*Adv. Appl. Prob.* **49,** 108–133 (2017)
*doi:*10.1017.apr.2016.81

# GENERALIZED MARKOV BRANCHING TREES

HARRY CRANE,[*] *Rutgers University*

## Abstract

Motivated by the gene tree/species tree problem from statistical phylogenetics, we extend the class of Markov branching trees to a parametric family of distributions on fragmentation trees that satisfies a *generalized Markov branching property*. The main theorems establish important statistical properties of this model, specifically necessary and sufficient conditions under which a family of trees can be constructed consistently as sample size grows. We also consider the question of attaching random edge lengths to these trees.

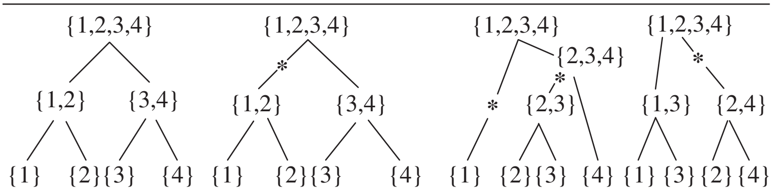
*Keywords:* Markov branching tree; exchangeable random partition; exchangeable fragmentation; beta-splitting model; coalescent process; hidden Markov model

2010 Mathematics Subject Classification: Primary 60G09; 60G05
Secondary 92D10

## 1. Introduction

In this paper we establish basic properties for the broad class of *generalized Markov branching tree models* introduced below. In particular, the main theorems characterize conditions under which generalized Markov branching tree models exhibit the properties of *label equivariance* and *consistency under subsampling*, which ensure that model-based inferences can be extended beyond the sample and are unaffected by arbitrary choices of labeling and sample size.

The proposed model is primarily motivated by phylogenetics applications, for which it incorporates a tree parameter in a natural and explicit way; see Figure 1.The model has further



| Species tree | Site 1 | Site 2 | Site 3 |
|---|---|---|---|
| 1 | A | C | A |
| 2 | A | G | C |
| 3 | T | G | A |
| 4 | T | T | C |

FIGURE 1: DNA sequence data for species labeled 1, 2, 3, 4. Mutations (*marked by* $*$) occur on each site specific tree and cause the resulting DNA configuration at that site. In applications, the array of DNA is observed, while the tree-valued sequence and its mutations are unobserved.

Received 3 September 2015; revision received 24 February 2016.

[*] Postal address: Department of Statistics and Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA. Email address: hcrane@stat.rutgers.edu

potential in hidden Markov modeling and tree search algorithms, as discussed in [19]. As a proper discussion requires substantial computational and scientific treatment, we leave these applications to future work. For consistency, we phrase the discussion in terms of a concrete problem in statistical phylogenetics; see Section 1.2.

### 1.1. Random tree models

Within probability, combinatorial stochastic process theory [27] establishes deep connections among classical population genetics [18], [23], coagulation and fragmentation processes [8], and Lévy processes [7]. In population genetics and phylogenetics applications, coalescent and fragmentation processes are among the most popular models [19]. Previous authors [1], [3], [4], [16], [17] have studied tree-valued Markov chains with different dynamics, but specifying an explicit data generating process on infinite phylogenetic trees is rarely the focus. See Evans's lecture notes [16] for a recent overview.

### 1.2. Phylogenetics motivation

In statistical phylogenetics, random tree models are most attractive among the many tools for inferring unknown hierarchical relationships in a population [19]. There are numerous competing approaches, such as the neighbor joining algorithm, but explicit probabilistic models for genetic variation offer additional benefits for interpreting perceived anomalies and estimating important quantities.

We assume there is one true evolutionary tree, called the *species tree*, which represents the relationships of all species to one another. Genetic data, e.g. mitochondrial DNA sequences, contain information about the species tree. Differences between DNA sequences reflect the species tree, but the relationship between genetic sequence data and the unknown species tree is obscured by recombination, migration between populations, and genetic drift [28], factors that cause the phylogenetic tree governing a particular site or gene to differ from the species tree. We call the tree associated to a given site or gene a *gene tree*.

As a consequence of biological processes such as recombination, each site along the chromosome corresponds to a gene tree. The observed DNA sequences are generated by mutations that occur on these gene trees. See Figure 1 for an illustration.

Previous authors [22], [28] note the discrepancy between gene and species trees and its effect on inference. In particular, McVean and Cardin [25] detailed the difficulty of approximating the coalescent under the influence of recombination; see [29] for recent developments. Because the authors modeled the relationship among the species tree, gene trees, and DNA sequences, likelihood-based methods are among the most attractive, but computational complexity remains a hindrance to their use in practice; see [19, Chapter 28].

### 1.3. Generalized Markov branching trees

We propose a probabilistic model for the latent gene tree sequence. *Generalized Markov branching trees* extend the theory of Markov branching trees [2], [10], [20], [21], [24] to a parametric model for phylogenetic inference. At a minimum, the proposed models have an interpretable, closed form density function, which remedies McVean and Cardin's issue. Special cases relate to classical theory of exchangeable random partitions [15], [18], [26] and coalescent processes [23].

In addition to the computational issues mentioned above, many tree models are devised without deference to basic logical properties of probabilistic models. In the main theorems below we establish necessary and sufficient conditions under which the generalized Markov

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

branching tree model is

- *label equivariant*, invariant under arbitrary relabeling of species, and
- *consistent under subsampling*, the gene tree for a subsample $[m] \subset [n]$ of $m \leq n$ species depends on the species tree only through its restriction to species labeled $[m] := \{1, \ldots, m\}$.

These properties coincide with the phylogenetic modeling axioms set forth by Aldous [2].

### 1.4. Outline

The paper is organized as follows. In Section 2 we give an informal description of the modeling framework and summarize the main theorems. Section 3 contains preliminary definitions and notation. In Section 4 we introduce generalized Markov branching tree models. In Section 5 we consider attaching random edge lengths.

## 2. Probabilistic framework

We now give an informal description of the main conclusions. A more formal discussion begins in Section 3. All figures show trees with binary splits, but the general treatment covers fragmentations with arbitrarily many children at each branch point.

Consider a collection of five species, labeled 1, 2, 3, 4, 5, which are related by the tree in Figure 2. This tree is interpreted from the bottom up: the bottom *leaves* are labeled by the species and any point where branches meet is labeled by the set of species below that point. In particular, from the tree in Figure 2 we see that species 1 and 2 are more closely related to each other than to species 3, 4, and 5, and that species 3 and 4 are more closely related to each other than to species 5. The root of the tree represents the ancestor of all species, called *eve*.

The principle of exchangeability prevails when modeling such a tree, that is, the probability of observing the tree in Figure 2 should not depend on arbitrary assignments of labels to species. The principle of consistency states that the tree governing the subsample $\{1, 2, 3, 4\}$ should agree with the tree in Figure 2, meaning the tree for species 1, 2, 3, 4 is just that tree obtained by deleting species 5, as in Figure 3. The consistency property permits the interpretation of the

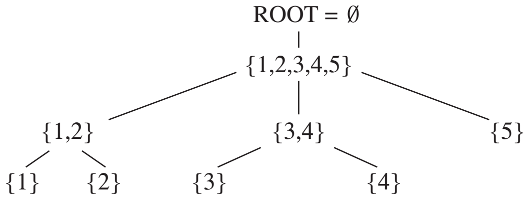


FIGURE 2.

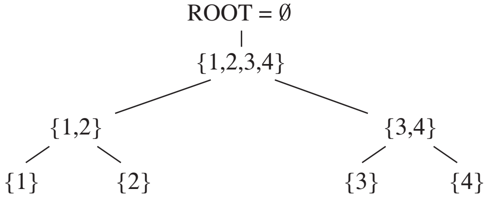


FIGURE 3.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

model as a data generating process for a larger population, from which the observed species have been sampled. Without knowledge to the contrary, we assume the population is infinite in size and labeled by the positive integers $\mathbb{N} := \{1, 2, \dots\}$.

Aldous [2] set forth the axioms of exchangeability and sampling consistency in phylogenetic modeling. In addition to these axioms, Aldous introduced the class of Markov branching trees, which are exchangeable, consistent random trees satisfying the *Markov branching property*, by which the branching below any vertex is conditionally independent of the rest of the tree. This property is natural in evolutionary modeling as it assumes that disjoint ancestral lines evolve independently of one another.

In Figure 1, we wish to model the latent tree sequence based on the true species tree, denoted $\boldsymbol{t}$. For this, we modify the family of Markov branching trees to a family of conditional probability distributions $Q(\cdot; \boldsymbol{t})$ on the space of phylogenetic trees, given the true species tree $\boldsymbol{t}$. In this case, the notions of exchangeability and sampling consistency give way to the conditions of *label equivariance* and *lack of interference*, respectively. Label equivariance implies that the distribution $Q(\cdot; \boldsymbol{t})$ does not depend on the assignment of labels to species, as long as the species are labeled consistently in both $\boldsymbol{t}$ and the realization $\boldsymbol{T} \sim Q(\cdot; \boldsymbol{t})$. Lack of interference works in accord with the assumption that missing observations occur completely at random, so that the observed gene tree for a subsample $[n] = \{1, \dots, n\}$ depends on the species tree $\boldsymbol{t}$ of the population $\mathbb{N}$ only through its restriction to the sample $[n] \subset \mathbb{N}$. We also modify the Markov branching property to the *conditional Markov branching property*, whereby the branching below each vertex is conditionally independent of the rest of the tree, given the species tree parameter $\boldsymbol{t}$.

The Markov branching property reduces much of the theory of Markov branching trees to a study of the associated *splitting rule*, which governs the branching below each point in the tree through a family of distributions on set partitions. We aim for the same mathematical and statistical tractability by specializing to a class of models for which the conditional distribution $Q(\cdot; \boldsymbol{t})$ depends on $\boldsymbol{t}$ through a (possibly random) sufficient statistic, which we take to be a random partition whose distribution depends on $\boldsymbol{t}$. In this way, the generalized Markov branching tree model is determined by a family of *conditional splitting rules*.

The expression of $Q(\cdot; \boldsymbol{t})$ in terms of a sufficient partition statistic strikes a balance between mathematical tractability and practical utility as well as connects our model to previous work on random trees. In the main theorems we establish necessary and sufficient conditions for constructing a phylogenetic tree model subject to the conditions of label equivariance, lack of interference, and the generalized Markov branching property. We also consider the task of consistently attaching edge lengths to these random trees, which incorporates the notion of evolutionary time into the model. We discuss practical issues for models of this kind in other related work [13], [14].

## 3. Preliminaries

### 3.1. Set partitions

Writing $A \subset_f \mathbb{N}$ to denote that $A$ is a finite subset of $\mathbb{N}$, a *partition* $\pi_A$ of $A$ is a collection $\{B_1, \dots, B_r\}$ of nonoverlapping, nonempty subsets for which $\bigcup_{i=1}^{r} B_i = A$. For any $\pi_A = \{B_1, \dots, B_r\}$, $B_1, \dots, B_r$ are called *blocks* and $\#\pi_A = r$ denotes the number of blocks. We write $\mathcal{P}_A$ for the collection of all partitions of $A$.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

For any partition $\pi \in \mathcal{P}_A$ and permutation $\sigma\colon A \to A$, we write $\pi^\sigma$ to denote the *relabeling* of $\pi$ by $\sigma$, where

(P) $i$ and $j$ are in the same block of $\pi^\sigma$ if and only if $\sigma^{-1}(i)$ and $\sigma^{-1}(j)$ are in the same block of $\pi$.

For $A' \subseteq A \subset_f \mathbb{N}$, we define the *restriction* of $\pi \in \mathcal{P}_A$ to $\mathcal{P}_{A'}$ by

$$\pi_{|A'} := \{A' \cap b\colon b \in \pi\} \setminus \{\varnothing\}. \tag{3.1}$$

We sometimes write $\boldsymbol{D}_{A',A}\colon \mathcal{P}_A \to \mathcal{P}_{A'}$ to denote the deletion map $\pi \mapsto \pi_{|A'}$. When $A' = [m]$ and $A = [n]$, $m \le n$, we write $\boldsymbol{D}_{m,n} = \boldsymbol{D}_{[m],[n]}$. For example, with $\pi = \{\{1,3,4\},\{2,6\},\{5\}\}$ and permutation $\sigma = (123)(456)$, in cycle notation, we have

$$\pi^\sigma = \{\{1,2,5\},\{3,4\},\{6\}\} \quad \text{and} \quad \pi_{|[4]} = \{\{1,3,4\},\{2\}\}.$$

The collection $\mathcal{P}_{\mathbb{N}}$ of partitions of $\mathbb{N}$ consists of sequences $(\pi_n,\ n \in \mathbb{N})$ of finite partitions for which $\pi_n \in \mathcal{P}_{[n]}$ and $\pi_m = \boldsymbol{D}_{m,n}\pi_n$ for every $m \le n$, for all $n \in \mathbb{N}$. For $\pi = (\pi_n) \in \mathcal{P}_{\mathbb{N}}$, we write $\boldsymbol{D}_n\pi = \pi_n$ to denote the restriction map $\mathcal{P}_{\mathbb{N}} \to \mathcal{P}_{[n]}$, for each $n \in \mathbb{N}$. We equip $\mathcal{P}_{\mathbb{N}}$ with the $\sigma$-field $\sigma\langle \boldsymbol{D}_n,\ n \in \mathbb{N}\rangle$ generated by these restriction maps.

### 3.2. Fragmentation trees

A *fragmentation* $\boldsymbol{t}_A$ of $A$ is a collection of subsets satisfying

(F1) $A \in \boldsymbol{t}_A$ and

(F2) if $\#A \ge 2$, then there is a partition $\pi_A = \{A_1, \ldots, A_r\}$ of $A$ such that

$$\boldsymbol{t}_A = \{A\} \cup \boldsymbol{t}_{A_1} \cup \cdots \cup \boldsymbol{t}_{A_r},$$

where $\boldsymbol{t}_{A_i}$ is a fragmentation of $A_i$ for each $i = 1, \ldots, r$.

We initialize the recursive definitions (F1) and (F2) by putting $\boldsymbol{t}_{\{i\}} := \{\{i\}, \varnothing\}$ for every $i \in \mathbb{N}$. We may call $\boldsymbol{t}_A$ a *fragmentation*, *fragmentation tree*, or *tree*. We write $\mathcal{T}_A$ to denote the set of fragmentations of $A$.

We can also regard a fragmentation $\boldsymbol{t}$ of $A \subset_f \mathbb{N}$ as a tree with a distinguished vertex 'ROOT' and leaves labeled by $A$. (For technical reasons, we identify ROOT with the empty set in the above definition.) By tracing the ancestral lineage from any leaf to ROOT, we label each internal vertex uniquely by the subset of leaves whose ancestral line passes through that vertex. Alternatively, if we begin at ROOT, we obtain a recursive fragmentation of the set of leaves. Below any vertex labeled $A' \subseteq A$ is a *subtree* $\boldsymbol{t}_{|A'}$ with leaves labeled in $A'$. In Figure 2 we see the leaf labeled tree corresponding to the fragmentation

$$\{\varnothing, \{1,2,3,4,5\}, \{1,2\}, \{3,4\}, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}.$$

For any $A' \subseteq A \subset_f \mathbb{N}$, the *restriction*, or *reduced subtree*, of $\boldsymbol{t} \in \mathcal{T}_A$ to $A'$ is defined by

$$\boldsymbol{t}_{|A'} := \boldsymbol{R}_{A',A}\boldsymbol{t} := \{b \cap A'\colon b \in \boldsymbol{t}\}. \tag{3.2}$$

For integers $1 \le m \le n$, we write $\boldsymbol{R}_{m,n}$ to denote the restriction map $\mathcal{T}_{[n]} \to \mathcal{T}_{[m]}$. Through these restriction maps, we can define an *infinite fragmentation* of $\mathbb{N}$ as a compatible sequence of finite trees $(\boldsymbol{t}_n,\ n \in \mathbb{N})$, that is, $\boldsymbol{t}_n \in \mathcal{T}_{[n]}$ and $\boldsymbol{R}_{m,n}\boldsymbol{t}_n = \boldsymbol{t}_m$ for all $m \le n$, for every $n \in \mathbb{N}$. For

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

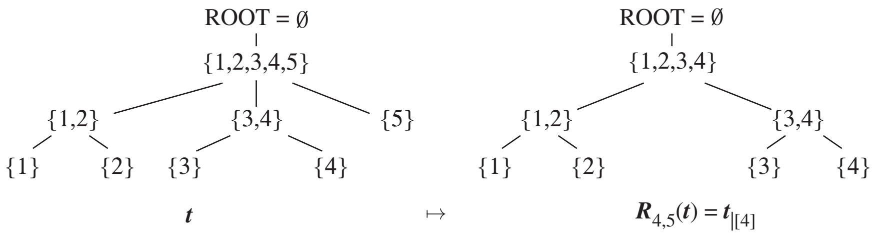


FIGURE 4.

each $n \in \mathbb{N}$, we denote the restriction of $\boldsymbol{t} := (\boldsymbol{t}_n,\ n \in \mathbb{N}) \in \mathcal{T}_{\mathbb{N}}$ to $\mathcal{T}_{[n]}$ by $\boldsymbol{R}_n\boldsymbol{t} := \boldsymbol{t}_n$. We indicate this space of infinite trees by $\mathcal{T}_{\mathbb{N}}$, which we equip with the $\sigma$-algebra $\sigma\langle \boldsymbol{R}_n,\ n \in \mathbb{N}\rangle$ that makes the restriction maps measurable. In Figure 4 we demonstrate the action of $\boldsymbol{R}_{4,5}\colon \mathcal{T}_{[5]} \to \mathcal{T}_{[4]}$ by removing species 5.

We mostly study exchangeable random fragmentations, which are invariant under relabeling of the leaves of the tree. In particular, the relabeling of $\boldsymbol{t} = \{b\} \in \mathcal{T}_{[n]}$ by any permutation $\sigma\colon [n] \to [n]$ is defined by $\boldsymbol{t}^{\sigma} := \{b^{\sigma},\ b \in \boldsymbol{t}\}$, where $b^{\sigma} := \{\sigma(i),\ i \in b\}$.

3.2.1. *Phylogenetic terminology.* For clarity we phrase much of our discussion in terms of the phylogenetics application in Figure 1. Let $\boldsymbol{t} \in \mathcal{T}_A$ be a fragmentation of $A \subset_f \mathbb{N}$. In phylogenetic terms, we call any $b \in \boldsymbol{t}$ an *ancestor* of $a \in \boldsymbol{t}$ if $a \subset b$. Therefore, $A$ is the unique element of $\boldsymbol{t}_A \in \mathcal{T}_A$ that is ancestral to all species, the so-called *most recent common ancestor* $\mathrm{MRCA}(\boldsymbol{t}_A) = A$. For $a \notin \{A, \varnothing\}$, the nonempty interval $(a, \mathrm{MRCA}(\boldsymbol{t})] := \{b \in \boldsymbol{t}\colon a \subset b\}$ has a unique minimal element $\mathrm{PA}(a) := \bigcap_{a \subset b \subseteq A} b$, called the *parent* of $a$. Conversely, except for singletons, every nonempty $a \in \boldsymbol{t}$ is the parent of some collection of nonempty subsets of $\boldsymbol{t}$, called the *children* of $a$ and denoted $\mathrm{CH}_{\boldsymbol{t}}(a) := \mathrm{PA}_{\boldsymbol{t}}^{-1}(a)$. Specifically, for any $A' \in \boldsymbol{t}_A$, the blocks of the partition $\pi_{A'}$ in (F2) are the *children* of $A'$. We call the children of $A = \mathrm{MRCA}(\boldsymbol{t}_A)$ the *root partition*, which we assign the special notation $\Pi_{\boldsymbol{t}_A} := \mathrm{CH}_{\boldsymbol{t}_A}(\mathrm{MRCA}(\boldsymbol{t}_A))$.

## 3.3. Markov branching trees

A *Markov branching tree* with $n$ leaves is a random fragmentation $\boldsymbol{T}$ of $[n]$ whose distribution satisfies the *Markov branching property*, i.e. for every $b \in \boldsymbol{T}$ with $\#b \geq 2$, the conditional distribution of $\mathrm{CH}_{\boldsymbol{T}}(b)$, given $b \in \boldsymbol{T}$, is independent of $\boldsymbol{T}_{|[n]\setminus b}$. In other words, any collection $\{\boldsymbol{T}_1, \dots, \boldsymbol{T}_r\}$ of nonoverlapping subtrees in $\boldsymbol{T}$ is conditionally independent given the nonoverlapping subsets $\{B_1, \dots, B_r\}$ labeling the leaves of each subtree. By this description, it is sufficient to specify a family $(p_b,\ b \subseteq [n])$ of *splitting rules*, where each $p_b$ is a probability distribution on the children of each possible parent in $\boldsymbol{T}$. In particular, each $p_b$ is a probability measure on $\mathcal{P}_b \setminus \{\mathbf{1}_b\}$, partitions of $b$ with the one block partition $\mathbf{1}_b$ removed.

A collection $p := (p_b,\ b \subseteq [n])$ of splitting rules determines a *Markov branching distribution* $Q_p^{[n]}$ on $\mathcal{T}_{[n]}$ as follows. To generate $\boldsymbol{T} \in \mathcal{T}_{[n]}$, we first generate its root partition $\Pi_{\boldsymbol{T}}$ from $p_{[n]}$. Given $\Pi_{\boldsymbol{T}} = \pi$, we iterate independently for each child $b \in \pi$, generating the children of each $b \in \pi$ from $p_b$. We repeat this procedure in subtrees until we reach the configuration of singletons. By conditional independence, $Q_p^{[n]}$ has the product form

$$Q_p^{[n]}(\boldsymbol{t}) = \prod_{\{b \in \boldsymbol{t}\colon \#b \geq 2\}} p_b(\mathrm{CH}_{\boldsymbol{t}}(b)), \qquad \boldsymbol{t} \in \mathcal{T}_{[n]}. \tag{3.3}$$

Since we deal exclusively with exchangeable models, it is sufficient to specify a collection $p := (p_m,\ 2 \leq m \leq n)$ of splitting rules indexed by $\{2, \dots, n\}$. For $m = 2, \dots, n$, $p_m$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

determines a splitting rule on $\mathcal{P}_{[m]} \setminus \{\mathbf{1}_{[m]}\}$ and, by exchangeability, $\mathcal{P}_b \setminus \{\mathbf{1}_b\}$ for every $b \subseteq [n]$ with $\#b = m$. To avoid unnecessary notation, we write $p_b$ to denote the splitting rule induced by $p_{\#b}$ on $\mathcal{P}_b \setminus \{\mathbf{1}_b\}$ through exchangeability.

A *Markov branching model* is a probability measure $Q_p$ on $\mathcal{T}_{\mathbb{N}}$ whose finite-dimensional distributions $(Q_p^{[n]},\ n \geq 1)$ have the form (3.3). Alternatively, we specify $Q_p$ from its finite-dimensional distributions $(Q_p^{[n]})_{n\geq 1}$, provided these are *consistent under subsampling*, i.e.

$$Q_p^{[m]} := Q_p^{[n]} \boldsymbol{R}_{m,n}^{-1} \quad \text{for all } m \leq n, \tag{3.4}$$

where $\boldsymbol{R}_{m,n} \colon \mathcal{T}_{[n]} \to \mathcal{T}_{[m]}$ is the restriction map defined in (3.2). Since the finite-dimensional distributions only depend on $p := (p_n,\ n \geq 2)$, $(Q_p^{[n]},\ n \geq 1)$ satisfies (3.4) if and only if

$$p_n(\pi) = p_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi)) + p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)})p_n(\pi) \quad \text{for all } \pi \in \mathcal{P}_{[n]}, \tag{3.5}$$

where $\boldsymbol{e}_n^{(n)} := \{\{1, \dots, n-1\}, \{n\}\}$ for every $n \geq 2$. In words, (3.5) reflects all possible ways to construct the root partition for a fragmentation $\boldsymbol{t}$ of $n+1$ elements, given the root partition $\pi$ of its restriction $\boldsymbol{t}_{|[n]}$ to $n$ elements. The first term reflects the probability that the root partitions are compatible, i.e. $\boldsymbol{D}_{n,n+1}(\Pi_{\boldsymbol{t}}) = \Pi_{\boldsymbol{t}_{|[n]}}$, while the second term reflects the possibility that the new element $n+1$ branches from $[n]$ immediately, i.e. $\Pi_{\boldsymbol{t}} = \{[n], \{n+1\}\}$, and then $[n]$ branches into $\pi$ in the subtree $\boldsymbol{t}_{|[n]}$. The model below extends the class of Markov branching trees to incorporate a tree parameter $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$, which represents the species tree for the population $\mathbb{N}$.

## 4. Generalized Markov branching trees

For $n \in \mathbb{N}$, let $p_n(\cdot; \pi)$ be a probability distribution on $\mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\}$ for every $\pi \in \mathcal{P}_{[n]}$. We call $p_n := (p_n(\cdot; \pi),\ \pi \in \mathcal{P}_{[n]})$ a *conditional splitting rule* on $\mathcal{P}_{[n]}$. Recalling the relabeling operation for partitions (P), we call a conditional splitting rule $p_n$ *exchangeable* if

$$p_n(\pi'; \pi) = p_n(\pi'^{\sigma}; \pi^{\sigma}) \quad \text{for all } \pi \in \mathcal{P}_{[n]},\ \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\} \tag{4.1}$$

for every permutation $\sigma \colon [n] \to [n]$. As long as $p_n$ is exchangeable, we can uniquely define a conditional splitting rule $p_b$, for every $b \subset \mathbb{N}$ with $\#b = n$.

For a collection $p := (p_m,\ 2 \leq m \leq n)$ of conditional splitting rules, we define a *conditional Markov branching distribution* on $\mathcal{T}_{[n]}$ by

$$Q_p^{[n]}(\boldsymbol{t}; \pi) := \prod_{\{b \in \boldsymbol{t} \colon \#b \geq 2\}} p_b(\mathrm{CH}_{\boldsymbol{t}}(b); \pi_{|b}), \qquad \boldsymbol{t} \in \mathcal{T}_{[n]},\ \pi \in \mathcal{P}_{[n]}. \tag{4.2}$$

In words, $Q_p^{[n]}(\cdot; \pi)$ incorporates a partition parameter into (3.3) by generating the children of each $b \in \boldsymbol{T} \sim Q_p^{[n]}(\cdot; \pi)$ from the conditional splitting rule that depends on the restriction of $\pi$ to $b$. The connection between the conditional Markov branching distribution and Markov branching trees should be clear by comparing (3.3) and (4.2). Specifically, if the conditional splitting rules depend trivially on the partition parameter, i.e. $p_b(\cdot; \pi) = p_b(\cdot; \pi')$ for all $\pi, \pi' \in \mathcal{P}_b$ and all $b \subseteq [n]$, then (4.2) reduces to (3.3).

In our phylogenetic model, the dependence of $Q_p^{[n]}(\cdot; \pi)$ on $\pi$ simplifies the dependence of the gene trees on the latent species tree $\boldsymbol{t}$. In terms of Figure 1, $\boldsymbol{t}$ is the species tree and each of the site specific gene trees is modeled by a distribution that depends on $\boldsymbol{t}$ by mixing with respect to the *sufficient partition statistic*, as in (4.3) below. We point out that the sufficient partition statistic has no tangible role in the data in Figure 1. Its only role is that it permits us

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

to incorporate dependence of a random tree on the latent species tree $\boldsymbol{t}$ in a mathematically and computationally tractable way.

To connect the distribution in (4.2) to a species tree $\boldsymbol{t} \in \mathcal{T}_{[n]}$, we specify a *sufficiency measure* $\mu_n(\cdot \mid \boldsymbol{t})$, which is a conditional probability distribution on the space of partitions of $[n]$, given $\boldsymbol{t} \in \mathcal{T}_{[n]}$. To model dependence of a random tree on a species tree $\boldsymbol{t}$, we define the $(\mu, p)$*-Markov branching distribution* $Q^{[n]}_{\mu,p}$ on $\mathcal{T}_{[n]}$ by mixing $Q^{[n]}_p$ in (4.2) with respect to $\mu_n$,

$$Q^{[n]}_{\mu,p}(\boldsymbol{t}'; \boldsymbol{t}) := \sum_{\pi \in \mathcal{P}_{[n]}} Q^{[n]}_p(\boldsymbol{t}'; \pi)\mu_n(\pi \mid \boldsymbol{t}), \quad \boldsymbol{t}, \boldsymbol{t}' \in \mathcal{T}_{[n]}. \tag{4.3}$$

We say that $\pi \sim \mu_n(\cdot \mid \boldsymbol{t})$ is *sufficient* for $\boldsymbol{t}$ in the sense that the conditional distribution of $\boldsymbol{T} \sim Q^{[n]}_{\mu,p}(\cdot; \boldsymbol{t})$, given $\pi \sim \mu_n(\cdot \mid \boldsymbol{t})$, depends only on $\pi$ and not on $\boldsymbol{t}$.

After conditioning on the realization of $\Pi \sim \mu_n(\cdot \mid \boldsymbol{t})$, $\boldsymbol{T} \sim Q^{[n]}_{\mu,p}(\cdot; \boldsymbol{t})$ depends only on a sufficient statistic, which is easier to handle than the more complicated structure of $\boldsymbol{t}$. After mixing with respect to $\mu_n(\cdot \mid \boldsymbol{t})$, the unconditional distribution of $\boldsymbol{T}$ depends on the entire species tree, and so our model establishes a direct link between the parameter of interest $\boldsymbol{t}$ and the latent gene tree sequence. In practice, the choice of sufficiency measure determines the nature of the random tree $\boldsymbol{T} \sim Q^{[n]}_{\mu,p}(\cdot; \boldsymbol{t})$. We discuss aspects of the choice of sufficiency measure in Section 4.1.

**Remark 4.1.** Though in some sense arbitrary, the decision to study models of the form (4.3) is natural from the perspective of both theory and applications. Without such a condition, the conditional distributions $Q_p(\cdot; \boldsymbol{t})$ can exhibit complicated dependence on $\boldsymbol{t}$ that is difficult to study rigorously.

To specify a statistical model for the latent tree sequence, we must specify a probability distribution for every finite sample size $n = 1, 2, \ldots$, and so we define a family $\mu = (\mu_n(\cdot \mid \boldsymbol{t}),\ n \in \mathbb{N})$ of finite-dimensional sufficiency measures. To ensure a model $(Q_{\mu,p}(\cdot; \boldsymbol{t}),\ \boldsymbol{t} \in \mathcal{T}_{\mathbb{N}})$ for trees labeled by $\mathbb{N}$, the finite-dimensional sufficiency measures must exhibit *lack of interference*, i.e.

$$\mu_m(\pi' \mid \boldsymbol{R}_{m,n}\boldsymbol{t}) = \mu_n(\boldsymbol{D}^{-1}_{m,n}(\pi') \mid \boldsymbol{t}) \quad \text{for all } \pi' \in \mathcal{P}_{[m]} \text{ and } \boldsymbol{t} \in \mathcal{T}_{[n]},$$

for all $m \leq n$. Under this condition, the collection $(\mu_n(\cdot \mid \boldsymbol{t}),\ n \in \mathbb{N})$ of finite-dimensional sufficiency measures determines a unique sufficiency measure $\mu(\cdot \mid \boldsymbol{t})$ on infinite partitions of $\mathbb{N}$, given $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$.

**Hypothesis 4.1.** Throughout the paper, we assume that the sufficiency measure $\mu(\cdot \mid \boldsymbol{t})$, $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$, is *label equivariant*, i.e.

$$\mu(\mathrm{d}\pi \mid \boldsymbol{t}) = \mu(\mathrm{d}\pi^{\sigma} \mid \boldsymbol{t}^{\sigma}), \quad \pi \in \mathcal{P}_{\mathbb{N}}, \boldsymbol{t} \in \mathcal{T}_{\mathbb{N}},$$

for all permutations $\sigma \colon \mathbb{N} \to \mathbb{N}$ fixing all but finitely many elements of $\mathbb{N}$, and has the *lack of interference* property,

$$\mu(\boldsymbol{D}^{-1}_n(\pi) \mid \boldsymbol{t}) = \mu(\boldsymbol{D}^{-1}_n(\pi) \mid \boldsymbol{t}^*) \quad \text{for all } \boldsymbol{t}, \boldsymbol{t}^* \in \mathcal{T}_{\mathbb{N}} \text{ for which } \boldsymbol{R}_n\boldsymbol{t} = \boldsymbol{R}_n\boldsymbol{t}^*,$$

for all $\pi \in \mathcal{P}_{[n]}$, for all $n \in \mathbb{N}$.

Alternatively, we can view $Q^{[n]}_{\mu,p}(\cdot; \boldsymbol{t})$ as a transition probability for a Markov chain on $\mathcal{T}_{[n]}$. The collection $(Q^{[n]}_{\mu,p},\ n \in \mathbb{N})$ determines a unique transition probability measure on $\mathcal{T}_{\mathbb{N}}$ if and only if

$$Q^{[n]}_{\mu,p}(\boldsymbol{t}'; \boldsymbol{R}_{n,n+1}\boldsymbol{t}^*) = Q^{[n+1]}_{\mu,p}(\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}'); \boldsymbol{t}^*), \quad \text{for all } \boldsymbol{t}^* \in \mathcal{T}_{[n+1]}, \boldsymbol{t}' \in \mathcal{T}_{[n]}, \tag{4.4}$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

for all $n \geq 1$ [9]. In this case, (4.3) can be expressed as

$$Q_{\mu,p}(\mathrm{d}\boldsymbol{t}'; \boldsymbol{t}) := \int_{\mathcal{P}_{\mathbb{N}}} Q_p(\mathrm{d}\boldsymbol{t}'; \pi)\mu(\mathrm{d}\pi \mid \boldsymbol{t}), \qquad \boldsymbol{t}, \boldsymbol{t}' \in \mathcal{T}_{\mathbb{N}}, \tag{4.5}$$

which we call a *generalized Markov branching model*, or $(\mu, p)$*-Markov branching model*. We call any random tree $\boldsymbol{T}$ with distribution in (4.5) a *generalized Markov branching tree*.

In Theorem 4.2, we show that (4.4) holds for $(Q^{[n]}_{\mu,p},\ n \geq 1)$ if and only if its conditional splitting rules $p$ satisfy

$$p_n(\pi'; \pi) = p_{n+1}(\boldsymbol{D}^{-1}_{n,n+1}(\pi'); \pi^*) + p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)p_n(\pi'; \pi), \qquad \pi, \pi' \in \mathcal{P}_{[n]}, \tag{4.6}$$

for all $\pi^* \in \boldsymbol{D}^{-1}_{n,n+1}(\pi)$, for all $n \in \mathbb{N}$. Intuitively, (4.6) is the Markovian extension of condition (3.5) for ordinary splitting rules.

### 4.1. Choice of sufficiency measure

Suitable choices of the sufficiency measure vary by application. We mention only a few natural choices for illustration.

In general, as long as the sufficiency measure is nondegenerate, it incorporates an inherently Bayesian feature into the model. For example, since Kingman's coalescent is a natural objective prior for $\boldsymbol{t}$, we can append edge lengths to $\boldsymbol{t}$ according to the law of a coalescent tree conditioned to have topology $\boldsymbol{t}$. Given the resulting tree with edge lengths $\boldsymbol{t}^\circ$, we let $\Pi$ be the partition in $\boldsymbol{t}^\circ$ observed at some random time $T^*$. Alternatively, we can append discrete edge lengths to $\boldsymbol{t}$, as in [10]. Denoting this tree by $\boldsymbol{t}^\bullet$, we can sample the partition in $\boldsymbol{t}^\bullet$ at any random integer-valued time, as above, or we can pick a meaningful fixed time. For example, sampling $\boldsymbol{t}^\bullet$ at time 1 gives the root partition of $\boldsymbol{t}^\bullet$, which affords the model a straightforward interpretation in terms of higher-order taxonomy.

Because the choice of sufficiency measure depends heavily on the application, we always leave $\mu$ unspecified, subject to Hypothesis 4.1. We focus instead on the conditional splitting rules and their implications for the model.

### 4.2. Examples of conditional splitting rules

In Section 4.3 we describe a simple class of conditional splitting rules by exploiting their relationship to partition-valued Markov chains. We now give a more concrete example that relates to previous work on Markov branching trees.

Gibbs measures play a special role in the study of fragmentation processes [6], [24]. For ordinary Markov branching trees, a *Gibbs splitting rule* has the form

$$p_n(\pi) \propto a_{\#\pi} \prod_{b \in \pi} \psi_{\#b}, \qquad \pi \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\},$$

for nonnegative sequences $\{\psi_n,\ n \geq 1\}$ and $\{a_n,\ n \geq 1\}$, where $\#\pi$ denotes the number of blocks of $\pi$. Our general description of Markov branching trees and their generalization puts no restriction on the number of children of each parent, but, in practice, we commonly assume only one speciation event can occur at any given time, which restricts the number of children to exactly two. This leads to the study of *binary Gibbs fragmentation trees*, which McCullagh *et al.* [24] characterized by Aldous's beta splitting rules [2],

$$p^\beta_n(\pi) := \frac{\Gamma(\beta + \#b_1 + 1)\Gamma(\beta + \#b_2 + 1)}{\Gamma(2\beta + n + 2)Z(n)}, \qquad \pi = \{b_1, b_2\} \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\},$$

where $\Gamma(\cdot)$ is the gamma function, $Z(n)$ is a normalizing constant, and $-2 \leq \beta \leq \infty$.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

Extending these ideas, we consider *Gibbs conditional splitting rules* with the form

$$p_n(\pi'; \pi) \propto a_{\#\pi'} \prod_{b \in \pi \wedge \pi'} \psi_{\#b}, \qquad \pi \in \mathcal{P}_{[n]},\ \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\},$$

where $\pi \wedge \pi'$ denotes the meet of $\pi$ and $\pi'$ in the partition lattice. In Theorem 4.3 below, we study these *Gibbs conditional splitting rules* in further detail and show that

$$p_n^{\beta}(\pi'; \pi) = 2 \frac{\prod_{b \in \pi \wedge \pi'} \beta^{\uparrow \#b}}{\prod_{b \in \pi} (2\beta)^{\uparrow \#b} - 2 \prod_{b \in \pi} \beta^{\uparrow \#b}}, \qquad \pi \in \mathcal{P}_{[n]},\ \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\},\ \beta > 0, \tag{4.7}$$

where $\beta^{\uparrow j} := \beta(\beta + 1) \cdots (\beta + j - 1)$, characterizes the class of exchangeable, consistent, binary Gibbs conditional splitting rules for the generalized Markov branching model.

The family in (4.7) is the analog of Aldous's beta splitting model to generalized Markov branching trees. By construction (4.11) below, these conditional splitting rules arise from the transition probabilities of the Ewens Markov chain [12].

### 4.3. Induced conditional splitting rules

An easy and natural way to specify a family of conditional splitting rules $(p_n,\ n \geq 2)$ is by the kernel of an exchangeable, consistent Markov chain on $\mathcal{P}_{\mathbb{N}}$. Let $(P_n,\ n \geq 2)$ be the finite-dimensional transition laws of an exchangeable, consistent, nondegenerate Markov chain on $\mathcal{P}_{\mathbb{N}}$, that is, $P_n(\pi'; \pi)$ is the probability of a transition from $\pi$ to $\pi'$ in $\mathcal{P}_{[n]}$, for each $n \in \mathbb{N}$, and $(P_n,\ n \geq 2)$ satisfy

$$P_2(\mathbf{1}_{[2]}; \pi) < 1 \quad \text{for all } \pi \in \mathcal{P}_{[2]}, \tag{4.8}$$

$$P_m(\pi'; \boldsymbol{D}_{m,n}\pi^*) = P_n(\boldsymbol{D}_{m,n}^{-1}(\pi'); \pi^*), \qquad \pi' \in \mathcal{P}_{[m]} \quad \text{for every } \pi^* \in \mathcal{P}_{[n]}, \tag{4.9}$$

and

$$P_n(\pi'^{\sigma}; \pi^{\sigma}) = P_n(\pi'; \pi), \qquad \pi, \pi' \in \mathcal{P}_{[n]} \quad \text{for all permutations } \sigma : [n] \to [n]. \tag{4.10}$$

From any such family, we can define conditional splitting rules $p := (p_n,\ n \geq 2)$ by conditioning a draw from $P_n(\cdot; \pi)$ to have at least two blocks, i.e.

$$p_n(\pi'; \pi) := \frac{P_n(\pi'; \pi)}{1 - P_n(\mathbf{1}_{[n]}; \pi)}, \qquad \pi \in \mathcal{P}_{[n]},\ \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\}. \tag{4.11}$$

Note that (4.9) is just the requirement that $(P_n,\ n \geq 2)$ is consistent under subsampling and thus determines the transition probabilities of a Markov chain on $\mathcal{P}_{\mathbb{N}}$.

**Proposition 4.1.** *Let $(P_n,\ n \in \mathbb{N})$ be finite-dimensional transition probability measures that satisfy (4.8)–(4.10). Then the conditional splitting rules $(p_n,\ n \geq 2)$ defined in (4.11) satisfy (4.6).*

*Proof.* Together, (4.8) and (4.9) imply that

$$P_n(\mathbf{1}_{[n]}; \pi) \leq P_2(\mathbf{1}_{[2]}; \boldsymbol{D}_{2,n}\pi) < 1, \qquad \text{for every } n \in \mathbb{N} \text{ and } \pi \in \mathcal{P}_{[n]},$$

and $p_n(\cdot; \pi)$ defined in (4.11) is a probability measure on $\mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\}$. Condition (4.6) follows easily from (4.9), i.e.

$$\begin{aligned} &p_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi'); \pi^*) + p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*) p_n(\pi'; \boldsymbol{D}_{n,n+1}\pi^*) \\ &\quad = \frac{P_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi'); \pi^*)}{1 - P_{n+1}(\mathbf{1}_{[n+1]}; \pi^*)} + \frac{P_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*)}{1 - P_{n+1}(\mathbf{1}_{[n+1]}; \pi^*)} \frac{P_n(\pi'; \boldsymbol{D}_{n,n+1}\pi^*)}{1 - P_n(\mathbf{1}_{[n]}; \boldsymbol{D}_{n,n+1}\pi^*)} \end{aligned}$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

$$= \frac{P_n(\pi'; \boldsymbol{D}_{n,n+1}\pi^*)}{1 - P_n(\mathbf{1}_{[n]}; \boldsymbol{D}_{n,n+1}\pi^*)} \left[ \frac{1 - P_n(\mathbf{1}_{[n]}; \boldsymbol{D}_{n,n+1}\pi^*) + P_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*)}{1 - P_{n+1}(\mathbf{1}_{[n+1]}; \pi^*)} \right]$$
$$= p_n(\pi'; \boldsymbol{D}_{n,n+1}\pi^*),$$

for every $\pi^* \in \mathcal{P}_{[n+1]}$ and $\pi' \in \mathcal{P}_{[n]}$, for all $n \in \mathbb{N}$. □

Conditional splitting rules constructed by Markov chains on $\mathcal{P}_{\mathbb{N}}$ play a special role in our theory; see Theorems 4.1 and 5.2 and the discussion of root partitions in the next section.

### 4.4. Root partitions

The root partition of a phylogenetic tree is the initial branching of the most recent common ancestor. It reflects the coarsest classification of species according to higher-order taxonomy. Aldous *et al.* [5] previously studied the problem of inferring higher-order taxa. In our study of Markov branching trees, root partitions play an important role because they determine the law of the branching below each vertex in the tree. We can also associate the existence of a root partition with special properties of the corresponding generalized Markov branching tree model, as we see in Theorem 5.2.

Formally, we say that $\boldsymbol{t} = (\boldsymbol{t}_n,\ n \in \mathbb{N}) \in \mathcal{T}_{\mathbb{N}}$ possesses a *root partition* if there exists $N \geq 1$ for which $(\Pi_{\boldsymbol{t}_n},\ n \geq N)$ is a compatible sequence of finite partitions, i.e. $\boldsymbol{D}_{n,n+1}\Pi_{\boldsymbol{t}_{n+1}} = \Pi_{\boldsymbol{t}_n}$ for all $n \geq N$. In this case, we write $\Pi_{\boldsymbol{t}} := \lim_{n\to\infty} \Pi_{\boldsymbol{t}_n}$ to denote the root partition of $\boldsymbol{t}$. Every finite tree possesses a root partition, but an infinite tree need not.

**Example 4.1.** (*Nonexistence of a root partition.*) The *infinite comb* $\boldsymbol{c}$ is defined by the collection $\boldsymbol{c} := (\boldsymbol{c}_n)_{n\geq 2}$, where $\Pi_{\boldsymbol{c}_n} = \boldsymbol{e}_n^{(n)} := \{[n-1], n\}$ for every $n \geq 2$. In this case, the sequence of finite root partitions is $(\boldsymbol{e}_n^{(n)})_{n\geq 2}$, for which $\boldsymbol{D}_{m,n}\boldsymbol{e}_n^{(n)} = \mathbf{1}_{[m]} \neq \boldsymbol{e}_m^{(m)}$ for every $m < n$; hence, $\lim_{n\to\infty} \Pi_{\boldsymbol{c}_n}$ does not exist.

In the following lemma, we define, for any Boolean statement $S$,

$$1(S) := \begin{cases} 1, & S \text{ holds}, \\ 0 & \text{otherwise}. \end{cases}$$

**Lemma 4.1.** *A fragmentation* $\boldsymbol{t} = (\boldsymbol{t}_n,\ n \in \mathbb{N})$ *of* $\mathbb{N}$ *possesses a root partition if and only if* $\sum_{n=2}^{\infty} 1(\Pi_{\boldsymbol{t}_n} = \boldsymbol{e}_n^{(n)}) < \infty$.

*Proof.* For fixed $\boldsymbol{t} = (\boldsymbol{t}_n,\ n \in \mathbb{N}) \in \mathcal{T}_{\mathbb{N}}$ and any $n \geq 2$, $\Pi_{\boldsymbol{t}_n}$ and $\Pi_{\boldsymbol{t}_{n+1}}$ are compatible unless $\Pi_{\boldsymbol{t}_{n+1}} = \boldsymbol{e}_{n+1}^{(n+1)}$. By the projective construction of $\mathcal{P}_{\mathbb{N}}$, $\lim_{n\to\infty} \Pi_{\boldsymbol{t}_n}$ exists if and only if $(\Pi_{\boldsymbol{t}_n},\ n \geq N)$ is a compatible sequence of partitions for some $N < \infty$, in which case

$$\sum_{n=2}^{\infty} 1(\Pi_{\boldsymbol{t}_n} = \boldsymbol{e}_n^{(n)}) \leq N < \infty. \qquad \square$$

In the following theorem, let $Q_{\mu,p}$ be the transition law of a $(\mu, p)$-Markov branching model with conditional splitting rules $(p_n,\ n \geq 2)$ and sufficiency measure $\mu$ satisfying Hypothesis 4.1.

**Theorem 4.1.** *Let* $(p_n,\ n \geq 2)$ *be a family of conditional splitting rules satisfying (4.1) and (4.6). Then* $(p_n,\ n \geq 2)$ *corresponds to transition probabilities* $(P_n,\ n \in \mathbb{N})$ *satisfying (4.8), (4.9), and (4.10) on* $(\mathcal{P}_{[n]},\ n \in \mathbb{N})$*, through (4.11), if and only if*

- *for every* $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$, $Q_{\mu,p}(\cdot; \boldsymbol{t})$*-almost every* $\boldsymbol{t}' \in \mathcal{T}_{\mathbb{N}}$ *possesses a root partition and*

- *for every $n \geq 2$, there exists a function $\tau_\infty \colon \mathcal{P}_{\mathbb{N}} \to (0, 1]$ such that, for all $n \geq 2$,*

$$\tau_\infty(\pi^*)Z(n, \pi^*) = \tau_\infty(\pi^{**})Z(n, \pi^{**}), \tag{4.12}$$

*for all $\pi^*, \pi^{**} \in \mathcal{P}_{\mathbb{N}}$ such that $\boldsymbol{D}_n\pi^* = \boldsymbol{D}_n\pi^{**}$, where*

$$Z(n, \pi) := \prod_{j=1}^{\infty}(1 - p_{n+j}(\boldsymbol{e}_{n+j}^{(n+j)}; \boldsymbol{D}_{n+j}\pi)), \qquad \pi \in \mathcal{P}_{\mathbb{N}},\ n \geq 2.$$

*Proof.* (i) *In the 'only if' direction.* Assume that $(p_n,\ n \geq 2)$ is given by (4.11) for some family $(P_n,\ n \in \mathbb{N})$ satisfying (4.8), (4.9), and (4.10). Then (4.1) is plainly satisfied. With $\boldsymbol{e}_n^{(n)} := \{[n-1], n\}$, we also have

$$1 - p_n(\boldsymbol{e}_n^{(n)}; \pi) = \frac{1 - P_{n-1}(\mathbf{1}_{[n-1]}; \boldsymbol{D}_{n-1,n}\pi)}{1 - P_n(\mathbf{1}_{[n]}; \pi)}, \tag{4.13}$$

for every $\pi \in \mathcal{P}_{[n]}$, for all $n \geq 2$. By (4.9) and Kolmogorov's extension theorem, $(P_n,\ n \in \mathbb{N})$ is determined by a unique Markov kernel $P_\infty$ on $\mathcal{P}_{\mathbb{N}}$ such that

$$P_n(\pi'; \boldsymbol{D}_n\pi^*) = P_\infty(\boldsymbol{D}_n^{-1}(\pi'); \pi^*), \qquad \pi' \in \mathcal{P}_{[n]}, \tag{4.14}$$

for all $\pi^* \in \mathcal{P}_{\mathbb{N}}$, for all $n \in \mathbb{N}$. Together with (4.8), (4.14) implies that

$$P_\infty(\mathbf{1}_{\mathbb{N}}; \pi) \leq P_2(\mathbf{1}_{[2]}; \boldsymbol{D}_2\pi) < 1 \quad \text{for every } \pi \in \mathcal{P}_{\mathbb{N}}.$$

We define

$$\tau_\infty(\pi) := 1 - P_\infty(\mathbf{1}_{\mathbb{N}}; \pi), \qquad \pi \in \mathcal{P}_{\mathbb{N}}, \tag{4.15}$$

which is everywhere positive. For every $\pi \in \mathcal{P}_{\mathbb{N}}$, (4.13) and (4.15) imply

$$\begin{aligned} Z(n, \pi) &:= \prod_{j=1}^{\infty}(1 - p_{n+j}(\boldsymbol{e}_{n+j}^{(n+j)}; \boldsymbol{D}_{n+j}\pi)) \\ &= \frac{1 - P_n(\mathbf{1}_{[n]}; \boldsymbol{D}_n\pi)}{1 - P_\infty(\mathbf{1}_{\mathbb{N}}; \pi)} \\ &= \frac{1 - P_n(\mathbf{1}_{[n]}; \boldsymbol{D}_n\pi)}{\tau_\infty(\pi)} \\ &> 0; \end{aligned}$$

and we observe (4.12).

We now take any $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$ and fix $\pi \in \mathcal{P}_{\mathbb{N}}$. For every $n \in \mathbb{N}$, (4.8) implies the existence of some $\pi_{**} \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\}$ for which $p_n(\pi_{**}; \boldsymbol{D}_n\pi) > 0$; whence,

$$\mathbb{P}\{\Pi_{\boldsymbol{T}'} \in \boldsymbol{D}_n^{-1}(\pi_{**})\} = p_n(\pi_{**}; \boldsymbol{D}_n\pi)Z(n, \pi) = \frac{P_n(\pi_{**}; \boldsymbol{D}_n\pi)}{\tau_\infty(\pi)} > 0.$$

But $\{\Pi_{\boldsymbol{T}'}$ exists$\}$ is a tail event with respect to the $\sigma$-field generated by the independent sequence $(\mathbf{1}\{\Pi_{\boldsymbol{T}'_{|[n]}} = \boldsymbol{e}_n^{(n)}\})_{n \geq 1}$. By Kolmogorov's 0-1 law together with

$$\mathbb{P}(\{\Pi_{\boldsymbol{T}'} \text{ exists}\}) = \mathbb{P}\bigg(\bigcup_{n \in \mathbb{N}} \bigcup_{\pi' \neq \mathbf{1}_{[n]}} \{\Pi_{\boldsymbol{T}'} \in \boldsymbol{D}_n^{-1}(\pi')\}\bigg) \geq \mathbb{P}\{\Pi_{\boldsymbol{T}'} \in \boldsymbol{D}_n^{-1}(\pi_{**})\} > 0,$$

$\Pi_{\boldsymbol{T}'}$ exists $Q_{\mu,p}(\cdot\,; \boldsymbol{t})$-almost surely.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

(ii) *In the 'if' direction.* Suppose that (4.12) holds and, for every $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$, $Q_{\mu,p}(\cdot; \boldsymbol{t})$-almost every $\boldsymbol{t}' \in \mathcal{T}_{\mathbb{N}}$ has a root partition. Fix $\boldsymbol{t} \in \mathcal{T}_{\mathbb{N}}$ and $\tau_\infty \colon \mathcal{P}_{\mathbb{N}} \to (0, 1]$ subject to these assumptions, and let $\boldsymbol{T}' \sim Q_{\mu,p}(\cdot; \boldsymbol{t})$. For each $n = 2, 3, \ldots$, define $E_n := \{\Pi_{\boldsymbol{R}_n \boldsymbol{T}'} = \boldsymbol{e}_n^{(n)}\}$, the event that the root partition of the reduced subtree $\boldsymbol{R}_n \boldsymbol{T}'$ is $\boldsymbol{e}_n^{(n)}$. (Note that the occurrence of $E_n$ signifies the regeneration of the root partition at stage $n$; that is, on $E_n$, the root partition of $\boldsymbol{T}'$ (if it exists) cannot be in the set $\boldsymbol{D}_n^{-1}(\mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\})$.) By Lemma 4.1, $\sum_{n\ge 2} 1(E_n) < \infty$ almost surely and

$$\infty > \mathbb{E}\biggl(\sum_{n\ge 2} 1(E_n)\biggr) = \sum_{n\ge 2} \mathbb{P}(E_n). \tag{4.16}$$

For every $\pi \in \mathcal{P}_{\mathbb{N}}$, the sequence $(Z(n+1, \pi),\ n \in \mathbb{N})$ is monotonically increasing and bounded above by 1. (In fact, $\lim_{n\to\infty} Z(n, \pi) = 1$.) It remains to show that $Z(n, \pi) > 0$ for all $n \ge 2$. For fixed $\pi \in \mathcal{P}_{\mathbb{N}}$,

$$0 < Z(2, \pi) = \exp\biggl\{\sum_{n=2}^{\infty} \log(1 - \mathbb{P}(E_n))\biggr\} \quad \iff \quad \sum_{n=2}^{\infty} \log(1 - \mathbb{P}(E_n)) > -\infty.$$

From (4.16), $\mathbb{P}(E_n) \to 0$ as $n \to \infty$ and, by Taylor's theorem, we can write

$$\log(1 - \mathbb{P}(E_n)) = -\mathbb{P}(E_n) + o(\mathbb{P}(E_n)) \quad \text{as } n \to \infty.$$

We conclude that

$$\sum_{n=2}^{\infty} \log(1 - \mathbb{P}(E_n)) \ge -\sum_{n=2}^{N} \mathbb{P}(E_n) + o(\mathbb{P}(E_n)) > -\infty. \tag{4.17}$$

Conditions (4.12) and (4.17) allow us to define $\tau_n \colon \mathcal{P}_{[n]} \to [0, 1]$ for each $n \in \mathbb{N}$: for every $\pi \in \mathcal{P}_{[n]}$, we can choose any $\pi^* \in \boldsymbol{D}_n^{-1}(\pi)$ and set

$$\tau_n(\pi) := \tau_\infty(\pi^*) Z(n, \pi^*) > 0.$$

For every $n \ge 2$ and $\pi \in \mathcal{P}_{[n+1]}$, the family $(\tau_n,\ n \ge 2)$ satisfies

$$\tau_n(\boldsymbol{D}_{n,n+1}\pi) = \tau_{n+1}(\pi)(1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi)).$$

From $(\tau_n,\ n \ge 2)$ and the conditional splitting rules $(p_n,\ n \ge 2)$, we define Markov kernels $(P_n,\ n \in \mathbb{N})$ on $(\mathcal{P}_{[n]},\ n \in \mathbb{N})$ by $P_1 \equiv 1$ and, for $n \ge 2$,

$$P_n(\pi'; \pi) := \begin{cases} \tau_n(\pi) p_n(\pi'; \pi), & \pi' \ne \mathbf{1}_{[n]}, \\ 1 - \tau_n(\pi), & \pi' = \mathbf{1}_{[n]}, \end{cases} \quad \text{for each } \pi \in \mathcal{P}_{[n]}.$$

Directly from its definition, $(P_n,\ n \in \mathbb{N})$ satisfies (4.8) and is exchangeable. To verify (4.9), we fix $n \in \mathbb{N}$ and $\pi^* \in \mathcal{P}_{[n+1]}$: for $\pi' \ne \mathbf{1}_{[n]}$,

$$\begin{aligned} P_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi'); \pi^*) &= \tau_{n+1}(\pi^*) p_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi'); \pi^*) \\ &= \tau_n(\boldsymbol{D}_{n,n+1}\pi^*) \frac{p_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi'); \pi^*)}{1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*)} \\ &= P_n(\pi'; \boldsymbol{D}_{n,n+1}\pi^*); \end{aligned}$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

otherwise,

$$P_{n+1}(\boldsymbol{D}^{-1}_{n,n+1}(\boldsymbol{1}_{[n]});\pi^*) = 1-\tau_{n+1}(\pi^*)(1-p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1};\pi^*)) = P_n(\boldsymbol{1}_{[n]};\boldsymbol{D}_{n,n+1}\pi^*).$$

Finally, the conditional splitting rules defined from $(P_n,\ n\in\mathbb{N})$ through (4.11) coincide with the conditional splitting rules $(p_n,\ n\geq 2)$, completing the proof. □

Theorem 4.1 comes into play in Section 5.4 when we consider attaching discrete edge lengths to generalized Markov branching trees. In that case, we establish a correspondence between generalized Markov branching trees with integer-valued edge lengths and exchangeable, consistent Markov chains on $\mathscr{P}_{\mathbb{N}}$.

## 4.5. Characterization of exchangeable generalized Markov branching trees

**Theorem 4.2.** *Let $\mu=(\mu_n,\ n\in\mathbb{N})$ be any sufficiency measure satisfying Hypothesis 4.1. Then an exchangeable collection $(Q^{[n]}_{\mu,p},\ n\in\mathbb{N})$ of generalized Markov branching distributions determines the law of an exchangeable generalized Markov branching tree on $\mathscr{T}_{\mathbb{N}}$ with distribution as in (4.5) if and only if $p$ satisfies (4.1) and (4.6).*

*Proof of the 'if' statement.* Suppose $p$ satisfies (4.1) and (4.6). Then exchangeability of $Q^{[n]}_{\mu,p}$, for each $n\in\mathbb{N}$, is clear. To establish consistency, we must verify (4.4). Let $\boldsymbol{t}_2$ denote the unique element of $\mathscr{T}_{[2]}$ and $\pi_2=\{\{1\},\{2\}\}=\frac{1}{2}$ denote the unique element of $\mathscr{P}_{[2]}\setminus\{\boldsymbol{1}_{[2]}\}$. For any $\pi\in\mathscr{P}_{[2]}$, we observe that

$$Q^{[2]}_{\mu,p}(\boldsymbol{t}_2;\boldsymbol{t}_2)=p_2(\pi_2;\pi)=1.$$

By (4.6), we have

$$\begin{aligned}1&=Q^{[3]}_{\mu,p}(\mathscr{T}_{[3]};\pi)\\&=Q^{[3]}_{\mu,p}(\boldsymbol{R}^{-1}_{2,3}(\boldsymbol{t}_2);\pi)\\&=p_3(\boldsymbol{D}^{-1}_{2,3}(\tfrac{1}{2});\pi)+p_3(\tfrac{12}{3};\pi)p_2(\tfrac{1}{2};\pi_{|[2]})\\&=p_2(\pi_2;\pi)\quad\text{ for any }\pi\in\mathscr{P}_{[3]}.\end{aligned}$$

Hence, $Q^{[2]}_{\mu,p}(\cdot;\pi_{|[2]})$ and $Q^{[3]}_{\mu,p}(\cdot;\pi)$ are consistent for all $\pi\in\mathscr{P}_{[3]}$.

For our induction hypothesis, we assume that

$$Q^{[n]}_p(\boldsymbol{t}';\boldsymbol{D}_{n,n+1}\pi^*)=Q^{[n+1]}_p(\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}');\pi^*)\quad\text{ for all }\pi^*\in\mathscr{P}_{[n+1]},\ \boldsymbol{t}'\in\mathscr{T}_{[n]},$$

holds for some $n\geq 2$. We fix $\boldsymbol{t},\boldsymbol{t}'\in\mathscr{T}_{[n]}$ and $\boldsymbol{t}^*\in\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t})$. For $\boldsymbol{t}''\in\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')$, we write $b^*\in\Pi_{\boldsymbol{t}'}$ to denote the block of $\Pi_{\boldsymbol{t}'}$ to which the element $n+1$ must be added to obtain $\Pi_{\boldsymbol{t}''}$, and we write $b^{**}\in\Pi_{\boldsymbol{t}''}$ to denote the block of $\Pi_{\boldsymbol{t}''}$ containing $n+1$. We also assume that $\pi^*\in\mathscr{P}_{[n+1]}$ and $\pi:=\pi^*_{|[n]}$ are the restrictions of an infinite partition drawn from $\mu(\cdot\mid\boldsymbol{t})$. Below, we use the recursive expression

$$Q^{[n]}_{\mu,p}(\boldsymbol{t}';\boldsymbol{t})=\sum_{\pi\in\mathscr{P}_{[n]}}p_n(\Pi_{\boldsymbol{t}'};\pi)\prod_{b\in\Pi_{\boldsymbol{t}'}}Q^b_p(\boldsymbol{t}'_{|b};\pi_{|b})\mu_n(\pi\mid\boldsymbol{t}),\qquad \boldsymbol{t},\boldsymbol{t}'\in\mathscr{T}_{[n]},$$

and the decomposition

$$\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')=\left\{\bigcup_{\{\pi''\in\boldsymbol{D}^{-1}_{n,n+1}(\Pi_{\boldsymbol{t}'})\}}\ \bigcup_{\{\boldsymbol{t}''\in\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')\colon\ \Pi_{\boldsymbol{t}''}=\pi''\}}\{\boldsymbol{t}''\}\right\}\cup\{\boldsymbol{t}_*\},\tag{4.18}$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

where $\boldsymbol{t}_*$ is the unique element of $\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')$ with $\Pi_{\boldsymbol{t}_*} = \boldsymbol{e}^{(n+1)}_{n+1}$. (Note that (4.18) is a disjoint union obtained by partitioning the elements of $\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')$ according to their root partition.) Conditional on $\pi^*$, the induction hypothesis implies that

$$
\begin{aligned}
&Q^{[n+1]}_p(\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}');\pi^*)\\
&\quad= \sum_{\boldsymbol{t}''\in\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')} Q^{[n+1]}_p(\boldsymbol{t}'';\pi^*)\\
&\quad= \sum_{\boldsymbol{t}''\in\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')} p_{n+1}(\Pi_{\boldsymbol{t}''};\pi^*)\prod_{b\in\Pi_{\boldsymbol{t}''}} Q^b_p(\boldsymbol{t}''_{|b};\pi^*_{|b})\\
&\quad= \sum_{\pi''\in\boldsymbol{D}^{-1}_{n,n+1}(\Pi_{\boldsymbol{t}'})}\ \sum_{\boldsymbol{t}''\in\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}')\colon \Pi_{\boldsymbol{t}''}=\pi''} p_{n+1}(\pi'';\pi^*)\\
&\qquad\times\prod_{b\in\pi''} Q^b_p(\boldsymbol{t}''_{|b};\pi^*_{|b}) + p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1};\pi^*)Q^{[n]}_p(\boldsymbol{t}';\pi)\\
&\quad= \sum_{\pi''\in\boldsymbol{D}^{-1}_{n,n+1}(\Pi_{\boldsymbol{t}'})} p_{n+1}(\pi'';\pi^*)\left[\prod_{b\neq b^*} Q^b_p(\boldsymbol{t}''_{|b};\pi^*_{|b})\right]\\
&\qquad\times\sum_{\boldsymbol{t}''\in\mathscr{A}} Q^{b^{**}}_p(\boldsymbol{t}'';\pi^*_{|b^{**}}) + p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1};\pi^*)Q^{[n]}_p(\boldsymbol{t}';\pi) &&(4.19)\\
&\quad= \sum_{\pi''\in\boldsymbol{D}^{-1}_{n,n+1}(\Pi_{\boldsymbol{t}'})} p_{n+1}(\pi'';\pi^*)\prod_{b\in\Pi_{\boldsymbol{t}'}} Q^b_p(\boldsymbol{t}''_{|b};\pi^*_{|b}) + p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1};\pi^*)Q^{[n]}_p(\boldsymbol{t}';\pi)\\
&\quad= \prod_{b\in\Pi_{\boldsymbol{t}'}} Q^b_p(\boldsymbol{t}'_{|b};\pi^*_{|b})[p_{n+1}(\boldsymbol{D}^{-1}_{n,n+1}(\Pi_{\boldsymbol{t}'});\pi^*) + p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1};\pi^*)p_n(\Pi_{\boldsymbol{t}'};\pi)] &&(4.20)\\
&\quad= p_n(\Pi_{\boldsymbol{t}'};\pi)\prod_{b\in\Pi_{\boldsymbol{t}'}} Q^b_p(\boldsymbol{t}'_{|b};\pi_{|b})\\
&\quad= Q^{[n]}_p(\boldsymbol{t}';\pi).
\end{aligned}
$$

The sum in (4.19) is over the set $\mathscr{A} := \boldsymbol{R}^{-1}_{b^*,b^{**}}(\boldsymbol{t}'_{|b^*})$ of all fragmentations $\boldsymbol{t}''$ of $b^{**}$ whose restriction to $b^*$ is $\boldsymbol{t}'_{|b^*}$, which is equal to $Q^{b^*}_p(\boldsymbol{t}'_{|b^*};\pi^*_{|b^*})$ by the induction hypothesis. By induction, we conclude that

$$
Q^{[n]}_p(\boldsymbol{t};\pi) = Q^{[n+1]}_p(\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t});\pi^*) \quad \text{for all } \pi^*\in\boldsymbol{D}^{-1}_{n,n+1}(\pi),
$$

for every $n \geq 1$. Hence, there exists a conditional distribution $Q_p(\cdot;\pi)$ on $\mathscr{T}_{\mathbb{N}}$ for every $\pi\in\mathscr{P}_{\mathbb{N}}$. By the lack of interference property of $\mu$, we conclude (4.4) for $(Q^{[n]}_{\mu,p},\ n\geq 1)$.

Conversely, if (4.4) holds for $(Q^{[n]}_{\mu,p},\ n\geq 1)$, then the above sequence of calculations is valid through line (4.20) by taking $\mu_n(\cdot \mid \boldsymbol{t})$ to be degenerate at $\pi^*\in\mathscr{P}_{[n+1]}$. If either $Q^{[n]}_p(\boldsymbol{t}';\pi) > 0$ or $Q^{[n]}_p(\boldsymbol{t}';\pi) = 0$ but $Q^b_p(\boldsymbol{t}'_{|b};\pi^*_{|b}) > 0$ for all $b\in\Pi_{\boldsymbol{t}'}$, then

$$
\begin{aligned}
&\prod_{b\in\Pi_{\boldsymbol{t}'}} Q^b_p(\boldsymbol{t}'_{|b};\pi_{|b})[p_{n+1}(\boldsymbol{D}^{-1}_{n,n+1}(\Pi_{\boldsymbol{t}'});\pi^*) + p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1};\pi^*)p_n(\Pi_{\boldsymbol{t}'};\pi)]\\
&\qquad= Q^{[n+1]}_p(\boldsymbol{R}^{-1}_{n,n+1}(\boldsymbol{t}');\pi^*)
\end{aligned}
$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

$$\begin{aligned} &= Q_p^{[n]}(\boldsymbol{t}'; \pi) \\ &= p_n(\Pi_{\boldsymbol{t}'}; \pi) \prod_{b \in \Pi_{\boldsymbol{t}'}} Q_p^b(\boldsymbol{t}'_{|b}; \pi_{|b}), \end{aligned}$$

which implies (4.6). Since, for any $\pi' \in \mathcal{P}_{[n]}$ with $p_n(\pi'; \pi) = 0$, we can always choose a collection of subtrees $\{\boldsymbol{t}'_b \colon b \in \pi'\}$ such that $Q_p^b(\boldsymbol{t}'_{|b}; \pi_{|b}) > 0$ for every $b \in \Pi_{\boldsymbol{t}'}$, these two cases suffice. □

## 4.6. Conditional Gibbs splitting rules

The conditional beta splitting rule in Section 4.2 specializes the more general class of *conditional Gibbs fragmentation processes*. For fixed $n \geq 2$, we say that a conditional splitting rule $p_n$ on $\mathcal{P}_{[n]} \setminus \{\mathbf{1}_n\}$ is of *Gibbs type* if, for each $\pi \in \mathcal{P}_{[n]}$,

$$p_n(\pi'; \pi) = \frac{a(\#\pi')}{Z(\pi)} \prod_{b \in \pi \wedge \pi'} \psi(\#b), \qquad \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_n\}, \tag{4.21}$$

for functions $a, \psi \colon \mathbb{N} \to \mathbb{R}^+$ and $Z \colon \mathcal{P}_{[n]} \to \mathbb{R}^+$. For binary models, we can put $a(2) = 1$ and $a(j) = 0$ for $j \neq 2$. From (4.21), conditional Gibbs rules are exchangeable, but a family $(p_n)_{n \geq 2}$ defined from the same $a$, $\psi$ need not satisfy the consistency condition (4.6).

For fixed $k \geq 2$, $-1 < \beta < \infty$, and $\pi \in \mathcal{P}_{[n]}$, we define the *conditional Dirichlet splitting rule*

$$p_n(\pi'; \pi) = \frac{k^{\downarrow \#\pi'}}{Z_n(\pi)} \prod_{b \in \pi} \frac{\prod_{b' \in \pi'} \Gamma(\beta + \#(b \cap b') + 1)/\Gamma(\beta + 1)}{\Gamma(k\beta + \#b + k)/\Gamma(k\beta + k)}, \qquad \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_n\}, \tag{4.22}$$

where $k^{\downarrow j} = k(k-1)\cdots(k-j+1)$ and

$$Z_n(\pi) = 1 - k \prod_{b \in \pi} \frac{\Gamma(\beta + \#b + 1)/\Gamma(\beta + 1)}{\Gamma(k\beta + \#b + k)/\Gamma(k\beta + k)}. \tag{4.23}$$

The conditional beta splitting rule (4.7) is the special case of (4.22) with $k = 2$.

**Theorem 4.3.** *Let $(Q^{[n]}(\cdot, \cdot))_{n \in \mathbb{N}}$ be a family of exchangeable, consistent Gibbs transition probabilities. Then its conditional Gibbs splitting rules $(p_n)_{n \geq 2}$ are given by*

$$p_n(\pi'; \pi) = \frac{k^{\downarrow \#\pi'}}{Z_n(\pi)} \prod_{b \in \pi} \frac{\prod_{b' \in \pi'} (\beta + 1)^{\uparrow \#(b \cap b')}}{(k\beta + k)^{\uparrow \#b}}, \qquad \pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\},$$

*for some $k < \infty$ and $\beta > -1$, or $p_n$ is the limit as $\beta \to \infty$, where $Z_n(\pi)$ is the normalizing constant in (4.23).*

*Proof.* For convenience, we write (4.21) in the equivalent form

$$p_n(\pi'; \pi) = \frac{a(\#\pi')}{Z(\pi)} \prod_{b \in \pi} \prod_{b' \in \pi'} \psi(\#(b \cap b')) \quad \text{for } \psi(0) = 1, \tag{4.24}$$

and we write $\psi(\pi) = \prod_{b \in \pi} \psi(\#b)$. Let $\pi \in \mathcal{P}_{[n]}$, $\pi' \in \mathcal{P}_{[n]} \setminus \{\mathbf{1}_n\}$, and $\pi^* \in \boldsymbol{D}^{-1}_{n,n+1}(\pi)$.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

The Gibbs form (4.21) and consistency (4.6) imply

$$\begin{aligned}\frac{a(\#\pi')}{Z(\pi)}\psi(\pi\wedge\pi')\bigg[1-\frac{a(2)}{Z(\pi^*)}\psi(\pi)\psi(1)\bigg]\\ =\frac{a(\#\pi'+1)}{Z(\pi^*)}\psi(\pi\wedge\pi')\psi(1)+\sum_{b\in\pi'}\frac{a(\#\pi')}{Z(\pi^*)}\psi(\pi^*\wedge\pi_b''),\end{aligned}$$

where $\pi_b''$ is obtained by inserting element $n+1$ in block $b\in\pi'$; whence, for any $\pi'$ in the support of $p_n(\cdot;\pi)$,

$$C(\pi^*):=\frac{Z(\pi^*)-a(2)\psi(\pi)\psi(1)}{Z(\pi)}=A(\#\pi')\psi(1)+\sum_{b'\in\pi'}\Psi(\#(b'\cap b^*)),$$

where $\Psi(i):=\psi(i+1)/\psi(i)$, $A(r):=a(r+1)/a(r)$, and $b^*\in\pi$ is the block into which $n+1$ is inserted to obtain $\pi^*$. The left-hand side depends only on $\pi^*$, which implies that $\Psi(i+1)-\Psi(i)$ is constant for all $i\geq 0$, and we may write $\Psi(i+1)-\Psi(i)=\beta$. For the moment, assume that $\Psi(i)=\alpha+\beta i$ with $\alpha,\beta>0$. We have

$$\psi(i)=\Psi(i-1)\Psi(i-2)\dots\Psi(0)=\prod_{j=0}^{i-1}(\alpha+\beta j)=\beta^i\frac{\Gamma(i+\alpha/\beta)}{\Gamma(\alpha/\beta)}.$$

There is no loss of generality in putting $\beta=1$ and $\gamma=\alpha/\beta>0$ to obtain $\psi(i)=\Gamma(i+\gamma)/\Gamma(\gamma)$, giving

$$\begin{gathered}C(\pi^*)=A(r)\psi(1)+\sum_{b'\in\pi'}\Psi(\#(b^*\cap b'))=\gamma A(r)+\gamma r+\#b^*,\\ r\leq\inf\{m\in\mathbb{N}\colon a(m+1)=0\};\end{gathered}$$

whence, $A(r):=\delta-r$. Without loss of generality, we assign $a(1)=0$ and $a(2)=\delta(\delta-1)$ so that

$$a(r)=a(2)\prod_{j=2}^{r-1}A(j)=\delta(\delta-1)\prod_{j=2}^{r-1}(\delta-j)=\delta^{\downarrow r}.$$

As $A(r)$ will be negative for large enough $r$, we must have $\#\pi'<k$ for some $k<\infty$ and $a(k+1)=0$ implies $\delta=k$.

In this case, the conditional Gibbs splitting rule is

$$p_n(\pi';\pi)=\frac{k^{\downarrow\#\pi'}}{Z(\pi)}\prod_{b\in\pi}\prod_{b'\in\pi'}\gamma^{\uparrow\#(b\cap b')}.$$

For $\alpha\in\mathbb{R}$, the $\alpha$-permanent of an $n\times n$ real-valued matrix $M$ is defined by

$$\operatorname{per}_\alpha M:=\sum_{\sigma\in\mathfrak{S}_n}\alpha^{\#\sigma}\prod_{j=1}^{n}M_{j,\sigma(j)},$$

where $\mathfrak{S}_n$ is the symmetric group of permutations acting on $[n]$ and $\#\sigma$ denotes the number of cycles of $\sigma\in\mathfrak{S}_n$. In [11] we have shown that

$$\operatorname{per}_\alpha M=\sum_{\pi\in\mathcal{P}_{[n]}}k^{\downarrow\#\pi}\prod_{b\in\pi}\operatorname{per}_{\alpha/k}M[b],$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

where $M[b]$ is the submatrix of $M$ with rows and columns indexed by $b$. For a partition $\pi \in \mathscr{P}_{[n]}$, we can write $\mathrm{per}_\alpha \pi = \prod_{b\in\pi} \alpha^{\uparrow\#b}$ by regarding $\pi$ as a 0-1 valued matrix with $(i, j)$ entry 1 if $i$ and $j$ are in the same block of $\pi$, and 0 otherwise. Hence,

$$\begin{aligned} Z(\pi) &= \sum_{\pi'\in\mathscr{P}_{[n]}\setminus\{\mathbf{1}_n\}} k^{\downarrow\#\pi'} \prod_{b\in\pi\wedge\pi'} \gamma^{\uparrow\#b} \\ &= \sum_{\pi'\in\mathscr{P}_{[n]}} k^{\downarrow\#\pi'} \prod_{b\in\pi\wedge\pi'} \mathrm{per}_\gamma \mathbf{1}_{b'} - k\prod_{b\in\pi} \gamma^{\uparrow\#b} \\ &= \mathrm{per}_{k\gamma}\,\pi - k\,\mathrm{per}_\gamma\,\pi. \end{aligned}$$

We have

$$p_n(\pi';\pi) = \frac{k^{\downarrow\#\pi'}}{Z^*(\pi)} \prod_{b\in\pi} \frac{\prod_{b'\in\pi'} \gamma^{\uparrow\#(b\cap b')}}{(k\gamma)^{\uparrow\#b}}, \tag{4.25}$$

where

$$Z^*(\pi) := \frac{\mathrm{per}_{k\gamma}\,\pi - k\,\mathrm{per}_\gamma\,\pi}{\mathrm{per}_{k\gamma}\,\pi}.$$

The following cases arise.

- $k < \infty$ and $0 < \gamma < \infty$. In this case, (4.25) is exactly (4.22) with $\beta = \gamma - 1$, i.e.

  $$p_n(\pi';\pi) = \frac{1}{Z^*(\pi)} \frac{k^{\downarrow\#\pi'}}{k^{\#\pi}} I_{\{\pi'\wedge\pi=\pi\}},$$

  a discrete-time coalescent chain conditioned not to transition to the trivial state $\mathbf{1}_n$, where $I_{\{\cdot\}}$ is the indicator function.

- $k < \infty$ and $\gamma = \infty$. Then $\gamma = \infty$ corresponds to $\beta = 0$, for which $\psi(i) = \alpha^i$ and $C(\pi^*) = \alpha A(r) + \alpha r'$; hence, $A(r') + r' = \delta > 0$ and $A(r') = \delta - r$ as above. In this case,

  $$p_n(\pi';\pi) = \frac{1}{Z(\pi)} \frac{k^{\downarrow\#\pi'}}{k^n}$$

  is the splitting rule obtained by conditioning a $k$-coupon collector partition to be nontrivial, the limit of (4.22) as $\beta \to \infty$.

In the Gibbs case, $\gamma = 0$ is not possible because it corresponds to $\Psi(0) = 0$. The case $k = \infty$ and $0 < \gamma \le \infty$ is prohibited as it corresponds to $p_n(\pi';\pi) = I_{\{\pi'=\mathbf{0}_n\}}$, the deterministic split into singletons, which is not of Gibbs type (4.21). □

**Corollary 4.1.** *The conditional Dirichlet splitting models (4.22) with $-1 < \beta \le \infty$ are the only consistent conditional Gibbs models.*

By taking limits, we can extend the parameter range of (4.25) to include the cases $k \in \mathbb{N}\cup\{\infty\}$ and $0 \le \gamma \le \infty$, but not all limiting cases have Gibbs type.

**Proposition 4.2.** *In considering limits in (4.25) as both $\gamma \downarrow 0$ and $k \uparrow \infty$, the following cases arise. Fix $n \in \mathbb{N}$, $\pi \in \mathscr{P}_{[n]}$, and $\pi' \in \mathscr{P}_{[n]} \setminus \{\mathbf{1}_n\}$.*

- *$k < \infty$ and $0 < \gamma < \infty$. In this case, (4.25) is exactly (4.22) with $\beta = \gamma - 1$.*

- *$k < \infty$ and $\gamma = \infty$. We have*
$$p_n(\pi'; \pi) = \frac{1}{Z(\pi)} \frac{k^{\downarrow \#\pi'}}{k^n},$$
*the $k$-coupon collector law conditioned to be nontrivial, also the limit of (4.22) as $\beta \to \infty$.*
- *$k \uparrow \infty$, $\gamma \downarrow 0$, and $\gamma k \to \theta \in (0, \infty)$. Equation (4.25) converges to*
$$p_n(\pi'; \pi, \theta) = \frac{\theta^{\#(\pi \wedge \pi')}}{Z^*(\pi)} \prod_{b \in \pi} \frac{\prod_{b' \in \pi'} \Gamma(\#(b \cap b'))}{\theta^{\uparrow \# b}} I_{\{\pi \wedge \pi' = \pi'\}},$$
*the splitting rule obtained by fragmenting the blocks of $\pi$ independently according to the Ewens distribution with parameter $\theta > 0$.*
- *$k = \infty$ and $0 < \gamma \leq \infty$. We have $p_n(\pi'; \pi) = I_{\{\pi' = \mathbf{0}_n\}}$, the deterministic split into singletons.*

**Remark 4.2.** Note that $\pi = \mathbf{1}_n$ translates each of the above cases to its corresponding unconditional Gibbs splitting rule.

## 5. Attaching edge lengths to generalized Markov branching trees

### 5.1. Weighted fragmentation trees

By associating each edge of $\boldsymbol{t}_A \in \mathcal{T}_A$ with a real number in $[0, \infty]$, we construct a tree with edge lengths. We can let edge lengths be real-valued, which corresponds to continuous-time evolution, or integer-valued, which corresponds to nonoverlapping generations in population genetics modeling. Appending edge lengths to fragmentation trees introduces further technical points. In the interest of space, we keep technicalities to a minimum.

**Definition 5.1.** A *weighted fragmentation* $\boldsymbol{t}^\circ$ of $A \subset_f \mathbb{N}$ is a pair $(\boldsymbol{t}, \boldsymbol{w})$ such that $\boldsymbol{t} \in \mathcal{T}_A$ and $\boldsymbol{w} = \{w_b,\ b \subseteq A\}$, with

(W1) $w_b \in (0, \infty]$ for all $b \in \boldsymbol{t}$,

(W2) $w_b = 0$ if and only if $b \notin \boldsymbol{t}$, and

(W3) $w_b = \infty$ if and only if $b$ is a singleton or the empty set.

For $b \in \boldsymbol{t}$, we call $w_b > 0$ the *weight*, or *length*, of $b$. We write $\mathcal{T}_A^\circ$ to denote the space of weighted trees with most recent common ancestor $A$. We call $\boldsymbol{t}^\circ$ a *discrete-weighted tree* if, in addition to (W2) and (W3),

(D) $w_b \in \{1, 2, \dots, \infty\}$ for all $b \in \boldsymbol{t}$.

We write $\mathcal{T}_A^\bullet \subset \mathcal{T}_A^\circ$ to denote the space of discrete weighted trees with MRCA $A$.

**Remark 5.1.** A weighted tree is a fragmentation tree with weights assigned to each of its edges. The infinite weight assigned to singletons reflects the fact that singletons undergo no further splitting and, thus, 'live forever'. We need not require that the weight associated to every singleton be infinite; however, these are the only trees that can arise as a result of the projection operation defined in (5.1) below.

In Figure 5 we give a pictorial representation of the tree $\boldsymbol{t}_5$ in Figure 2 with edge lengths $w_{\{1,2,3,4,5\}} = 1$, $w_{\{1,2\}} = 3$, and $w_{\{3,4\}} = 2$ attached.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

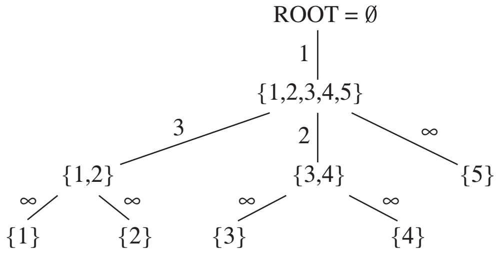


FIGURE 5.

Analogous to the unweighted case, we define the restriction of $\boldsymbol{t}^\circ \in \mathcal{T}_A$ by removing elements and elongating branches as appropriate. Formally, for $A' \subseteq A$, we define the restriction of $\boldsymbol{t}^\circ := (\boldsymbol{t}, \boldsymbol{w}) \in \mathcal{T}^\circ_A$ to $\mathcal{T}^\circ_{A'}$ by $\boldsymbol{R}^\circ_{A',A}\boldsymbol{t}^\circ = \boldsymbol{t}^\circ_{|A'} := (\boldsymbol{R}_{A',A}\boldsymbol{t}, \boldsymbol{w}')$, where $\boldsymbol{R}_{A,A'}$ is defined in (3.2) and $\boldsymbol{w}' := \{w'_b,\ b \subseteq A'\}$ is given by

$$w'_b := \sum_{\{b' \in \boldsymbol{t}\colon b' \cap A' = b\}} w_{b'}, \qquad b \subseteq A'. \tag{5.1}$$

We write $\mathcal{T}^\circ_{\mathbb{N}}$ to denote the space of weighted $\mathbb{N}$-trees, which consists of collections $(\boldsymbol{t}^\circ_n,\ n \in \mathbb{N})$ with $\boldsymbol{t}^\circ_n \in \mathcal{T}^\circ_{[n]}$ and $\boldsymbol{t}^\circ_m = \boldsymbol{R}^\circ_{m,n}\boldsymbol{t}^\circ_n$ for all $m \le n$, for every $n \in \mathbb{N}$. As usual, we equip $\mathcal{T}^\circ_{\mathbb{N}}$ and $\mathcal{T}^\bullet_{\mathbb{N}}$ with the smallest $\sigma$-field so that these restriction maps are measurable.

## 5.2. Attaching edge lengths to generalized Markov branching trees

Previous authors have considered the task of attaching both continuous, exponentially distributed [21], [24] and discrete, geometrically distributed [10] edge lengths to Markov branching trees. We now consider the analogous task for generalized Markov branching trees. As before, we let $p = (p_n,\ n \ge 2)$ be an exchangeable, consistent conditional splitting rule and $\mu$ be a sufficiency measure satisfying Hypothesis 4.1. We could allow the sufficiency measure to depend on a weighted fragmentation $\boldsymbol{t}^\circ = (\boldsymbol{t}, \boldsymbol{w}) \in \mathcal{T}_{\mathbb{N}}$, in which case we modify Hypothesis 4.1 accordingly. This generalization is straightforward to carry out and does not affect our conclusions, so we omit it.

Given $\boldsymbol{t}^\circ := (\boldsymbol{t}, \boldsymbol{w}) \in \mathcal{T}^\circ_{\mathbb{N}}$, we define a distribution on $\mathcal{T}^\circ_{\mathbb{N}}$ as follows.

- First, generate $\pi \sim \mu(\cdot \mid \boldsymbol{t}^\circ)$.
- Given $\pi$, we generate $\boldsymbol{T}' \sim Q_p(\cdot; \pi)$.
- Given $\boldsymbol{T}' = \boldsymbol{t}'$ and $\pi$, assign edge lengths $W_b$ to each $b \in \boldsymbol{t}'$ conditionally independently according to a distribution $F_b(\cdot; \pi)$ on $[0, \infty]$.

To maintain the generalized Markov branching property, the edge length distribution must be memoryless, which immediately restricts $F_b$ to either the exponential or geometric distribution. In the next two sections, we obtain necessary and sufficient conditions for the edge length distributions $(F_b)_{b \subset_f \mathbb{N}}$, given a family of conditional splitting rules.

## 5.3. Generalized Markov branching trees with continuous edge weights

Let $(p_n,\ n \ge 2)$ be the conditional splitting rules of a $(\mu, p)$-Markov branching model on $\mathcal{T}_{\mathbb{N}}$ and let $(\lambda_n,\ n \ge 0)$ be a collection of exponential rate functions, $\lambda_n\colon \mathcal{P}_{[n]} \to [0, \infty)$ for each $n = 0, 1, \ldots$, such that $\lambda_0 = \lambda_1 = 0$ and

$$\lambda_n(\pi) = \lambda_n(\pi^\sigma), \qquad \pi \in \mathcal{P}_{[n]} \quad \text{for all permutations } \sigma\colon [n] \to [n],\ n \ge 2. \tag{5.2}$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

Given $\boldsymbol{t}^\circ := (\boldsymbol{t}, \boldsymbol{w}) \in \mathcal{T}^\circ_{[n]}$, we define $Q^{[n]\circ}_{\mu,p,\lambda}(\cdot; \boldsymbol{t}^\circ)$ as the law of $\boldsymbol{T}'^\circ$ generated as follows.

(C1) We take $\pi \sim \mu(\cdot \mid \boldsymbol{t}^\circ)$ and, given $\pi$, $\boldsymbol{T}' \sim Q^{[n]}_p(\cdot; \pi)$.

(C2) Given $\boldsymbol{T}' = \boldsymbol{t}'$, $\pi$, and $\boldsymbol{t}^\circ$, we generate $\boldsymbol{W}' := \{W'_b,\ b \in \boldsymbol{t}'\}$, a mutually independent collection of exponential random variables, where $W'_b$ is exponentially distributed with rate parameter $\lambda_b(\pi_{|b})$, for each $b \in \boldsymbol{t}'$.

(C3) We put $\boldsymbol{T}'^\circ := (\boldsymbol{t}', \boldsymbol{w}')$, where $\boldsymbol{w}' := \{W'_b,\ b \subseteq [n]\}$ is defined from $\{W'_b,\ b \in \boldsymbol{t}'\}$ in (C2) and $W'_b \equiv 0$ for every $b \notin \boldsymbol{t}'$.

**Theorem 5.1.** *Let $p := (p_n,\ n \geq 2)$ be a family of conditional splitting rules satisfying (4.1) and (4.6) and let $\mu$ satisfy Hypothesis 4.1. The following are equivalent.*

(i) *There exists a collection $\lambda := (\lambda_n,\ n \geq 0)$ of exponential rate functions such that $(Q^{\circ[n]}_{\mu,p,\lambda},\ n \in \mathbb{N})$ are the finite-dimensional distributions of an exchangeable generalized Markov branching model on $\mathcal{T}^\circ_{\mathbb{N}}$.*

(ii) *The family $\lambda := (\lambda_n,\ n \in \mathbb{N})$ of exponential rate functions satisfies $\lambda_0 = \lambda_1 = 0$, (5.2), and*

$$\lambda_n(\boldsymbol{D}_{n,n+1}\pi^*) = \lambda_{n+1}(\pi^*)(1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)) \tag{5.3}$$

*for every $\pi^* \in \mathcal{P}_{[n+1]}$, for all $n \geq 1$.*

(iii) *The families $p = (p_n,\ n \geq 2)$ and $\lambda = (\lambda_n,\ n \geq 1)$ are determined by a family of exchangeable Markovian transition rates $(q_n,\ n \in \mathbb{N})$ on $(\mathcal{P}_{[n]},\ n \in \mathbb{N})$ satisfying (4.9) and*

$$q_n(\mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\}; \pi) < \infty \quad \textit{for every } \pi \in \mathcal{P}_{[n]}, \textit{ for all } n \in \mathbb{N}. \tag{5.4}$$

*In this case, we have $\lambda_n(\pi) = q_n(\mathcal{P}_{[n]} \setminus \{\mathbf{1}_{[n]}\}; \pi)$ and $p_n(\pi'; \pi) = q_n(\pi'; \pi)/\lambda_n(\pi)$, $\pi' \neq \mathbf{1}_{[n]}$; moreover, the finite-dimensional conditional densities $(Q^{[n]\circ}_{\mu,p,\lambda},\ n \in \mathbb{N})$ associated to $Q^\circ_{\mu,p,\lambda}$ are*

$$\begin{aligned} &Q^{[n]\circ}_{\mu,p,\lambda}(\mathrm{d}\boldsymbol{t}'^\circ; \boldsymbol{t}^\circ) \\ &\quad = \int_{\mathcal{P}_{\mathbb{N}}} \prod_{\{b \in \boldsymbol{t}' :\ \#b \geq 2\}} q_b(\mathrm{CH}_{\boldsymbol{t}'}(b); \pi_{|b}) \mathrm{e}^{\{-w'_b q_b(\mathcal{P}_b \setminus \{\mathbf{1}_b\}; \pi_{|b})\}} \mathrm{d}w'_b \mu(\mathrm{d}\pi \mid \boldsymbol{t}^\circ) \end{aligned} \tag{5.5}$$

*for every $\boldsymbol{t}^\circ$, $\boldsymbol{t}'^\circ = (\boldsymbol{t}', w') \in \mathcal{T}^\circ_{[n]}$.*

*Proof.* Throughout this proof, we assume that $\boldsymbol{T}'^\circ \sim Q^\circ_{\mu,p,\lambda}(\cdot; \boldsymbol{t}^\circ)$, for some $\boldsymbol{t}^\circ \in \mathcal{T}^\circ_{\mathbb{N}}$, wherever it appears.

(i) $\Longleftrightarrow$ (ii). From Theorem 4.2, the induced kernel $Q_{\mu,p}$ on $\mathcal{T}_{\mathbb{N}}$ is exchangeable and satisfies (4.4) if and only if $p$ satisfies (4.1) and (4.6). To establish the consistency of $(Q^{[n]\circ}_{\mu,p,\lambda},\ n \geq 1)$, the independent random edge weights $(W_b,\ b \in \boldsymbol{T}')$ must satisfy

$$W_b \overset{\mathrm{D}}{=} W_{b\cup\{n+1\}} + W_b I_{E_{b\cup\{n+1\}}} \quad \text{for all } b \subset_f \mathbb{N},$$

where '$\overset{\mathrm{D}}{=}$' denotes *equality in law*, $E_{b\cup\{n+1\}} := \{\Pi_{\boldsymbol{T}'_{|b\cup\{n+1\}}} = \{b, \{n+1\}\}\}$. By exchangeability of $Q_{\mu,p}$, it suffices that

$$W_n \overset{\mathrm{D}}{=} W_{n+1} + W_n I_{n+1} \quad \text{for every } \pi^* \in \mathcal{P}_{[n+1]}, \text{ for all } n \in \mathbb{N}, \tag{5.6}$$

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

where $\{W_n, W_{n+1}, I_{n+1}\}$ are mutually independent, $W_n \sim \operatorname{Exp}(\lambda_n(\boldsymbol{D}_{n,n+1}\pi^*))$, $W_{n+1} \sim \operatorname{Exp}(\lambda_{n+1}(\pi^*))$, and $I_{n+1} \sim \operatorname{Bern}(p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*))$.
The characteristic function of $X \sim \operatorname{Exp}(\lambda)$ is

$$\psi_X(t) := \mathbb{E}\mathrm{e}^{\mathrm{i}tX} = \frac{\lambda}{\lambda - \mathrm{i}t}.$$

Fix $\pi^* \in \mathcal{P}_{[n+1]}$ and put $\pi = \boldsymbol{D}_{n,n+1}\pi^*$. Then

$$\begin{aligned}
&\mathbb{E}(\mathrm{e}^{\mathrm{i}t(W_{n+1}+W_nI_{n+1})}) \\
&\quad = \frac{\lambda_{n+1}(\pi^*)}{\lambda_{n+1}(\pi^*) - \mathrm{i}t}\biggl[\frac{\lambda_n(\pi)}{\lambda_n(\pi) - \mathrm{i}t}p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*) + 1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*)\biggr] \\
&\quad = \frac{\lambda_n(\pi)}{\lambda_n(\pi) - \mathrm{i}t}\biggl[\frac{\lambda_{n+1}(\pi^*)}{\lambda_{n+1}(\pi^*) - \mathrm{i}t}\,\frac{\lambda_n(\pi) - \mathrm{i}t(1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*))}{\lambda_n(\pi)}\biggr]. \qquad (5.7)
\end{aligned}$$

If (5.3), then (5.7) simplifies to $\lambda_n(\pi)/(\lambda_n(\pi) - \mathrm{i}t)$, the characteristic function of $W_n \sim \operatorname{Exp}(\lambda_n(\pi))$. On the other hand, if (5.7) is equal to $\psi_{W_n}(t)$, then

$$\frac{\lambda_{n+1}(\pi^*)}{\lambda_{n+1}(\pi^*) - \mathrm{i}t} = \frac{\lambda_n(\pi)}{\lambda_n(\pi) - \mathrm{i}t(1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*))}.$$

We have defined $\lambda_0 = \lambda_1 = 0$ so that (5.6) is obvious for $n = 1$. Definition 5.1 forces $\lambda_n(\cdot) > 0$ for all $n \geq 2$; so, for every $\pi^* \in \mathcal{P}_{[n+1]}$, there is a unique $\alpha_{\pi^*} = \alpha > 0$ such that $\alpha\lambda_n(\boldsymbol{D}_{n,n+1}\pi^*) = \lambda_{n+1}(\pi^*)$. We have

$$\frac{\alpha}{\alpha}\,\frac{\lambda_n(\pi)}{\lambda_n(\pi) - \mathrm{i}t(1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*))} = \frac{\lambda_{n+1}(\pi^*)}{\lambda_{n+1}(\pi^*) - \alpha\mathrm{i}t(1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*))},$$

which simplifies to (5.3). Since this relation must hold for all $\boldsymbol{t}^\circ \in \mathcal{T}_{\mathbb{N}}^\circ$, it must be satisfied for all $\pi^* \in \mathcal{P}_{[n+1]}$.

(ii) $\Longleftrightarrow$ (iii) Assume (ii). For every $n \in \mathbb{N}$, we define, for each $\pi \in \mathcal{P}_{[n]}$,

$$Q_n(\pi'; \pi) = \begin{cases} \lambda_n(\pi)p_n(\pi'; \pi), & \pi' \neq \mathbf{1}_{[n]}, \\ -\lambda_n(\pi), & \pi' = \mathbf{1}_{[n]}\,. \end{cases}$$

Finiteness of $\lambda_n$ for each $n \in \mathbb{N}$ implies (5.4). Also, (4.6) and (5.3) imply (4.9): for $\pi' \neq \mathbf{1}_{[n]}$,

$$\begin{aligned}
\lambda_{n+1}(\pi^*)p_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\pi'); \pi^*) &= \lambda_{n+1}(\pi^*)(1 - p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*))p_n(\pi'; \pi) \\
&= \lambda_n(\pi)p_n(\pi'; \pi);
\end{aligned}$$

otherwise, for $\pi' = \mathbf{1}_{[n]}$,

$$Q_{n+1}(\boldsymbol{D}_{n,n+1}^{-1}(\mathbf{1}_{[n]}); \pi^*) = -\lambda_{n+1}(\pi^*) + \lambda_{n+1}(\pi^*)p_{n+1}(\boldsymbol{e}_{n+1}^{(n+1)}; \pi^*) = -\lambda_n(\pi).$$

Thus, $(Q_n,\ n \in \mathbb{N})$ satisfies (4.9) and (5.4). The finite-dimensional transition density (5.5) follows by our procedure for embedding in $\mathcal{T}_{[n]}^\circ$ through (C1)–(C3) preceding the theorem. The converse is immediate.

This completes the proof. □

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

**Corollary 5.1.** *Any exchangeable generalized Markov branching tree on $\mathcal{T}_{\mathbb{N}}$ can be consistently embedded in $\mathcal{T}^{\circ}_{\mathbb{N}}$.*

*Proof.* Setting $\lambda_1(\{1\}) = 0$, $\lambda_2(\pi) = 1$ for all $\pi \in \mathcal{P}_{[2]}$, and defining $\lambda_n$ recursively from (5.3) completes the proof. □

### 5.4. Generalized Markov branching trees with discrete edge weights

Let $(p_n,\ n \geq 2)$ be the conditional splitting rules associated to an exchangeable Markov branching model on $\mathcal{T}_{\mathbb{N}}$ and let $(\tau_n,\ n \geq 0)$ be a collection of geometric success functions, $\tau_n \colon \mathcal{P}_{[n]} \to [0, 1]$, satisfying $\tau_0 = \tau_1 = 0$ and (5.2). Given $\boldsymbol{t}^\bullet := (\boldsymbol{t}, \boldsymbol{w}) \in \mathcal{T}^\bullet_{[n]}$, we define $Q^{[n]\bullet}_{\mu,p,\tau}(\cdot; \boldsymbol{t}^\bullet)$ as the law of $\boldsymbol{T}'^\bullet$ generated in three steps (G1), (G2), (G3), where (G1) and (G3) coincide with (C1) and (C3), respectively, and

(G2) given $\boldsymbol{T}' = \boldsymbol{t}'$, $\pi$, and $\boldsymbol{t}^\bullet$, we generate $\{W_b,\ b \in \boldsymbol{t}'\}$ to be a mutually independent collection of geometric random variables, where $W_b$ has the geometric distribution with success probability $\tau_b(\pi_{|b})$, for each $b \in \boldsymbol{t}'$.

**Theorem 5.2.** *Let $p := (p_n,\ n \geq 2)$ be a family of conditional splitting rules satisfying (4.1) and (4.6) and let $\mu$ be a sufficiency measure on $\mathcal{P}_{\mathbb{N}}$ satisfying Hypothesis 4.1. The following are equivalent.*

(i) *There exists a collection $\tau := (\tau_n,\ n \geq 0)$ of geometric success functions such that $(Q^{[n]\bullet}_{\mu,p,\tau},\ n \in \mathbb{N})$ are the finite-dimensional distributions of an exchangeable generalized Markov branching model on $\mathcal{T}^\bullet_{\mathbb{N}}$.*

(ii) *The family $\tau := (\tau_n,\ n \geq 0)$ satisfies $\tau_0 = \tau_1 = 0$, (5.2), and, for each $\pi \in \mathcal{P}_{[n]}$,*

$$\tau_n(\pi) = \tau_{n+1}(\pi^*)(1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)) \quad \textit{for all } \pi^* \in \boldsymbol{D}^{-1}_{n,n+1}(\pi), \textit{ for all } n \geq 2. \tag{5.8}$$

(iii) *The families $(p_n,\ n \geq 2)$ and $(\tau_n,\ n \geq 0)$ are associated, through (4.11), to some family $(P_n,\ n \in \mathbb{N})$ of Markovian transition probabilities on $(\mathcal{P}_{[n]},\ n \in \mathbb{N})$ satisfying (4.8) and (4.9), and $\tau_n(\pi) = 1 - P_n(\mathbf{1}_{[n]}; \pi)$, $\pi \in \mathcal{P}_{[n]}$, for every $n \in \mathbb{N}$. Moreover, the finite-dimensional conditional distributions $(Q^{[n]\bullet}_{\mu,p,\tau},\ n \in \mathbb{N})$ associated to $Q^\bullet_{\mu,p,\tau}$ are*

$$Q^{[n]\bullet}_{\mu,p,\tau}(\boldsymbol{t}'^\bullet; \boldsymbol{t}^\bullet) = \int_{\mathcal{P}_{\mathbb{N}}} \prod_{\{b \in \boldsymbol{t}' \colon \#b \geq 2\}} P_b(\mathrm{CH}_{\boldsymbol{t}'}(b); \pi_{|b}) P_b(\mathbf{1}_b; \pi_{|b})^{w'_b - 1} \mu(\mathrm{d}\pi \mid \boldsymbol{t}^\bullet), \tag{5.9}$$
$$\boldsymbol{t}^\bullet, \boldsymbol{t}'^\bullet := (\boldsymbol{t}', \boldsymbol{w}') \in \mathcal{T}^\bullet_{[n]}.$$

(iv) *$Q^\bullet_{\mu,p,\tau}(\cdot; \boldsymbol{t}^\bullet)$-almost every $\boldsymbol{t}'^\bullet \in \mathcal{T}^\bullet_{\mathbb{N}}$ possesses a root partition, for every $\boldsymbol{t}^\bullet \in \mathcal{T}^\bullet_{\mathbb{N}}$.*

(v) *$\lambda_\infty(\pi) := \lim_{n\to\infty} \lambda_n(\pi_{|[n]}) < \infty$ for all $\pi \in \mathcal{P}_{\mathbb{N}} \setminus \{\mathbf{1}_{\mathbb{N}}\}$, where $\lambda := (\lambda_n,\ n \geq 2)$ is defined recursively by $\lambda_2 \equiv 1$ and*

$$\lambda_{n+1}(\pi^*) := \frac{\lambda_n(\pi^*_{|[n]})}{(1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))}, \quad \textit{for every } \pi^* \in \mathcal{P}_{[n+1]},\ n \geq 2.$$

*Proof.* The argument closely follows the proof of Theorem 5.1. Only condition (5.3) and (iii) differ slightly.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

(i) $\Longleftrightarrow$ (ii). As in Theorem 5.1, consistency of $(Q^{[n]\bullet}_{\mu,p,\tau},\ n \geq 1)$ holds if and only if

$$W_n \overset{\mathrm{D}}{=} W_{n+1} + W_n I_{n+1} \quad \text{for all } \boldsymbol{t}^\circ \in \mathcal{T}^\circ_{\mathbb{N}}, \tag{5.10}$$

where $\{W_n, W_{n+1}, I_{n+1}\}$ are mutually independent, $W_n \sim \mathrm{geom}(\tau^*_n(\pi))$, $W_{n+1} \sim \mathrm{geom}(\tau^*_{n+1}(\pi^*))$, and $I_{n+1} \sim \mathrm{Bern}(p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))$, for $\pi^* \in \mathcal{P}_{[n+1]}$ and $\pi := \boldsymbol{D}_{n,n+1}\pi^*$.

The probability generating function of $X \sim \mathrm{geom}(p)$, $0 < p < 1$, is

$$G_X(s) := \mathbb{E}s^X = \frac{ps}{1-(1-p)s}, \qquad s < \frac{1}{1-p}.$$

We have defined $\tau_1 \equiv 0$ so that (5.10) is trivial for $n = 1$. By Definition 5.1, $0 < \tau_n \leq 1$ for all $n \geq 2$. Letting $\pi$, $\pi^*$, and $I_{n+1}$ be as in (5.10) and writing $\sigma_n(\cdot) := 1 - \tau_n(\cdot)$, we have

$$\begin{aligned}
&\mathbb{E}(s^{W_{n+1}+W_n I_{n+1}})\\
&\quad = \mathbb{E}(s^{W_{n+1}})\mathbb{E}(s^{W_n I_{n+1}})\\
&\quad = \frac{s\tau_{n+1}(\pi^*)}{1-\sigma_{n+1}(\pi^*)s}\biggl[\frac{s\tau_n(\pi)p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)}{1-s\sigma_n(\pi)} + 1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)\biggr]\\
&\quad = \frac{s\tau_n(\pi)}{1-s\sigma_n(\pi)}\frac{s\tau_{n+1}(\pi^*)}{1-s\sigma_{n+1}(\pi^*)}\\
&\qquad \times \frac{1}{s\tau_n(\pi)}[s\tau_n(\pi)p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*) + 1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)\\
&\qquad\qquad - s\sigma_n(\pi)(1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))]\\
&\quad = \frac{s\tau_n(\pi)}{1-s\sigma_n(\pi)}\frac{s\tau_{n+1}(\pi^*)}{1-s\sigma_{n+1}(\pi^*)}\frac{s\tau_n(\pi) + (1-s)(1-p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))}{s\tau_n(\pi)}.
\end{aligned}$$

Hence, (5.8) implies (5.10). On the other hand, if (5.10) then

$$\frac{s\tau_{n+1}(\pi^*)}{1-s\sigma_{n+1}(\pi^*)} = \frac{s\tau_n(\pi)}{s\tau_n(\pi) + (1-s)(1-p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))}.$$

As $\tau_n(\pi)$ and $\tau_{n+1}(\pi^*)$ are both strictly positive, there is a unique $\alpha_{\pi^*} = \alpha > 0$ such that $\alpha\tau_n(\pi) = \tau_{n+1}(\pi^*)$. The above expression simplifies to

$$\begin{aligned}
\frac{s\tau_{n+1}(\pi^*)}{1-s\sigma_{n+1}(\pi^*)} &= \frac{\alpha}{\alpha}\frac{s\tau_n(\pi)}{s\tau_n(\pi) + (1-s)(1-p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))}\\
&= \frac{s\tau_{n+1}(\pi^*)}{s\tau_{n+1}(\pi^*) + (1-s)\alpha(1-p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*))},
\end{aligned}$$

for which we need

$$\alpha(1 - p_{n+1}(\boldsymbol{e}^{(n+1)}_{n+1}; \pi^*)) = 1,$$

establishing (5.8).

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

(ii) $\Longleftrightarrow$ (iii). Assuming (ii), we can define $P_n : \mathcal{P}_{[n]} \times \mathcal{P}_{[n]} \to [0, 1]$, $n \in \mathbb{N}$, by

$$P_n(\pi'; \pi) = \begin{cases} p_n(\pi'; \pi)\tau_n(\pi), & \pi' \neq \mathbf{1}_{[n]}, \\ 1 - \tau_n(\pi), & \pi' = \mathbf{1}_{[n]}. \end{cases}$$

The finite-dimensional transition law (5.9) follows from the construction of $Q^{\bullet}_{\mu,p,\tau}(\cdot; \boldsymbol{t}^{\bullet})$ in (G1)–(G3). Corollary 4.1 gives the converse.

(iii) $\Longleftrightarrow$ (iv). This follows from Theorem 4.1.

(ii) $\Longleftrightarrow$ (v). This is clear since (5.8) implies the functions $\lambda_n$ constructed from (5.3) converge to a finite value for each $\pi \in \mathcal{P}_{\mathbb{N}}$.

This completes the proof. □

**Remark 5.2.** Note that the essential difference between Theorems 5.1 and 5.2 occurs in parts (iv) and (v) in Theorem 5.2. Whereas trees with continuous edge lengths need not have a well-defined root partition, those with discrete edge lengths do.

## Acknowledgements

This work was partially supported by NSF grants CNS-1523785 and CAREER DMS-1554092.

## References

[1] Abraham, R., Delmas, J.-F. and He, H. (2012). Pruning Galton–Watson trees and tree-valued Markov processes. *Ann. Inst. H. Poincaré Prob. Statist.* **48,** 688–705.

[2] Aldous, D. (1996). Probability distributions on cladograms. In *Random Discrete Structures* (IMA Vol. Math. Appl. **76**), Springer, New York, pp. 1–18.

[3] Aldous, D. (1998). Tree-valued Markov chains and Poisson Galton–Watson distributions. In *Microsurveys in Discrete Probability* (DIMACS Ser. Discrete Math. Theoret. Comput. Sci. **41**), American Mathematical Society, Providence, RI, pp. 1–20.

[4] Aldous, D. and Pitman, J. (1998). Tree-valued Markov chains derived from Galton–Watson processes. *Ann. Inst. H. Poincaré Statist.* **34,** 637–686.

[5] Aldous, D., Krikun, M. and Popovic, L. (2008). Stochastic models for phylogenetic trees on higher-order taxa. *J. Math. Biol.* **56,** 525–557.

[6] Berestycki, N. and Pitman, J. (2007). Gibbs distributions for random partitions generated by a fragmentation process. *J. Statist. Phys.* **127,** 381–418.

[7] Bertoin, J. (1996). *Lévy Processes*. Cambridge University Press.

[8] Bertoin, J. (2006). *Random Fragmentation and Coagulation Processes* (Camb. Stud. Adv. Math. **102**). Cambridge University Press.

[9] Burke, C. J. and Rosenblatt, M. (1958). A Markovian function of a Markov chain. *Ann. Math. Statist.* **29,** 1112–1122.

[10] Crane, H. (2013). Consistent Markov branching trees with discrete edge lengths. *Electron. Commun. Prob.* **18,** 14pp.

[11] Crane, H. (2013). Some algebraic identities for the $\alpha$-permanent. *Linear Algebra Appl.* **439,** 3445–3459.

[12] Crane, H. (2014). The cut-and-paste process. *Ann. Prob.* **42,** 1952–1979.

[13] Crane, H. (2015). Clustering from categorical data sequences. *J. Amer. Statist. Assoc.* **110,** 810–823.

[14] Crane, H. (2015). Generalized Ewens–Pitman model for Bayesian clustering. *Biometrika* **102,** 231–238.

[15] Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.* **31,** 1–19. (Rejoinder: **31,** 37–39.)

[16] Evans, S. N. (2008). *Probability and Real Trees* (Lecture Notes Math. **1920**). Springer, Berlin.

[17] Evans, S. N. and Winter, A. (2006). Subtree prune and regraft: a reversible real tree-valued Markov process. *Ann. Prob.* **34,** 918–961.

[18] Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3,** 87–112.

[19] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer, Sunderland.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press

[20] Haas, B. and Miermont, G. (2012). Scaling limits of Markov branching trees with applications to Galton–Watson and random unordered trees. *Ann. Prob.* **40,** 2589–2666.

[21] Haas, B., Miermont, G., Pitman, J. and Winkel, M. (2008). Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. *Ann. Prob.* **36,** 1790–1837.

[22] Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoret. Pop. Biol.* **23,** 183–201.

[23] Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

[24] McCullagh, P., Pitman, J. and Winkel, M. (2008). Gibbs fragmentation trees. *Bernoulli* **14,** 988–1002.

[25] McVean, G. and Cardin, N. (2005). Approximating the coalescent with recombination. *Phil. Trans. R. Soc. London* **360,** 1387–1393.

[26] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Prob. Theory Relat. Fields* **102,** 145–158.

[27] Pitman, J. (2006). *Combinatorial Stochastic Processes* (Lecture Notes Math. **1875**). Springer, Berlin.

[28] Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105,** 437–460.

[29] Wang, Y. *et al.* (2014). A new method for modeling coalescent processes with recombination. *BMC Bioinformatics* **15,** 12pp.

https://doi.org/10.1017/apr.2016.81 Published online by Cambridge University Press