple state trees

So far we proposed a model, *multiple state ERM tree*, that extends the discrete ERM tree to a tree with multiple states, and showed that long term statistics on the number of different types of cherries is sufficient to recover rate parameters in the model. This random tree is an appropriate model when the underlying biological process is such that the value of any particular state does not affect the evolutionary rates at which the branchings in the tree happen (e.g. speciation and extinction rates are state independent). In this case the shape of the tree is not affected by the state values along the ancestral process, and can be modelled by multiple state propagation on an ERM tree.

In case speciation and extinction rates are state dependent a more complex model is needed. We next propose a model, *multiple state Yule tree*, for extending the continuous time Yule tree (i.e. a Yule tree with edge lengths) to a pure birth tree with multiple states and with state transitions. This random tree model naturally appears as the ancestral tree of a multi-type birth and death branching process, which we first show to be true by applying a procedure of pruning extinct lineages in the full genealogy of the branching process. As was the case for the discrete multiple state ERM tree, we show that model parameters are recoverable from the long term statistics on the number of cherries of different types.

The continuous time multiple state Yule tree model is an extension of the discrete multiple state ERM tree model: if the former has time independent rates, then ignoring its edge lengths produces the multiple state ERM tree (c.f. Remark 15). Importantly, in both models the counts of different types of cherries represent simple signatures of the correlation structure inherent in the evolution of trees with multiple states. Consequently, the collection of cherry and pendant counts can be used as an effective tool in identifying features of the underlying biological process (c.f. Corollaries 11 and 25). This explains the motivation behind our derivation of asymptotic results for these mathematical objects, and their applicability in inference and model testing for evolutionary processes.

4 Ancestral tree of a multi-type birth-death process

Consider a random tree with edge lengths, constructed from an originating node, using a pure birth process. By rescaling time one can relate any such tree to one whose birth rate is 1, even when the rate is time varying. The distribution of this tree is called 'Yule' tree (first considered in the biological context by Yule 1924), and has been used extensively as a null model in investigating speciation process. This is due to the fact that its distribution is precisely that of the ancestral tree reconstructed from any birth-death branching process with constant rates (Nee et al. 1994)—an *ancestral tree* is obtained from a full tree of the process by pruning away all the branches without any extant species. When the branch lengths of a Yule tree are ignored (given the same length) this produces the uniform distribution on ranked tree shapes (a *ranked tree* is one in which the order of branching events matters) with labelled tips, and when the ranking is also ignored it produces the (single-type) ERM distribution on binary trees (Aldous 1996).

We consider a multiple state version of this tree obtained as the ancestral tree reconstructed from a multi-type birth-death process. Let $\mathbf{Z} = (\mathbf{Z}(t))_{t \geq 0}$ denote a multi-type birth-death process on $\mathcal{K} = \{1, \ldots, k\}$ types (i.e. states), whose coordinates provide the count of different types in the population $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_k(t))$. Let $T > 0$ and let $\mathcal{Z}$ denote the full tree of $(\mathbf{Z}(t))_{0 \leq t \leq T}$. Let $\mathcal{W}$ denote the ancestral tree obtained by pruning away all lineages of $\mathcal{Z}$ which do not have any extant lineages at time T (the law of $\mathcal{W}$ depends on T but for simplicity we omit T from its notation). An illustration of an ancestral tree associated with a multi-type birth-death process is shown in Fig. 3. Let $\mathbf{W} = (\mathbf{W}(t))_{0 \leq t \leq T}$, $\mathbf{W}(t) = (W_1(t), \ldots, W_k(t))$ denote the population size process of the ancestral tree $\mathcal{W}$ (clearly we have $\forall i$, $\forall t \in [0, T]$: $W_i(t) \leq Z_i(t)$ and $W_i(T) = Z_i(T)$). We call $\mathbf{W}$ the *reconstructed ancestral process* of $\mathbf{Z}$ and the first intuitive observation is that the ancestral tree obtained by such pruning is Markovian (detailed proof is in the Appendix). We then derive its law, which turns out to be a multiple state pure birth process with time varying rates and an added ability to switch states along a single lineage.
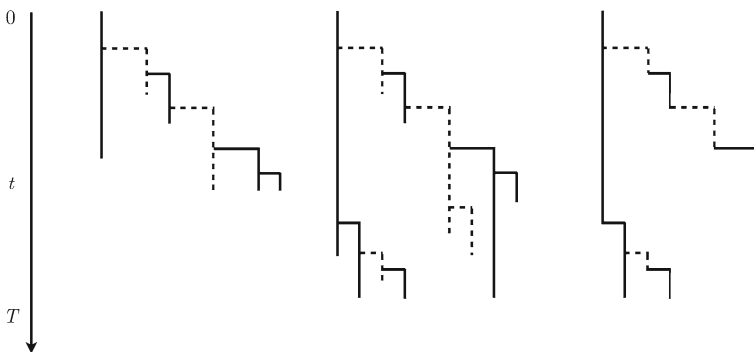

Fig. 3 A tree of a two-type birth-death process, where *dashed* and *solid lines* represent type 1 and type 2 lineages, respectively; the *left* most tree is that of the process observed until some time $t < T$, the *center* tree is of the same process observed until time T, the *right* most tree is the ancestral tree associated with the process surviving to T

Lemma 12 *The reconstructed ancestral process W of Z is a Markov process.*

Proposition 13 *Assume the multi-type birth-death process Z has birth rates $\{b_i^{ij}\}_{i,j \in \{1,\ldots,k\}}$ (b_i^{ij} = rate at which any type i gives birth to a type j) and death rates $\{d_i\}_{i \in \{1,\ldots,k\}}$ (d_i = rate at which any type i dies). Then, for any $T > 0$, the reconstructed ancestral process W is a pure birth process with birth rates $\{q_i^{ij}(t)\}_{i,j \in \{1,\ldots,k\}}$ (q_i^{ij} = rate at which state i gives birth to state j) and state transition rates $\{q_i^j(t)\}_{i \in \{1,\ldots,k\}}$ (q_i^j = rate at which state i changes into state j) at time $t \in [0, T]$, given by*

$$q_i^{ij}(t) = b_i^{ij}\left(1 - p_{e_j(t,T)}^{\mathbf{0}}\right) \ \forall i, j, \quad q_i^j(t)$$

$$= b_i^{ij}\left(1 - p_{e_j(t,T)}^{\mathbf{0}}\right) \frac{p_{e_i(t,T)}^{\mathbf{0}}}{1 - p_{e_i(t,T)}^{\mathbf{0}}} \ \forall i \neq j \tag{4}$$

where $p_{e_i(t,T)}^{\mathbf{0}} = \mathbb{P}[Z(T) = 0 | Z(t) = e_i]$ are the extinction probabilities for Z.

Remark 14 When there is only one state, for example i, this reduces to a pure birth process with time varying birth rate $b_i(1 - p_{e_i}^{\mathbf{0}})$ as previously established (Nee et al. 1994). The extinction probabilities $\{p_{e_i}^{\mathbf{0}}(t, T)\}_{i \in \{1,\ldots,k\}}$ can be shown to satisfy a system of differential equations (Mode 1962; Jones 2011)

$$\frac{dp_{e_i}^{\mathbf{0}}(t, T)}{dt} = d_i - \left(\sum_{j=1}^k b_i^{ij} + d_i\right) p_{e_i}^{\mathbf{0}}(t, T) + \sum_{j=1}^k b_i^{ij} p_{e_i}^{\mathbf{0}}(t, T) p_{e_j}^{\mathbf{0}}(t, T),$$

$$i = 1, \ldots, k.$$

Proof Since the reconstructed ancestral process $(W(t))_{t \geq 0}$ is a Markov process, it suffices to show that its only transitions are changes of the form $\{e_i, i = 1, \ldots, k\}$ and $\{e_j - e_i, i \neq j = 1, \ldots, k\}$ and calculate their rates. The set of possible transition changes for Z, and the fact that $||W(t)||$ (the sum of its coordinates) is non-decreasing, imply the form of changes for W: an addition of e_i occurs iff there is a birth event and both the new lineage and the parent lineage survive to T, an addition of $e_i - e_j$, occurs iff there is a birth event and only the new lineage survives to T (see Fig. 3 for an example).

Considering the possible values of the underlying birth-death process Z for a transition in $(t, t + \Delta t]$, using (12), we get

$$\mathbb{P}[W_{t+\Delta t} = w + e_j \mid W(t) = w]$$

$$= \frac{\sum_{z \geq w} \mathbb{P}[W(t + \Delta t) = w + e_j, W(t) = w, Z(t) = z]}{\sum_{z \geq w} \mathbb{P}[W(t) = w, Z(t) = z]}$$

$$= \frac{\sum_{z \geq w} \mathbb{P}[Z(t) = z] C_{z,w} \sum_{i=1}^k w_i b_i^{ij} \Delta t \left(1 - p^{\mathbf{0}}(t + \Delta t, T)\right)^{w+e_i} p^{\mathbf{0}}(t + \Delta t, T)^{z-w} + o(\Delta t)}{\sum_z \mathbb{P}[Z(t) = z] C_{z,w} (1 - p^{\mathbf{0}}(t, T))^w p^{\mathbf{0}}(t, T)^{z-w}}$$

$$= \sum_{i=1}^k w_i b_i^{ij} (1 - p_{e_j}^{\mathbf{0}}(t, T)) \Delta t + o(\Delta t). \tag{5}$$

where we used $p^{\mathbf{0}}(t, T)^w := \prod_{i=1}^k p_{e_i}^{\mathbf{0}}(t, T)^{w_i}$, $(1 - p^{\mathbf{0}}(t, T))^w := \prod_{i=1}^k (1 - p_{e_i}^{\mathbf{0}}(t, T))^{w_i}$.

Similarly for $i \neq j$

$$\mathbb{P}[W(t + \Delta t) = w + e_j - e_i \mid W(t) = w]$$

$$= \frac{\sum_z \mathbb{P}[Z(t) = z]C_{z,w} w_i b_i^{ij} \Delta t (1 - p^0(t + \Delta t, T))^{w + e_j - e_i} p^0(t + \Delta t, T)^{z - w + e_i} + o(\Delta t)}{\sum_z \mathbb{P}(Z(t) = z)C_{z,w}(1 - p^0(t, T))^w p^0(t, T)^{z - w}}$$

$$= \frac{w_j b_i^{ij}(1 - p_{e_j}^0(t, T)) p_{e_i}^0(t, T)}{1 - p_{e_i}^0(t, T)} \Delta t + o(\Delta t). \tag{6}$$

Transition rates (5) and (6) correspond to those of a pure birth process allowing for state transitions along the lineages as claimed in (4). $\qquad\square$

In continuous time $t \in [0, T)$ nodes of different states have different time varying *weights*, such that at any time the probability of a node of certain state is chosen to be the next node with a branch-point (binary or unary) is proportional to this weight. The weight of a node of state i is $a_i(t) = q_i(t)/\sum_{\ell=1}^k q_\ell(t)$ where

$$q_i(t) = \frac{1}{1 - p_{e_i}^0(t, T)} \left(\sum_{j=1}^k b_i^{ij}(1 - p_{e_j}^0(t, T)) - b_i^{ii} p_{e_i}^0(t, T)(1 - p_{e_i}^0(t, T)) \right)$$

is the overall rate of events for state i. The probabilities of a node of state i having a binary branch-point (with states i and j) versus a unary branch-point (of state $j \neq i$) are

$$p_i^{ij}(t) = \frac{b_i^{ij}(1 - p_{e_j}^0(t, T))(1 - p_{e_i}^0(t, T))}{\sum_{\ell=1}^k b_i^{i\ell}(1 - p_{e_\ell}^0(t, T)) - b_i^{ii} p_{e_i}^0(t, T)(1 - p_{e_i}^0(t, T))} \quad \forall i, j,$$

$$p_i^j(t) = \frac{b_i^{ij}(1 - p_{e_j}^0(t, T)) p_{e_i}^0(t, T)}{\sum_{\ell=1}^k b_i^{i\ell}(1 - p_{e_\ell}^0(t, T)) - b_i^{ii} p_{e_i}^0(t, T)(1 - p_{e_i}^0(t, T))} \quad \forall i \neq j.$$

Contrary to the single state case, it is not possible to rescale time and relate this to a Yule process with constant rates of births and of state transitions, because the rate at which the time needs to be rescaled depends on the state of the node that was involved in the last branching event. This information is dependent on the randomness of the tree and is not simply a deterministic function of time as it is in the single state case. Consequently, ignoring the edge lengths and possibly the ranking of branching events in these trees does not produce any logical model on multiple state discrete trees. Topologically it also results in multiple state discrete trees which are no longer regular binary ones, as in addition to binary branch-points they also have unary branch-points (with the type attached being necessarily different). Figure 4 illustrates obtaining such a discrete tree.

However, as in the single state case, near the present ($t \approx T$) probabilities of extinction $p^0(t, T)$ are approximately zero, and birth and state transition rates in the ancestral tree are approximately constant $q_i^{ij} \approx b_i^{ij}$, $\forall i, j$, and $q_i^j \approx 0$, $\forall i \neq j$. This allows one to infer birth rates of the process using results on constant rate multiple state Yule trees described in the next section (see Corollary 25). Knowing the values of lineage through time plots for different types ($Z(t), 0 \leq t \leq T$) will then allow one to also infer death rates of the process.
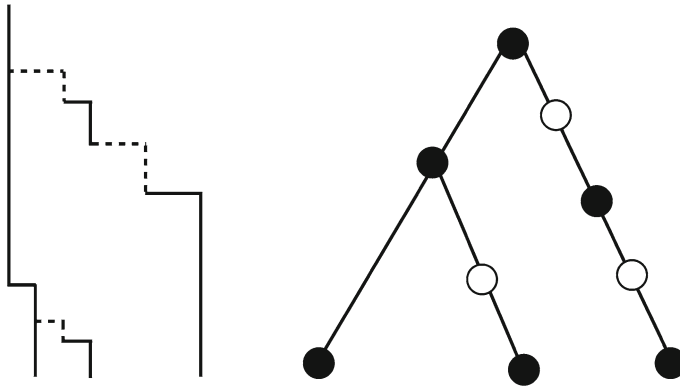
Fig. 4 The ancestral tree from Fig. 3 and the corresponding discrete two-type tree with branch-points and state transitions, obtained by ignoring edge-lengths and ranking in the former, where *dashed* and *solid lines* represent types 1 and 2 ancestors, respectively

Remark 15 If we consider multiple state Yule trees whose birth rates $\{q_i^{ij}\}_{i,j\in\{1,...,k\}}$ and state transition rates $\{q_i^j\}_{i\neq j\in\{1,...,k\}}$ are constant, ignoring edge lengths results in a useful model on multiple state discrete trees: each node of type i is chosen to be the next branch-point with probability proportional to its weight

$$a_i = \frac{q_i}{\sum_{\ell=1}^k q_\ell}, \quad \text{where } q_i = \sum_{j=1}^k q_i^{ij} + \sum_{\substack{j=1 \\ j\neq i}}^k q_i^j;$$

once chosen the branch-point is binary with attached leaves of types i, j, or unary with attached leaf of type $j \neq i$, respectively, with probabilities

$$p_i^{ij} = \frac{q_i^{ij}}{q_i} \ \forall i, j, \quad \text{and} \quad p_i^j = \frac{q_i^j}{q_i} \ \forall i \neq j.$$

The distribution of different types of cherries and pendants in the tree should provide information about its birth and state transition rates. However, the approaches for obtaining their distribution using generating functions and recursive relations (when the number of leaves is finite), as well as the Pólya urn approach for their asymptotic distribution (as the number of leaves grows), are completely unwieldly. The more appropriate approach is to analyze distributions of different types of cherries and pendants in the original continuous time trees as shown in the next Section.

5 Multiple state Yule trees with state transitions

We consider a multiple state birth process with state transitions constructed using time-dependent birth rates $\{q_i^{j_1 j_2}(t)\}_{i,j_1,j_2\in\{1,...,k\}}$ and state transition rates $\{q_i^j(t)\}_{i\neq\in\{1,...,k\}}$, and call its associated tree a *multiple state Yule tree* with state transitions. For generality, we allow for the birth events to result in an instantaneous change

of state for the parent node as well, so that birth rates for a parent node of state i are indexed in the superscript by any $j_1, j_2 \in \{1, \ldots, k\}$ giving a birth event of type ij_1j_2 (rather than only having birth events of types iij as in ancestral trees of the previous section). Consequently, each birth event is a branch-point (with no special designation in the continuing lineages) and in order not to distinguish between different planar embeddings we will w.l.o.g. assume that $j_1 \leq j_2$ (as in the multiple state ERM case). For k states this model has $k^2(k+1)/2 + k(k-1)$ parameters.

Due to state transitions in the model (producing unary branch-points) we need to precise a definition of cherries and pendants in such a tree. Since the sequence of state transition events along a lineage is typically not available in data, we will focus on the states at the topological end-points of the structure. We first let the topology of the tree be defined only by binary branch-points, while unary branch-points are ignored. The cherries and pendants are then defined in this topology as they would be in a regular binary tree. This means that the type of each cherry and each pendant is defined by the state values at the end nodes of the cherry or pendant. respectively. Figure 4 illustrates a two-state Yule tree with state transitions which has only one cherry of type 222 and only one pendant of type 22. In general there are $k^2(k+1)/2$ different types of cherries (we don't differentiate between different planar embeddings of a cherry type), k^2 different types of pendants (sequence of state transitions along a lineage can revert to the original state), and k different leaf states.

5.1 Moments of the number of different types of cherries and pendants

For a multiple state Yule tree with state transitions, we let $N_1(t), \ldots, N_k(t)$ denote the number of leaves of states $1, \ldots, k$, respectively, at time t. Let $C_i^{j_2 j_2}(t)$ denote the number of cherries of type ij_1j_2, and $L_i^j(t)$ the number of pendants of type ij at time t. We next consider their means, which are relatively straightforward, although quite complicated, to calculate.

Lemma 16 *Let* $\boldsymbol{v}(t) = (v_1(t), \ldots, v_k(t))$ *be the vector of leaf means,* $v_i(t) := \mathbb{E}[N_i(t)], \forall i.$ *Then,* $\forall t \geq 0$

$$\frac{d\boldsymbol{v}(t)}{dt} = \boldsymbol{B}(t)\boldsymbol{v}(t)$$

where $\boldsymbol{B}(t)$ *is the* $k \times k$ *matrix with entries*

$$[\boldsymbol{B}(t)]_{\ell_1, \ell_2} = \begin{cases} q_{\ell_1}^{\ell_1 \ell_1}(t) - \displaystyle\sum_{\substack{i \leq j \\ i, j \neq \ell_1}} q_{\ell_1}^{ij}(t) - \displaystyle\sum_{i \neq \ell_1} q_{\ell_1}^{i}(t) & \text{when } \ell_1 = \ell_2 \\ 2q_{\ell_2}^{\ell_1 \ell_1}(t) + q_{\ell_2}^{\ell_1}(t) + \displaystyle\sum_{j < \ell_1} q_{\ell_2}^{j\ell_1}(t) + \displaystyle\sum_{j > \ell_1} q_{\ell_2}^{\ell_1 j}(t) & \text{when } \ell_1 \neq \ell_2. \end{cases}$$

Proof The matrix formulation is equivalent to the claim that each $v_\ell(t)$ for $1 \leq \ell \leq k$ satisfies

$$\frac{dv_\ell(t)}{dt} = \sum_{i \neq \ell} \left(2q_i^{\ell\ell}(t) + q_i^\ell(t) + \sum_{j<\ell} q_i^{j\ell}(t) + \sum_{j>\ell} q_i^{\ell j}(t) \right) v_i(t)$$

$$+ \left(q_\ell^{\ell\ell}(t) - \sum_{\substack{i \leq j \\ i,j \neq \ell}} q_\ell^{ij}(t) - \sum_{i \neq \ell} q_\ell^i(t) \right) v^\ell(t).$$

To see why this is true, observe that in the time interval $(t, t + \Delta t)$ the number of leaves of state ℓ increases by 2 iff we have a birth event of type $i\ell\ell$ for some $i \neq \ell$. It increases by 1 iff we have a birth event of type $ij\ell$ for some $i, j \neq \ell$ or for $i = j = \ell$, or if we apply a change of type $i\ell$ for some $i \neq \ell$. The number of leaves of state ℓ decreases by 1 only by having birth events of types ℓij or by having state changes ℓi for $i, j \neq \ell$.

Remark 17 If $\boldsymbol{B}(t)$ in Lemma 16 is such that it commutes with $\int_0^t \boldsymbol{B}(\tau)dt\tau \ \forall t \geq 0$, then the vector of leaf means can be given explicitly as

$$\boldsymbol{v}(t) = \int_0^t \exp\{\boldsymbol{B}(\tau)\}d\tau \ \boldsymbol{v}(0), \quad \text{moreover } \boldsymbol{v}(t) = \exp\{\boldsymbol{B}t\} \ \boldsymbol{v}(0)$$

if $\boldsymbol{B}(t)$ is a constant (time-independent) matrix $\boldsymbol{B}$.

Let $\rho(t) := \sum_{i=1}^k v_i(t)$. By adding up counts for all different leaves we obtain the following.

Corollary 18 *Assume $\boldsymbol{B}(t)$ commutes with $\int_0^t \boldsymbol{B}(\tau)d\tau \ \forall t \geq 0$, then*

$$\rho(t) = \boldsymbol{1}^\mathsf{T} \int_0^t \exp\{\boldsymbol{B}(\tau)\}d\tau \ \boldsymbol{v}(0), \quad \text{and} \quad \rho(t)$$

$$= \boldsymbol{1}^\mathsf{T} \exp\{\boldsymbol{B}t\}\boldsymbol{v}(0) \ \text{if } \boldsymbol{B}(t) \equiv \boldsymbol{B} \ \forall t \geq 0.$$

We next give the mean number of cherries whose branch-point is of state ℓ. The mean number of cherries with branch-points of other states can be obtained analogously.

Proposition 19 *Let $\boldsymbol{\mu}_\ell(t) = (\mu_\ell^{11}(t), \ldots, \mu_\ell^{kk}(t))$ be the vector of cherry means $\mu_\ell^{ij}(t) := \mathbb{E}[C_\ell^{11}(t)]$ of types ℓij, for $i \leq j \in \{1, \ldots, k\}$. Then, $\forall t \geq 0$*

$$\frac{d\boldsymbol{\mu}_\ell(t)}{dt} = \boldsymbol{A}_\ell(t)\boldsymbol{\mu}_\ell(t) + \boldsymbol{q}_{(\ell)}(t)v_\ell(t),$$

where

$$\boldsymbol{q}_{(\ell)}(t) := [q_\ell^{11}(t), q_\ell^{12}(t), \ldots, q_\ell^{kk}(t)]^\mathsf{T}$$

and $\boldsymbol{A}_\ell(t)$ is a $\binom{k+1}{2} \times \binom{k+1}{2}$ matrix with entries

$[\boldsymbol{A}_\ell(t)]_{\ell ij, \ell mn}$

$$= \begin{cases} -(q_i(t) + q_j(t)), & \text{when } (m, n) = (i, j) \\ \delta_{m,i} q_n^i(t) + \delta_{n,i} q_m^i(t), & \text{when } (m, n) \neq (i, j), \ i = j \\ \delta_{m,i} q_n^j(t) + \delta_{m,j} q_n^i(t) + \delta_{n,i} q_m^j(t) + \delta_{n,j} q_m^i(t), & \text{when } (m, n) \neq (i, j), \ i \neq j. \end{cases}$$

where, for $i \in \{1, \ldots, k\}$, $q_i(t)$ is the overall rate of events occurring to a lineage of state i

$$q_i(t) := \sum_{j \leq \ell} q_i^{j\ell}(t) + \sum_{j \neq i} q_i^j(t),$$

and where entries in the matrix $\boldsymbol{A}_\ell(t)$ are ordered in a consistent way with that of types in the vectors $\boldsymbol{\mu}_\ell(t)$, $\boldsymbol{q}_{(\ell)}(t)$.

Proof We show that each $\mu_\ell^{ij}(t)$ satisfies the following differential equation when $i = j$

$$\frac{d\mu_\ell^{ij}(t)}{dt} = \sum_{\substack{m \leq n \\ (m,n) \neq (i,i)}} \left(\delta_{m,i} q_n^i(t) + \delta_{n,i} q_m^i(t) \right) \mu_\ell^{mn} - 2q_i(t)\mu_\ell^{ii}(t) + q_\ell^{ii}(t)v_\ell(t),$$

and when $i \neq j$ it satisfies

$$\frac{d\mu_\ell^{ij}(t)}{dt} = \sum_{\substack{m \leq n \\ (m,n) \neq (i,j)}} \left(\delta_{m,i} q_n^j(t) + \delta_{m,j} q_n^i(t) + \delta_{n,i} q_m^j(t) + \delta_{n,j} q_m^i(t) \right) \mu_\ell^{mn}(t)$$

$$- (q_i(t) + q_j(t))\mu_\ell^{ij}(t) + q_\ell^{ij}(t)v_\ell(t).$$

This can be seen from the fact that the number of cherries of type ℓij will increase by 1 iff a cherry of type ℓij is added by a birth event to a lineage of state ℓ, or there is a state transition along a lineage of a cherry which from a cherry of some different type produces a cherry of type ℓij. The number of cherries of type ℓij will decrease by 1 iff there is a state transition along a lineage of a type ℓij cherry, or there is a birth event along one of its lineages producing a cherry of some different type.

Remark 20 The matrix $\boldsymbol{A}_\ell(t)$ is diagonally dominant by columns, as: in a column ℓmn every rate of the form $q_m^i(t)$ and every rate of the form $q_n^i(t)$ appears exactly once (when $n = m$ each one appears twice) and the sum of these rates is less than or equal to $q_n(t) + q_m(t)$. We will use this fact in upcoming proofs.

Proposition 21 *Let $\boldsymbol{\gamma}(t) = (\gamma_1^1(t), \ldots, \gamma_k^k(t))$ be the vector of pendant means $\gamma_i^j(t) := \mathbb{E}[L_i^j(t)]$. Then, $\forall t \geq 0$*

$$\frac{d\boldsymbol{\gamma}(t)}{dt} = \boldsymbol{C}(t)\boldsymbol{\gamma}(t) + \boldsymbol{U}(t)\boldsymbol{\mu}(t),$$

where $C(t)$ is a $k^2 \times k^2$ matrix with entries

$$[C(t)]_{\ell m, ij} = \begin{cases} -q_m(t) & \text{when } (\ell, m) = (i, j) \\ q_j^m(t) & \text{when } \ell = i, \ m \neq j \\ 0 & \text{otherwise.} \end{cases}$$

and $U(t)$ is a $k^2 \times \binom{k+1}{2}$ matrix with entries

$$[U(t)]_{\ell m, \ell' i j} = \begin{cases} 2 \displaystyle\sum_{j_1 \leq j_2} q_m^{j_1 j_2}(t) & \text{when } \ell = \ell', \ m = i = j \\ \displaystyle\sum_{j_1 \leq j_2} q_i^{j_1 j_2}(t) & \text{when } \ell = \ell', \ m = j > i \\ \displaystyle\sum_{j_1 \leq j_2} q_j^{j_1 j_2}(t) & \text{when } \ell = \ell', \ m = i < j \\ 0 & \text{otherwise.} \end{cases}$$

Proof We show that each $\gamma_\ell^m(t)$ satisfies

$$\frac{d\gamma_\ell^m(t)}{dt} = \sum_{j \neq m} q_j^m(t)\gamma_\ell^j(t) - q_m(t)\gamma_\ell^m(t) + \sum_{i < m}\left(\sum_{j_1 \leq j_2} q_i^{j_1 j_2}(t)\right)\mu_\ell^{im}(t)$$

$$+ \sum_{i > m}\left(\sum_{j_1 \leq j_2} q_i^{j_1 j_2}(t)\right)\mu_\ell^{mi}(t) + 2\sum_{j_1 \leq j_2} q_m^{j_1 j_2}(t)\mu_\ell^{mm}(t).$$

To see this, observe that the number of pendant edges of type ℓm will increase by 1 if a state transition of type jm occurs on a pendant edge of type ℓj. Also, it will increase by 1 if a birth event happens adding any cherry of type $ij_1 j_2$ to any cherry of type ℓim ($i \neq m$), or if a cherry of type $mj_1 j_2$ is added to a cherry of type ℓmm. The number of pendant edges of type ℓm will decrease by 1 iff any birth event adding a cherry or state transition occurs on a pendant edge of type ℓm.

5.2 Long time asymptotics for the number of cherries and pendants

We next consider what happens to the tree structure of the multiple state Yule process with state transitions as $t \to \infty$. The random total number of leaves $\sum_{i=1}^k N_i(t)$ grows as well, so we need to consider the fraction of different types of cherries and pendants. We start with results in case the birth $\{q_i^{j_1 j_2}\}_{i, j_1 \leq j_2 \in \{1,...,k\}}$ and state transition $\{q_i^j\}_{i \neq j \in \{1,...,k\}}$ rates in the process are constant (time independent) and then generalize to the time varying case.

The matrix $B(t)$ from Lemma 16 has nonnegative entries, except possibly for those on the diagonal. By the Perron-Frobenius theorem, if it is irreducible, there exists a largest simple eigenvalue $\lambda(t)$ of $B(t)$ with right and left eigenvectors $u(t)$, $v(t)$,

respectively. We can assume that $\mathbf{1} \cdot \boldsymbol{u}(t) = 1$. In case that $\boldsymbol{B}(t) \equiv \boldsymbol{B}$ is time independent, we have the following result. Recall $\rho(t)$ is the mean total number of leaves in the tree.

Lemma 22 *If $\boldsymbol{B}(t) \equiv \boldsymbol{B}$ and is irreducible, then $\boldsymbol{\eta}_\ell(t) := \rho(t)^{-1}\boldsymbol{\mu}_\ell(t)$, if it converges, satisfies*

$$\lim_{t \to \infty} \boldsymbol{\eta}_\ell(t) = -u_\ell(\boldsymbol{A}_\ell - \lambda \boldsymbol{I})^{-1}\boldsymbol{q}_{(\ell)},$$

where λ is the largest real eigenvalue of $\boldsymbol{B}$ with corresponding right eigenvector $\boldsymbol{u} = (u_1, \ldots, u_k)$. Furthermore, if $\boldsymbol{\eta}(t) = (\boldsymbol{\eta}_1(t), \ldots, \boldsymbol{\eta}_k(t))$, then $\boldsymbol{\eta}^\star(t) := \rho(t)^{-1}\boldsymbol{\gamma}(t)$, if it converges, satisfies

$$\lim_{t \to \infty} \boldsymbol{\eta}^\star(t) = -(\boldsymbol{C} - \lambda \boldsymbol{I})^{-1}\boldsymbol{U} \lim_{t \to \infty} \boldsymbol{\eta}(t).$$

Proof Using $\boldsymbol{\mu}_\ell(t) = \rho(t)\boldsymbol{\eta}_\ell(t)$ in the differential equation for $\mu_\ell(t)$ from Proposition 19 we get

$$\frac{d\boldsymbol{\eta}_\ell(t)}{dt} = \left(\boldsymbol{A}_\ell - \rho(t)^{-1}\frac{d\rho(t)}{dt}\boldsymbol{I}\right)\boldsymbol{\eta}_\ell(t) + \rho(t)^{-1}\boldsymbol{q}_{(\ell)}(t)v_\ell(t).$$

Assuming that $\lim_{t \to \infty} \boldsymbol{\eta}_\ell(t)$ exists, taking limit as $t \to \infty$ on both sides and using the fact that $\boldsymbol{\eta}_\ell(t)$ is continuous, we get

$$0 = \left(\boldsymbol{A}_\ell - \lim_{t \to \infty}\rho(t)^{-1}\frac{d\rho(t)}{dt}\boldsymbol{I}\right)\lim_{t \to \infty}\boldsymbol{\eta}_\ell(t) + \boldsymbol{q}_{(\ell)}\lim_{t \to \infty}\rho(t)^{-1}v_\ell(t). \qquad (7)$$

Let $\boldsymbol{J}$ denote the Jordan representation form of the matrix $\boldsymbol{B}$, so that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}$ and $\exp\{\boldsymbol{B}\} = \boldsymbol{P}\exp\{\boldsymbol{J}\}\boldsymbol{P}^{-1}$. By Corollary 18,

$$\begin{aligned}
\lim_{t \to \infty}\rho(t)^{-1}\frac{d\rho(t)}{dt} &= \lim_{t \to \infty}\frac{\mathbf{1}^\mathsf{T}\boldsymbol{B}\exp\{\boldsymbol{B}t\}\boldsymbol{v}(0)}{\mathbf{1}^\mathsf{T}\exp\{\boldsymbol{B}t\}\boldsymbol{v}(0)} \\
&= \frac{\mathbf{1}^\mathsf{T}\lambda\boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{v}(0)}{\mathbf{1}^\mathsf{T}\boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{v}(0)} = \lambda.
\end{aligned}$$

Similarly, using Lemma 16,

$$\lim_{t \to \infty}\rho(t)^{-1}v_\ell(t) = \frac{e_\ell\boldsymbol{u}\boldsymbol{v}^\mathsf{T}e_a}{\mathbf{1}^\mathsf{T}\boldsymbol{u}\boldsymbol{v}^\mathsf{T}e_a} = u_\ell.$$

We claim that $(\boldsymbol{A}_\ell - \lambda\boldsymbol{I})$ is invertible. This is true because $\boldsymbol{A}_\ell$ is diagonally dominant by columns (see Remark 20), and $\lambda \geq 0$ ($\rho(t)$ is positive and increasing), which means that $(\boldsymbol{A}_\ell - \lambda\boldsymbol{I})$ is diagonally dominant by columns as well. Hence, from (7) we get

$$\lim_{t \to \infty}\boldsymbol{\eta}_\ell(t) = -u_\ell(\boldsymbol{A}_\ell - \lambda\boldsymbol{I})^{-1}\boldsymbol{q}_{(\ell)}.$$

The proof for $\lim_{t \to \infty} \boldsymbol{\eta}^\star(t)$ follows in the analogous steps, using $\boldsymbol{\gamma}(t) = \rho(t)\boldsymbol{\eta}^\star(t)$, the differential equation for $\boldsymbol{\gamma}(t)$ from Proposition 21 and the fact that $\boldsymbol{C} - \lambda \boldsymbol{I}$ is also diagonally dominant.

To obtain a version of this result in the time varying case we need to make some assumptions on the behaviour of birth and state transition rates in the long term limit.

Theorem 23 *If all the birth rates and state transition rates in the long term converge to limits $\{\lim_{t \to \infty} q_i^{j_1 j_2}(t)\}_{i, j_1 \le j_2 \in \{1,...,k\}}$ and $\{\lim_{t \to \infty} q_i^j(t)\}_{i \ne j \in \{1,...,k\}}$ such that the matrix $\lim_{t \to \infty} \boldsymbol{B}(t)$ is irreducible with maximum eigenvalue λ and corresponding right and left eigenvectors $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively; then, assuming the limits below exist,*

$$\boldsymbol{w}_\ell := \lim_{t \to \infty} \boldsymbol{\eta}_\ell(t) = -u_\ell \lim_{t \to \infty} (\boldsymbol{A}_\ell(t) - \lambda \boldsymbol{I})^{-1} \lim_{t \to \infty} \boldsymbol{q}_{(\ell)}(t),$$

and

$$\boldsymbol{w}^\star := \lim_{t \to \infty} \boldsymbol{\eta}^\star(t) = - \lim_{t \to \infty} (\boldsymbol{C}(t) - \lambda \boldsymbol{I})^{-1} \boldsymbol{U}(t) \lim_{t \to \infty} \boldsymbol{\eta}(t),$$

Proof The proof is similar to that in the constant rate case. Replacing $\boldsymbol{\mu}_\ell(t) = \rho(t)\boldsymbol{\eta}_\ell(t)$ in the differential equation for $\mu_\ell(t)$ from Proposition 19, we get

$$\frac{d\boldsymbol{\eta}_\ell(t)}{dt} = \left(\boldsymbol{A}_\ell(t) - \rho(t)^{-1}\frac{d\rho(t)}{dt}\boldsymbol{I}\right)\boldsymbol{\eta}_\ell(t) + \rho(t)^{-1}\boldsymbol{q}_{(\ell)}(t)v_\ell(t)$$

Since $\lim_{t \to \infty} \boldsymbol{\eta}_\ell(t)$ exists, taking $t \to \infty$ on both sides, we get

$$0 = \left(\lim_{t \to \infty} \boldsymbol{A}_\ell(t) - \lim_{t \to \infty} \rho(t)^{-1}\frac{d\rho(t)}{dt}\boldsymbol{I}\right) \lim_{t \to \infty} \boldsymbol{\eta}_\ell(t)$$
$$+ \lim_{t \to \infty} \boldsymbol{q}_{(\ell)}(t) \lim_{t \to \infty} \rho(t)^{-1}v_\ell(t) \tag{8}$$

From Lemma 16 we have $\frac{d\boldsymbol{v}(t)}{dt} = \boldsymbol{B}(t)\boldsymbol{v}(t)$, and defining $\boldsymbol{\beta}(t) := \rho(t)^{-1}\boldsymbol{v}(t)$, we have

$$\frac{d\boldsymbol{\beta}(t)}{dt} = \boldsymbol{B}(t)\boldsymbol{\beta}(t) - \rho(t)^{-1}\frac{d\rho(t)}{dt}\boldsymbol{\beta}(t),$$

which taking $t \to \infty$ on both sides gives

$$\lim_{t \to \infty} \rho(t)^{-1}\frac{d\rho(t)}{dt} \lim_{t \to \infty} \boldsymbol{\beta}(t) = \lim_{t \to \infty} \boldsymbol{B}(t) \lim_{t \to \infty} \boldsymbol{\beta}(t).$$

By assumption the matrix $\lim_{t \to \infty} \boldsymbol{B}(t)$ has all finite entries and is irreducible, hence the vector $\lim_{t \to \infty} \boldsymbol{\beta}(t)$ only has positive entries and the Perron-Frobenius Theorem implies that this vector is the eigenvector $\boldsymbol{u}$ and that $\lambda = \lim_{t \to \infty} \rho(t)^{-1}\frac{d\rho(t)}{dt}$.

We now claim that $(\boldsymbol{A}_\ell(t) - \lambda \boldsymbol{I})$ is invertible. This is true because $\boldsymbol{A}_\ell(t)$ is diagonally dominant by columns (again see Remark 20), and $\lambda \ge 0$ (since $\rho(t)$ is positive and

increasing), which means that $(A_\ell(t) - \lambda I)$ is diagonally dominant by columns as well. Hence, (8) implies

$$\lim_{t \to \infty} \boldsymbol{\eta}_\ell(t) = -u_\ell \lim_{t \to \infty} (A_\ell(t) - \lambda I)^{-1} \lim_{t \to \infty} \boldsymbol{q}_{(\ell)}(t)$$

as claimed.

The proof for $\lim_{t \to \infty} \boldsymbol{\eta}^\star(t)$ follows in the analogous steps, replacing $\boldsymbol{\gamma}(t) = \rho(t)\boldsymbol{\eta}^\star(t)$ in the differential equation for $\boldsymbol{\gamma}(t)$ from Proposition 21 and using the fact that $C - \lambda I$ is also diagonally dominant.

Remark 24 In the special case that $\forall t \geq 0$ the matrices $\boldsymbol{B}(t)$ are irreducible, mutually diagonalizable, to matrices $\boldsymbol{D}(t)$, and have the same right and left eigenvectors $\boldsymbol{u}, \boldsymbol{v}$ for their corresponding maximum eigenvalues $\lambda(t)$, we can give a shorter proof: from Corollary 18;

$$
\begin{aligned}
\lim_{t \to \infty} \frac{\frac{d\rho(t)}{dt}}{\rho(t)} &= \lim_{t \to \infty} \frac{\mathbf{1}^\mathsf{T} \boldsymbol{B}(t) \exp\{\int_0^t \boldsymbol{B}(\tau)d\tau\}\boldsymbol{v}(0)}{\mathbf{1}^\mathsf{T} \exp\{\int_0^t \boldsymbol{B}(\tau)d\tau\}\boldsymbol{v}(0)} \\
&= \lim_{t \to \infty} \frac{\mathbf{1}^\mathsf{T} \boldsymbol{P} \boldsymbol{D}(t) \boldsymbol{P}^{-1} \boldsymbol{P} \exp\{\int_0^t \boldsymbol{D}(\tau)d\tau\}\boldsymbol{P}^{-1}\boldsymbol{v}(0)}{\mathbf{1}^\mathsf{T} \boldsymbol{P} \exp\{\int_0^t \boldsymbol{D}(\tau)d\tau\}\boldsymbol{P}^{-1}\boldsymbol{v}(0)} \\
&= \lim_{t \to \infty} \frac{\mathbf{1}^\mathsf{T} \boldsymbol{P} \boldsymbol{D}(t) \exp\{\int_0^t \boldsymbol{D}(\tau)d\tau\}\boldsymbol{P}^{-1}\boldsymbol{v}(0)}{\mathbf{1}^\mathsf{T} \boldsymbol{P} \exp\{\int_0^t \boldsymbol{D}(\tau)d\tau\}\boldsymbol{P}^{-1}\boldsymbol{v}(0)} \\
&= \lim_{t \to \infty} \lambda(t) \frac{\mathbf{1}^\mathsf{T} \boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{v}(0)}{\mathbf{1}^\mathsf{T} \boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{v}(0)} = \lim_{t \to \infty} \lambda(t),
\end{aligned}
$$

since the dominating terms are only those involving $e^{\int_0^t \lambda(\tau)d\tau}$ with $\boldsymbol{u}, \boldsymbol{v}$ as right and left eigenvectors of $\boldsymbol{B}(t)$ respectively; also,

$$\lim_{t \to \infty} \frac{v_\ell(t)}{\rho(t)} = \frac{\boldsymbol{e}_\ell \boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{e}_a}{\mathbf{1}^\mathsf{T} \boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{e}_a} = u_\ell,$$

and substituting these in (8) gives the desired result.

The asymptotic results allow one to infer the birth and state transition rate parameters of the models based on the number of cherries and pendants. Note that in the constant rate case, we have $k^2(k+1)/2 + k(k-1)$ parameters, and we have $k^2(k+1)/2 + k^2$ statistics which satisfy the relation: $2\sum_{\ell, i \leq j} \eta_\ell^{ij}(t) + \sum_{i \neq j} \eta_i^j(t) = 1$. One nonetheless needs some form of additional information in order to infer the model parameters, as in the following result. Let

$$r_i(t) := \sum_{j_1 \leq j_2} q_i^{j_1, j_2}(t), \quad i \in \{1, \dots, k\}$$

denote the overall birth rates for each state.

Corollary 25 *If the long term birth rates $\lim_{t\to\infty} r_i(t)$ and the maximum real eigenvalue λ of $\lim_{t\to\infty} B(t)$ are known, then the limits of the birth and state transition rates can be expressed in terms of the limiting fractions of cherries and pendants $w = \lim_{t\to\infty} \eta(t)$ and $w^\star = \lim_{t\to\infty} \eta^\star(t)$, where $w_\ell = [w_\ell^{11}, \dots w_\ell^{kk}]^T$, $\forall \ell \in \{1, \dots, k\}$ and $w^\star = [w_1^1, \dots, w_k^k]$.*

Proof Observe that we can express C and U in terms of $r_i(t)$ as

$$[C(t)]_{\ell m, ij} = \begin{cases} -r_m(t) - \sum_{i \neq m} q_m^i(t) & \text{when } (\ell, m) = (i, j). \\[2ex] q_j^m(t) & \text{when } \ell = i, \ m \neq j. \\[2ex] 0 & \text{otherwise.} \end{cases}$$

and,

$$[U(t)]_{\ell m, \ell' ij} = \begin{cases} 2r_m(t) & \text{when } \ell = \ell', \ m = i = j. \\[2ex] r_i(t) & \text{when } \ell = \ell', \ m = j > i. \\[2ex] r_j(t) & \text{when } \ell = \ell', \ m = i < j. \\[2ex] 0 & \text{otherwise.} \end{cases}$$

Since λ is known, Theorem 23 implies we have

$$\lim_{t\to\infty} (C(t) - \lambda I) w^\star + \lim_{t\to\infty} U(t) w = 0$$

a linear system which, knowing the values of $\lim_{t\to\infty} r_i(t)$ and λ, and given the values of w and $w^\star$ from statistics of cherries and pendants, depends only on the limits of the state transition rates $\lim_{t\to\infty} q_i^j(t)$, $i \neq j \in \{1, \dots, k\}$.

For each solution of this system in terms of the limiting state transition rates, we will have the values of $\lim_{t\to} q_{(\ell)}(t)$ for $\ell \in \{1, \dots, k\}$ which can subsequently be used in each of the systems

$$\lim_{t\to\infty} (A_\ell(t) - \lambda I) w_\ell + u_\ell \lim_{t\to\infty} q_{(\ell)}(t) = 0, \quad \forall \ell \in \{1, \dots, k\},$$

which is in fact a linear system in the branching rates because

$$u_\ell = 2 \sum_i w_i^{\ell\ell} + \sum_{i,j<\ell} w_i^{j\ell} + \sum_{i,j>\ell} w_i^{\ell j} + \sum_{i\neq\ell} w_i^\ell.$$

It is therefore possible to get solutions of this system in terms of the limiting birth rates by expressing in terms of vectors w, w^* as claimed. □

Remark 26 In the special case that the overall birth rates $r_i(t) \equiv r_i \; \forall t$ are constants and $r_i \equiv r \; \forall i \in \{1, \ldots, k\}$ are independent of state, the maximum eigenvalue of $\boldsymbol{B}$ is simply $\lambda = r$, so in order to infer the birth and state transition rates we only need to know the overall growth rate r and the statistics on the fractions of cherries and pendants.

5.3 Some simple special cases of multiple state Yule models

To illustrate how the asymptotic fractions of cherries and pendants can be used to infer the birth $\{q_i^{j_1 j_2}\}_{i, j_1 \le j_2 \in \{1, \ldots, k\}}$ and state transition $\{q_i^j\}i \ne j \in \{1, \ldots, k\}$ rates in the model we consider two particular cases of the *'symmetric change of state'* models with $k = 2$. We will assume that the two overall birth rates $r_i = \sum_{1 \le j_1 \le j_2 \le 2} q_i^{j_1 j_2}$ are independent of the state $r_1 = r_2 =: r$, that birth rates are symmetric in parent state $\{q_1^{11} = q_2^{22}, q_1^{12} = q_2^{12}\}$, and that the same holds for state transition rates $\{q_1^2 = q_2^1\}$. We consider the following two such models:

(a) *'symmetric cladogenetic transitions'* model in which change in state can only occur at birth events and occurs independently for the offspring and parent:

$$q_1^2 = q_2^1 = 0, \quad q_1^{11} = q_2^{22} = r(1-p)^2, \quad q_1^{12} = q_2^{12} = r \, 2p(1-p),$$
$$q_1^{22} = q_2^{11} = rp^2$$

where $p \in (0, 1)$ is the probability of state change at a birth event;

(b) *'symmetric anagenetic transitions'* model in which change in state can only occur along the lineage:

$$q_1^{11} = q_2^{22} = r, \; q_1^{12} = q_1^{22} = q_2^{12} = q_2^{11} = 0, \quad q_1^2 = q_2^1 = rp$$

where p is the relative rate of state transition along a lineage.

(a) Since in the cladogenetic case all state transition rates are zero, by Corollary 25 we only need to solve the system of equations $\{(\boldsymbol{A}_\ell - \lambda \boldsymbol{I})\boldsymbol{w}_\ell + u_\ell \boldsymbol{q}_{(\ell)} = \boldsymbol{0}\}_{\ell=1,2}$ for the rates $\boldsymbol{q}_{(1)}, \boldsymbol{q}_{(2)}$. We have that the matrix $\boldsymbol{B}$ is

$$\boldsymbol{B} = \begin{bmatrix} q_1^{11} - q_1^{22} - q_1^2 & 2q_2^{11} + q_2^{12} + q_2^1 \\ 2q_1^{22} + q_1^{12} + q_1^2 & q_2^{22} - q_2^{11} - q_2^1 \end{bmatrix} = r \begin{bmatrix} 1 - 2p & 2p \\ 2p & 1 - 2p \end{bmatrix},$$

with eigenvalue $\lambda = r$ and corresponding right eigenvector $u = [1/2, 1/2]^{\mathrm{T}}$. We have $q_1 = q_2 = r$ and the matrices $\boldsymbol{A}_1, \boldsymbol{A}_2$ are

$$\boldsymbol{A}_1 = \boldsymbol{A}_2 = \begin{bmatrix} -(q_1 + q_1) & 0 & 0 \\ 0 & -(q_1 + q_2) & 0 \\ 0 & 0 & -(q_2 + q_2) \end{bmatrix} = \begin{bmatrix} -2r & 0 & 0 \\ 0 & -2r & 0 \\ 0 & 0 & -2r \end{bmatrix}.$$

Solving the above system of equations for $q_{(1)}, q_{(2)}$ in terms of the asymptotic fractions of cherries and pendants $\boldsymbol{w}_\ell = [w_\ell^{11}, w_\ell^{12}, w_\ell^{22}]$ for $\ell \in \{1, 2\}$ and $\boldsymbol{w}^\star =$

$[w_1^1, w_1^2, w_2^1, w_2^2]$ gives $q_{(\ell)} = -2(A_\ell - rI)w_\ell = -3A_\ell w_\ell = 6r[w_\ell^{11}, w_\ell^{12}, w_\ell^{22}]^T$, for $\ell = 1, 2$. Birth rates then are

$$q_\ell^{j_1 j_2} = 6r w_\ell^{j_1 j_2} \quad \text{for } \ell, j_1 \le j_2 \in \{1, 2\}$$

This implies that the asymptotic fractions of cherries together with p satisfy

$$p = 1 - \sqrt{6w_1^{11}} = \sqrt{6w_1^{22}} = \frac{1}{2}\left(1 \pm \sqrt{1 - 12w_1^{12}}\right).$$

Note that if we ignore edge lengths in this tree, we essentially get the random discrete tree arising from the symmetric Markov propagation model, briefly discussed at the end of Sect. 2.3, in which the propagation matrix S is symmetric with $s_{12} = s_{21} = p$, $s_{11} = s_{22} = 1 - p$.

(b) In the anagenetic case, by Corollary 25 we need to solve the system of equations $(C - \lambda I)w^\star + Uw = 0$ for the state transition rates $q_1^2 = q_2^1 = rp$. The matrix B is

$$B = \begin{bmatrix} q_1^{11} - q_1^{22} - q_1^2 & 2q_2^{11} + q_2^{12} + q_2^1 \\ 2q_1^{22} + q_1^{12} + q_1^2 & q_2^{22} - q_2^{11} - q_2^1 \end{bmatrix} = r\begin{bmatrix} 1 - p & p \\ p & 1 - p \end{bmatrix},$$

with eigenvalue $\lambda = r$ and corresponding right eigenvector $u = [1/2, 1/2]^T$. Also $q_1 = q_2 = r + rp$, the matrix $C - \lambda I = C - rI$ is

$$C - rI = \begin{bmatrix} -q_1 - r & q_2^1 & 0 & 0 \\ q_1^2 & -q_2 - r & 0 & 0 \\ 0 & 0 & -q_1 - r & q_2^1 \\ 0 & 0 & q_1^2 & -q_2 - r \end{bmatrix}$$

$$= r\begin{bmatrix} -(2 + p) & p & 0 & 0 \\ p & -(2 + p) & 0 & 0 \\ 0 & 0 & -(2 + p) & p \\ 0 & 0 & p & -(2 + p) \end{bmatrix}$$

and the matrix U is

$$U = \begin{bmatrix} 2q_1^{11} & q_2^{22} & 0 & 0 & 0 & 0 \\ 0 & q_1^{11} & 2q_2^{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2q_1^{11} & q_2^{22} & 0 \\ 0 & 0 & 0 & 0 & q_1^{11} & 2q_2^{22} \end{bmatrix} = \begin{bmatrix} 2r & r & 0 & 0 & 0 & 0 \\ 0 & r & 2r & 0 & 0 & 0 \\ 0 & 0 & 0 & 2r & r & 0 \\ 0 & 0 & 0 & 0 & r & 2r \end{bmatrix}$$

Solving the above system for $q_1^2 = q_2^1 = rp$ in terms of the asymptotic fractions of cherries and pendants $w_\ell = [w_\ell^{11}, w_\ell^{12}, w_\ell^{22}]$ for $\ell \in \{1, 2\}$ and $w^\star = [w_1^1, w_1^2, w_2^1, w_2^2]$ implies that the asymptotic fractions of cherries and pendants as well as p satisfy

$$p = \frac{2w_1^{11} + w_1^{12} - 2w_1^1}{w_1^1 - w_1^2} = \frac{w_1^{12} + 2w_1^{22} - 2w_1^2}{w_1^2 - w_1^1} = \frac{2w_2^{22} + w_2^{12} - 2w_2^2}{w_2^2 - w_2^2}$$
$$= \frac{w_2^{12} + 2w_2^{22} - 2w_2^2}{w_2^2 - w_2^1}.$$

The matrices A_1, A_2 are

$$A_1 = A_2 = \begin{bmatrix} -(q_1 + q_1) & 0 & 0 \\ 0 & -(q_1 + q_2) & 0 \\ 0 & 0 & -(q_2 + q_2) \end{bmatrix}$$
$$= r \begin{bmatrix} -2(1 + p) & 0 & 0 \\ 0 & -2(1 + p) & 0 \\ 0 & 0 & -2(1 + p) \end{bmatrix}$$

and the value of p should make the system of equations $\{(A_\ell - rI)w_\ell + \frac{1}{2}q_{(\ell)} = 0\}_{\ell=1,2}$ with $q_{(1)} = [r, 0, 0]^T$, $q_{(2)} = [0, 0, r]^T$ a consistent one. With p as above, birth and state transition rates then are

$$q_1^{11} = q_2^{22} = r, \ q_1^{12} = q_1^{22} = q_2^{12} = q_2^{11} = 0, \quad q_1^2 = q_2^1 = rp.$$

Our results can be used together with what is previously known about predictive accuracy of a reconstruction method, such as maximum parsimony, majority rule and maximum likelihood, for the ancestral states. Predictive accuracy is measured in terms of the expected value (over all sample trees in the random model) of the probability that the predicted state of the root is correct. There are a number of known results (Gascuel and Steel 2014; Mossel and Steel 2014) on when a reconstruction method for the state of the root in the tree is more accurate than a uniform guess on its value. For the above models of symmetric change of state (with $k = 2$) the results of Gascuel and Steel (2014) state that the predictive accuracy of the maximum parsimony method is asymptotically 1/2 iff $r \leq 6s$; and the predictive accuracy of any method is asymptotically 1/2 if $r \leq 4s$; where $s = rp$ denotes the substitution rate in this symmetric propagation model. Results of Mossel and Steel (2014) state that majority rule is more accurate than a uniformly random guess iff $r > 4s$. Our expressions for $p = s/r$ allow one to approximately determine whether in a given tree the state of the root can be accurately predicted by one of these methods or not.

Unfortunately, our results cannot be used in the extreme *'asymmetric change of state'* models, such as:

(c) *'asymmetric cladogenetic change'* model with

$$q_1^2 = q_2^1 = 0, \quad q_1^{11} = r, q_1^{12} = q_1^{22} = 0, q_2^{22} = r(1 - p)^2, q_2^{12} = r2p(1 - p),$$
$$q_2^{11} = rp^2$$

where $p \in (0, 1)$ is the probability of state change from 2 to 1 at a birth event; here the matrix $B = r \begin{bmatrix} 1 & 2p \\ 0 & 1 - 2p \end{bmatrix}$ has maximal eigenvalue $\lambda = r$ but it is reducible; and

(d) *'asymmetric anagenetic irreversible transition'* model with

$$q_1^{11} = q_2^{22} = r, \; q_1^{12} = q_1^{22} = q_2^{12} = q_2^{11} = 0, \quad q_1^2 = 0, \; q_2^1 = rp$$

here p is the relative rate of state transition along state 2 lineage; and the matrix $\boldsymbol{B} = r \begin{bmatrix} 1 & p \\ 0 & 1-p \end{bmatrix}$ is reducible as well.

We stress that the reason why we cannot use limiting fractions of different cherries and pendants in order to infer rates in the above two cases is due to extreme asymmetry in these two examples. Namely, had we chosen any less extreme form of asymmetry with no zero entries in $\boldsymbol{B}$ we could have applied our results just as we had in the symmetric cases (a) and (b).

5.4 Comparison for numbers of cherries in different models

Consider a general multiple state Yule tree on $k = 2$ states but without state transitions. Its overall birth rates $q_1(t) = \sum_{j_1 \le j_2} q_1^{j_1 j_2}(t)$ and $q_2(t) = \sum_{j_1 \le j_2} q_2^{j_1 j_2}(t)$ are generally not the same, which implies that the probabilities at which lineages of each state are chosen to be the next one to give birth are not the same (there is 'non-neutrality' in states). Let $a_1(t) := q_1(t)/(q_1(t) + q_2(t))$ and $a_2(t) := q_2(t)/(q_1(t) + q_2(t)) = 1 - a_1(t)$ denote the weights proportional which lineages of states 1 and 2, respectively, get chosen to give birth (see Remark 15). For any two such models $\{q_i^{j_1 j_2}(t)\}_{i, j_1 \le j_2 \in \{1,2\}}$ and $\{q'_i^{j_1 j_2}(t)\}_{i, j_1 \le j_1 \in \{1,2\}}$ we can compare the weights a_1 and a'_1 of choosing state 1 lineages. We provide a comparison between the asymptotic fraction of different types of cherries $\boldsymbol{w}_1 = [w_1^{11}, w_1^{12}, w_1^{22}]^{\mathrm{T}}$ and $\boldsymbol{w}_2 = [w_2^{11}, w_2^{12}, w_2^{22}]^{\mathrm{T}}$ in the two models based on the comparison of their weights $a_1(t)$ and $a_2(t) = 1 - a_1(t)$ of choosing a lineage of different states to give birth.

Proposition 27 *Assume that the birth rates* $\{q_i^{j_1 j_2}(t)\}_{i, j_1 \le j_2 \in \{1,2\}}$ *and* $\{q'_i^{j_1 j_2}(t)\}_{i, j_1 \le j_2 \in \{1,2\}}$ *in the two models are such that, their limits* $q_\ell^{j_1 j_2} := \lim_{t \to \infty} q_\ell^{j_1 j_2}(t)$, $q_\ell := \lim_{t \to \infty} q_\ell(t)$ *satisfy*

$$\frac{q_1^{11} - q_1^{22}}{q_1} = 1 + \frac{q_2^{11} - q_2^{22}}{q_2}, \quad \frac{q'_1^{11} - q'_1^{22}}{q'_1} = 1 + \frac{q'_2^{11} - q'_2^{22}}{q'_2} \tag{9}$$

Then, the asymptotic proportions of cherries of type 1 and type 2 in the two models satisfy monotonicity in terms of weights a_1 *and* a'_1 *given by*

$$a_1 < a'_1 \quad \Rightarrow \quad w_1^{j_1 j_2} < w'_1^{j_1 j_2}, \; \forall j_1 \le j_2 \text{ and } w_2^{j_1 j_2} > w'_2^{j_1 j_2}, \; \forall j_1 \le j_2$$

where $a_1 := \lim_{t \to \infty} a_1(t)$, $a'_1 := \lim_{t \to \infty} a'_1(t)$ *denote the limiting weights of state 1 lineages.*

Proof Theorem 23 implies that the vectors $\boldsymbol{w}_\ell = [w_\ell^{11}, w_\ell^{12}, w_\ell^{22}]$ for $\ell \in \{1, 2\}$ satisfy

$$\boldsymbol{w}_\ell = \lim_{t \to \infty} \boldsymbol{\eta}_\ell(t) = -u_\ell \lim_{t \to \infty} (\boldsymbol{A}_\ell(t) - \lambda \boldsymbol{I})^{-1} \lim_{t \to \infty} \boldsymbol{q}_{(\ell)}.$$

Since we are considering multiple state Yule models with state transition rates $q_1^2(t) = q_2^1(t) = 0$, the matrix $\boldsymbol{A}(t)$ from Proposition 19 depends only on $q_1(t)$ and $q_2(t)$.

Let $p_\ell^{j_1 j_2} := q_\ell^{j_1 j_2}/q_\ell$ denote the probabilities that a birth event at a lineage of state ℓ results in states j_1, j_2 (see Remark 15) in the limit as $t \to \infty$. Then $q_\ell^{j_1 j_2} = p_\ell^{j_1 j_2} a_\ell(q_1 + q_2)$ and the assumption on the limiting birth rates becomes

$$p_1^{11} + p_2^{22} = 1 + p_1^{22} + p_2^{11}$$

Substituting all this into the equations for $\boldsymbol{w}_\ell$ above, and using $a_2 = 1 - a_1$, we obtain expressions for $\boldsymbol{w}_1, \boldsymbol{w}_2$ that are written entirely in terms of probabilities $p_\ell^{j_1 j_2}$ and weight a_1:

$$w_1^{11} = \frac{a_1 p_1^{11} \left(p_1^{11} - p_1^{22}\right)}{2 p_1^{11} a_1 + a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 1},$$

$$w_1^{12} = \frac{a_1 p_1^{12} \left(p_1^{11} - p_1^{22}\right)}{2 p_1^{11} a_1 - a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 2},$$

$$w_1^{22} = \frac{a_1 p_1^{22} \left(p_1^{11} - p_1^{22}\right)}{2 p_1^{11} a_1 - 3 a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 3},$$

$$w_2^{11} = \frac{p_2^{11} \left(1 - p_1^{11} + p_1^{22}\right)(1 - a_1)}{2 p_1^{11} a_1 + a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 1},$$

$$w_2^{12} = \frac{p_2^{12} \left(1 - p_1^{11} + p_1^{22}\right)(1 - a_1)}{2 p_1^{11} a_1 - a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 2},$$

$$w_2^{11} = \frac{p_2^{22} \left(1 - p_1^{11} + p_1^{22}\right)(1 - a_1)}{2 p_1^{11} a_1 - 3 a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 3}.$$

In order to prove the monotonicity of $w_\ell^{j_1 j_2}$ as a function of a_1 for $\ell, j_1, j_2 \in \{1, 2\}$, it suffices to check their first derivate with respect to a_1:

$$\frac{\partial w_1^{11}}{\partial a_1} = \frac{p_1^{11} \left(p_1^{11} - p_1^{22}\right)\left(1 - p_1^{11} + p_1^{22}\right)}{\left(2 p_1^{11} a_1 + a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 1\right)^2} > 0,$$

$$\frac{\partial w_1^{12}}{\partial a_1} = \frac{p_1^{12} \left(p_1^{11} - p_1^{22}\right)\left(2 - p_1^{11} + p_1^{22}\right)}{\left(2 p_1^{11} a_1 - a_1 - 2 p_1^{22} a_1 - p_1^{11} + p_1^{22} + 2\right)^2} > 0,$$

$$\frac{\partial w_1^{22}}{\partial a_1} = \frac{p_1^{22} \left(p_1^{11} - p_1^{22}\right) \left(3 - p_1^{11} + p_1^{22}\right)}{\left(2p_1^{11}a_1 - 3a_1 - 2p_1^{22}a_1 - p_1^{11} + p_1^{22} + 3\right)^2} > 0,$$

$$\frac{\partial w_2^{11}}{\partial a_1} = \frac{-p_2^{11} \left(2 - \left(p_1^{11} - p_1^{22}\right) \left(1 + p_1^{11} - p_1^{22}\right)\right)}{\left(2p_1^{11}a_1 + a_1 - 2p_1^{22}a_1 - p_1^{11} + p_1^{22} + 1\right)^2} < 0,$$

$$\frac{\partial w_2^{12}}{\partial a_1} = \frac{-p_2^{12} \left(1 - \left(p_1^{11} - p_1^{22}\right)^2\right)}{\left(2p_1^{11}a_1 - a_1 - 2p_1^{22}a_1 - p_1^{11} + p_1^{22} + 2\right)^2} < 0,$$

$$\frac{\partial w_2^{22}}{\partial a_1} = \frac{-p_2^{22} \left(p_1^{11} - p_1^{22}\right) \left(1 - p_1^{11} + p_1^{22}\right)}{\left(2p_1^{11}a_1 - 3a_1 - 2p_1^{22}a_1 - p_1^{11} + p_1^{22} + 3\right)^2} < 0,$$

and the result follows. $\qquad\square$

The assumption for the limiting birth rates can be satisfied in relevant models. For example, in a process where at a birth event states of the two continuing lineages are assigned according to a Markov process with transition probabilities $(s_{ij})_{i,j\in\{1,2\}}$, the probabilities $p_\ell^{j_1 j_2}$ are

$$p_1^{11} = (1 - s_{12})^2, \quad p_1^{12} = 2(1 - s_{12})s_{12}, \quad p_1^{22} = (s_{12})^2,$$
$$p_2^{22} = (1 - s_{21})^2, \quad p_2^{12} = 2(1 - s_{21})s_{21}, \quad p_2^{11} = (s_{21})^2,$$

and the assumption (9) is equivalent to $s_{12} + s_{21} = 1/2$.

Remark 28 In the special case when the overall birth rates are equal $a_1 = a_2 = 1/2$ ('neutral' underlying tree shape), if edge lengths in the multiple state Yule tree without state transitions are ignored the resulting distribution on the tree is that of a corresponding multiple state ERM model. Accordingly, the asymptotic fractions of cherries we obtained in the proof of Proposition 27 are in fact the same as asymptotic fractions obtained in Theorem 6 for the multiple state ERM trees with probabilities $\{p_i^{j_1 j_2}\}_{i,\, j_1 \le j_2 \in \{1,2\}}$.

Appendix

Proof of Lemma 12. In the event that $\mathbf{Z}(T) = 0$ there is nothing to prove, so we consider $\mathbf{W}$ on the event $\mathbf{Z}(T) \ne 0 \Leftrightarrow \mathbf{W}(0) \ne 0$ (and $\mathbf{W}(T) \ne 0$ as well).

For any $n \ge 1$ let $0 \le t_0 \le t_1 \le \cdots \le t_n \le T$, we denote the joint distribution of $\mathbf{W}$ at these times by

$$P_{t_0; t_1, \ldots, t_n}(z_0; \mathbf{w}_1, \ldots, \mathbf{w}_n) = \mathbb{P}\left[\mathbf{W}(t_j) = \mathbf{w}_j,\ 1 \le j \le n \mid \mathbf{Z}(t_0) = z_0\right].$$

We first show, by induction, that $\forall n \geq 1$

$$P_{t_0;t_1,\ldots,t_n}(z_0; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$$
$$= P_{t_0;t_1,\ldots,t_{n-1}}(z_0; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-1}) \frac{P_{t_0;t_{n-1},t_n}(z_0; \boldsymbol{w}_{n-1}, \boldsymbol{w}_n)}{P_{t_0;t_{n-1}}(z_0; \boldsymbol{w}_{n-1})}. \tag{10}$$

This is evident for $n = 2$. Assume the equation is true $\forall i \leq n - 1$ with $n > 2$. Notice that

$$P_{t_0;t_1,\ldots,t_n}(z_0; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$$
$$= \sum_{z_1 \geq \boldsymbol{w}_1} \mathbb{P}[\boldsymbol{Z}(t_1) = z_1 | \boldsymbol{Z}(t_0) = z_0] P_{t_1;t_1,\ldots,t_n}(z_1; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n). \tag{11}$$

The branching property of the birth-death process $\boldsymbol{Z}$ guarantees independence of its subtrees originating from non-overlapping subsets of individuals present at any time t_1. Since all individuals surviving at time T must be descendants of the process $\boldsymbol{W}$, we have

$$P_{t_1;t_1,\ldots,t_n}(z_1; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) = \mathbb{P}\left[\boldsymbol{W}(t_j) = \boldsymbol{w}_j, \ 1 \leq j \leq n \,\big|\, \boldsymbol{Z}(t_1) = z_1\right]$$
$$= C_{z_1,\boldsymbol{w}_1} \mathbb{P}\left[\boldsymbol{W}(t_j)\right.$$
$$= \boldsymbol{w}_j, \ 1 \leq j \leq n \,\big|\, \boldsymbol{Z}(t_1) = \boldsymbol{w}_1\right] p^{\mathbf{0}}_{z_1 - \boldsymbol{w}_1}(t_1, T)$$
$$= C_{z_1,\boldsymbol{w}_1} P_{t_1;t_1,\ldots,t_n}(\boldsymbol{w}_1; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) p^{\mathbf{0}}_{z_1 - \boldsymbol{w}_1}(t_1, T) \tag{12}$$

where $C_{z_1,\boldsymbol{w}_1}$ denotes the combinatorial number of distinct ways of choosing $\boldsymbol{w}_1$ out of z_1 individuals, and $p^{\mathbf{0}}_z(t, T) = \mathbb{P}[\boldsymbol{Z}(T) = 0 | \boldsymbol{Z}(t) = z]$ is the extinction probability by time T of the process $\boldsymbol{Z}$ started at time t with $\boldsymbol{Z}(t) = z$.

Given $\boldsymbol{Z}(t_1) = \boldsymbol{w}_1$, the process $(\boldsymbol{Z}(t))_{t \geq t_1}$ is the sum of birth-death processes defined by subtrees $\{\mathcal{T}^{(i)}\}$, $i = 1, \ldots, |\boldsymbol{w}_1|$, originated by one of each of the $|\boldsymbol{w}_1|$ individuals at time t_1. We may assume that each $\mathcal{T}^{(i)}$ is started by an individual of state $\tau^{(i)}$, where $\tau^{(1)}, \ldots, \tau^{(|\boldsymbol{w}_1|)}$ is some ordering of the $|\boldsymbol{w}_1|$ surviving originator states. Probability for the surviving lineages is

$$P_{t_1;t_1,\ldots,t_n}(\boldsymbol{w}_1; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$$
$$= \mathbb{P}\left[\boldsymbol{W}(t_j) = \boldsymbol{w}_j, \ 1 \leq j \leq n \,\big|\, \boldsymbol{Z}(t_1) = \boldsymbol{w}_1\right]$$
$$= \mathbb{P}\left[\boldsymbol{W}(t_j)(\mathcal{T}^{(i)}) \neq 0 \ \forall i, \ \sum_{i=1}^{|\boldsymbol{w}_1|} \boldsymbol{W}(t_j)(\mathcal{T}^{(i)}) = \boldsymbol{w}_j, \ \forall 2 \leq j \leq n\right]$$

where $\boldsymbol{W}(t)(\mathcal{T}^{(i)})$ denotes the number of individuals of $\mathcal{T}^{(i)}$ at time t which have a surviving lineage at time T. Since the subtrees $\mathcal{T}^{(i)}$ are independent

$$P_{t_1;t_1,\ldots,t_n}(\boldsymbol{w}_1; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$$
$$= \sum_{\substack{\forall 2 \leq j \leq n, \ \left(\boldsymbol{w}_j^{(i)}\right)_{1 \leq i \leq |\boldsymbol{w}_1|}: \\ \boldsymbol{w}_j^{(i)} > 0, \ \sum_{i=1}^{|\boldsymbol{w}_1|} \boldsymbol{w}_j^{(i)} = \boldsymbol{w}_j}} \prod_{i=1}^{|\boldsymbol{w}_1|} P_{t_1;t_2,\ldots,t_n}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_2^{(i)}, \ldots, \boldsymbol{w}_n^{(i)}\right), \tag{13}$$

where $\boldsymbol{e}_i$ denotes the unit k-dimensional vector whose i-th coordinate is 1 and all other coordinates are 0, and the summation is over all possible decompositions of $\boldsymbol{w}_j$ into vectors $(\boldsymbol{w}_j^{(i)})_{i=1,\ldots,|\boldsymbol{w}_1|}$ with all nonzero coordinate values, for each $j = 2, \ldots, n$. By the inductive hypothesis (10) for $n-1$, the probabilities in the product on the right side are equal to

$$
P_{t_1;t_2,\ldots,t_n}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_2^{(i)}, \ldots, \boldsymbol{w}_n^{(i)}\right)
$$

$$
= P_{t_1;t_2,\ldots,t_{n-1}}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_2^{(i)}, \ldots, \boldsymbol{w}_{n-1}^{(i)}\right) \frac{P_{t_1;t_{n-1},t_n}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_{n-1}^{(i)}, \boldsymbol{w}_n^{(i)}\right)}{P_{t_1;t_{n-1}}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_{n-1}^{(i)}\right)}
$$

$$
= P_{t_1;t_2,\ldots,t_{n-1}}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_2^{(i)}, \ldots, \boldsymbol{w}_{n-1}^{(i)}\right) \mathbb{P}\left[\boldsymbol{W}(t_n) = \boldsymbol{w}_n^{(i)} | \boldsymbol{W}(t_{n-1})\right.
$$

$$
= \boldsymbol{Z}(t_{n-1}) = \boldsymbol{w}_{n-1}^{(i)}\Big]
$$

where the last equality follows from (11) and (12) since

$$
P_{t_1;t_{n-1},t_n}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_{n-1}^{(i)}, \boldsymbol{w}_n^{(i)}\right)
$$

$$
= \mathbb{P}\left[\boldsymbol{W}(t_{n-1}) = \boldsymbol{w}_{n-1}^{(i)}, \boldsymbol{W}(t_n) = \boldsymbol{w}_n^{(i)} | \boldsymbol{Z}(t_1) = \boldsymbol{e}_{\tau^{(i)}}\right]
$$

$$
= \sum_{z_{n-1} \geq \boldsymbol{w}_{n-1}} \mathbb{P}\left[\boldsymbol{Z}(t_n - 1) = z_{n-1} | \boldsymbol{Z}(t_1) = \boldsymbol{e}_{\tau^{(i)}}\right] C_{z_{n-1};\boldsymbol{w}_{n-1}^{(i)}} p_{z_{n-1}-\boldsymbol{w}_{n-1}}^{\mathbf{0}}(t_{n-1}, T)
$$

$$
\times \mathbb{P}\left[\boldsymbol{W}(t_n) = \boldsymbol{w}_n^{(i)} | \boldsymbol{W}(t_{n-1}) = \boldsymbol{Z}(t_{n-1}) = \boldsymbol{w}_{n-1}^{(i)}\right]
$$

$$
= \mathbb{P}\left[\boldsymbol{W}(t_n) = \boldsymbol{w}_n^{(i)} | \boldsymbol{W}(t_{n-1}) = \boldsymbol{Z}(t_{n-1}) = \boldsymbol{w}_{n-1}^{(i)}\right] P_{t_1;t_{n-1}}\left(\boldsymbol{e}_{\tau^{(i)}}; \boldsymbol{w}_{n-1}^{(i)}\right).
$$

As the first factor on the right side above does not depend on $(\boldsymbol{w}_n^{(i)})_{i=1,\ldots,|\boldsymbol{w}_1|}$ the sum in (13) may be split into outer sums, over $2 \leq j \leq n-1$, and an inner sum, over $j = n$ that is equal to

$$
\sum_{\substack{(\boldsymbol{w}_n):\boldsymbol{w}_n^{(i)}>0,\\ \sum_{i=1}^{|\boldsymbol{w}_1|} \boldsymbol{w}_n^{(i)}=\boldsymbol{w}_n}} \prod_{i=1}^{|\boldsymbol{w}_1|} \mathbb{P}\left[\boldsymbol{W}(t_n) = \boldsymbol{w}_n^{(i)} | \boldsymbol{W}(t_{n-1}) = \boldsymbol{Z}(t_{n-1}) = \boldsymbol{w}_{n-1}^{(i)}\right].
$$

By the same argument using splitting over independent subtrees, but this time splitting the individuals at time t_{n-1} into subsets of sizes $(\boldsymbol{w}_{n-1}^{(i)})_{i=1,\ldots,|\boldsymbol{w}_1|}$, we can show that this sum contributes to the outer sums a factor of

$$
\mathbb{P}\left[\boldsymbol{W}(t_n) = \boldsymbol{w}_n | \boldsymbol{W}(t_{n-1}) = \boldsymbol{Z}(t_{n-1}) = \boldsymbol{w}_{n-1}\right] = \frac{P_{t_0;t_{n-1},t_n}(z_0; \boldsymbol{w}_{n-1}, \boldsymbol{w}_n)}{P_{t_0;t_{n-1}}(z_0; \boldsymbol{w}_{n-1})},
$$

where the last equality follows again from Eqs. (11) and (12), and combining with the outer sums in (13) implies

$$P_{t_1;t_1,\ldots,t_n}\left(\boldsymbol{w}_1;\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n\right)$$
$$= P_{t_1;t_1,\ldots,t_{n-1}}\left(\boldsymbol{w}_1;\boldsymbol{w}_1,\ldots,\boldsymbol{w}_{n-1}\right)\frac{P_{t_0;t_{n-1},t_n}\left(\boldsymbol{z}_0;\boldsymbol{w}_{n-1},\boldsymbol{w}_n\right)}{P_{t_0;t_{n-1}}\left(\boldsymbol{z}_0;\boldsymbol{w}_{n-1}\right)},$$

as wanted. By using once again Eqs. (11) and (12), this becomes Eq. (10) for step n. Equation (10) may be written in terms of conditional probabilities as

$$\mathbb{P}\left[\boldsymbol{W}\left(t_n\right)=\boldsymbol{w}_n \,\middle|\, \boldsymbol{W}\left(t_j\right)=\boldsymbol{w}_j,\ 1\leq j\leq n-1,\ \boldsymbol{Z}\left(t_0\right)=\boldsymbol{z}_0\right]$$
$$= \mathbb{P}\left[\boldsymbol{W}\left(t_n\right)=\boldsymbol{w}_n \,\middle|\, \boldsymbol{W}\left(t_{n-1}\right)=\boldsymbol{w}_{n-1},\ \boldsymbol{Z}\left(t_0\right)=\boldsymbol{z}_0\right]$$

which implies the Markov property for $(\boldsymbol{W}(t))_{t\geq 0}$. □

References

Aldous DJ (1996) Probability distributions on cladograms. Random Discrete Structures, (IMA Volumes Math Appl 76), pp 1–18

Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat Sci 16(1):23–34

Aldous D, Popovic L (2005) A critical branching process model for biodiversity. Adv Appl Probab 37:1094–1115. doi:10.1239/aap/1134587755

Fitzjohn RG (2010) Quantitative traits and diversification. Syst Biol 59:619–633

Fitzjohn RG (2012) What drives biological diversification? Detecting traits under species selection. University of British Columbia, PhD Thesis

Gascuel O, Steel M (2014) Predicting the ancestral character changes in a tree is typically easier than predicting the root state. Syst Biol 63(3):421–435

Goldberg EE, Igic B (2012) Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709

Goldberg EE, Lancaster LT, Ree RH (2011) Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Syst Biol 60:451–465

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

Janson S (2004) Functional limit theorems for multitype branching processes and generalized Pólya urns. Stoch Process Appl 110(2):177–245

Jones G (2011) Calculations for multi-type age-dependent binary branching processes. J Math Biol 63(1):33–56

Lambert A, Popovic L (2013) The coalescent point-process of branching trees. Ann Appl Probab 23(1):99–144. doi:10.1214/11-AAP820

Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. Syst Biol 56(5):701–710

McKenzie A, Steel M (2000) Distributions of cherries for two models of trees. Math Biosci 164(1):81–92

Mode CJ (1962) Some multi-dimensional birth and death processes and their applications in population genetics. Int Biometric Soc 18(4):543–567

Mooers AO, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 72(1):31–54

Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In: Gascuel O (ed) Mathematics of evolution and phylogeny, chap 14. Oxford University Press, Oxford, pp 384–412

Mossel E, Steel M (2014) Majority rule has transition ration 4 on Yule trees under a 2-state symmetric model. J Theor Biol 18(360):315–318

Nee S, May RH, Harvey PH (1994) The reconstructed evolutionary process. Philos Trans Roy Soc B 344(1309):305–311

NG J, Smith SD (2014) How traits shape trees: new approaches for detecting character state-dependent lineage diversification. J Evol Bio. doi:10.1111/jeb.12460

Popovic L, Rivas M (2014) The coalescent point-process of multi-type branching trees. Stoch Process Appl 124(12):4120–4148

Smythe RT (1996) Central limit theorems for urn models. Stoch Process Appl 65(1):115–137

Yule GU (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