

A unified multi-resolution coalescent: Markov lumpings of the Kingman-Tajima n -coalescent

Raazesh Sainudiin* and Tanja Stadler†

Biomathematics Research Centre*, University of Canterbury,
Private Bag 4800, Christchurch, New Zealand
`r.sainudiin@math.canterbury.ac.nz` and

ETH Zürich, Institut f. Integrative Biologie†,
CHN H 72, Universitätstrasse 16, 8092 Zürich, Switzerland
`tanja.stadler@env.ethz.ch`

UCDMS 2009/4. Some rights reserved.



This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 New Zealand Licence.

To view a copy of this license, visit
<http://creativecommons.org/licenses/by-nc-sa/3.0/nz/>.

April 5, 2009
Version 2 Revised on October 12, 2009

Abstract

In this paper, we formulate six different resolutions of a continuous-time approximation of the Wright-Fisher sample genealogical process. We derive Markov chains for the six different approximations in the spirit of J.F.C. Kingman. These Markov chains are essential for inference methods. Two of the resolutions are the well-known n -coalescent and the lineage death process due to Kingman. Two other resolutions were mentioned by Kingman and Tajima, but never explicitly formalized. Another two resolutions are novel, and embed the genealogical objects of Kingman and Tajima into a general framework via the theory of lumped Markov chains. We show that any sample genealogical Markov chain is amenable to Kingman's n -coalescent approximation if it has the lineage death chain as its lumped Markov chain. We formulate a lumped n -coalescents graph that embodies multiple n -coalescent resolutions of the underlying sample genealogical process and leads to computationally efficient inference.

1 Introduction

Kingman's n -coalescent [15, 14] is a process of central importance in mathematical population genetics. The n -coalescent is a continuous-time Markov chain formulation for a limiting approximation of the genealogical history of a labeled sample of size n from a Wright-Fisher population [7, 32] of a large and constant size N . The state space of the n -coalescent is \mathbb{C}_n , the set of all set partitions of the label set $\mathcal{L} = \{1, 2, \dots, n\}$. At time zero, the n integer labels in \mathcal{L} are all in separate "blocks." As time t advances into the past, each transition that allows two blocks merging into a single block happens at rate one, until the process reaches a terminal state in which all integer labels are together. If one considers just the discrete skeleton or the embedded jump chain of this Markov chain, then at each time step one picks two blocks of the partition at random and merges them together, until there is just a single block after $n - 1$ time steps.

In this paper, we consider six variants or genealogical resolutions of this coalescent process. They are briefly introduced below.

- The *vintaged and labeled* n -coalescent $\{B^\dagger(t)\}$ of § 3.1 is the same as the process described above except that, at all times, each block of the partition has an associated number called the *vintage*, which records the time step or coalescent epoch in which the block was created. Its state space \mathbb{B}_n is an augmentation of \mathbb{C}_n with coalescent vintage tags. This is the Kingman-Tajima n -coalescent.

- The *unvintaged and labeled n -coalescent* $\{C^\uparrow(t)\}$ of § 3.2 is obtained from $\{B^\uparrow(t)\}$ by dropping the vintage. This is the standard Kingman's n -coalescent. Every sequence of states in \mathbb{C}_n that is visited by this process is an element of \mathcal{C}_n , the set of n -coalescent sequences or c -sequences. A c -sequence induces a *ranked, rooted, binary tree* with leaves labeled by \mathcal{L} as defined in 2.1 and depicted in Figure 3.
- The *vintaged and sized n -coalescent* $\{D^\uparrow(t)\}$ of § 3.3 is obtained from $\{B^\uparrow(t)\}$ by keeping track only of the vintage and the size of each block of the partition, and dropping the integer labels $1, 2, \dots, n$. Its state space \mathbb{D}_n is an ordered integer partition.
- The *vintaged and shaped n -coalescent* $\{G^\uparrow(t)\}$ of § 3.4 is obtained from $\{D^\uparrow(t)\}$ by keeping track only of the vintages of the blocks at each time step, and throwing away the sizes of the blocks. The state space \mathbb{G}_n is contained in the vertices of the hypercube $\{0, 1\}^{n-1}$. The sequence of states visited by this process gives Tajima's *evolutionary relationships* [28, Figures 1-4], which resolve genealogical histories up to *ranked, rooted, binary tree shapes* as defined in 2.1 and depicted in Figure 3. This is Tajima's n -coalescent.
- The *unvintaged and sized n -coalescent* $\{F^\uparrow(t)\}$ of § 3.5 is obtained from $\{C^\uparrow(t)\}$ by just keeping track of how many blocks there are of each size. This process is also known as the *label-killed n -coalescent* [15, (5.2)] or *unlabeled n -coalescent* [23] or *family-size process* [13, 30, p. 136-137] on \mathbb{F}_n , the integer partitions of n .
- The *pure death process* $\{H^\uparrow(t)\}$ of § 2.2.2 is obtained from any of the other five processes above by just keeping track of the number of blocks or the number of ancestral sample lineages in $\mathbb{H}_n = \{n, n-1, \dots, 1\}$.

Using the theory of lumped Markov chains [12, § 6.3, p. 123], we formalise a unified multi-resolution coalescent as the *lumped n -coalescents graph* (Definition 4.1). It is a partially ordered graph whose nodes represent the six coalescents described above and whose edges describe Markov lumpings. We show that a Markov chain $\{B^\uparrow(t)\}$ on \mathbb{B}_n called (i) the *vintaged and labeled n -coalescent* or the Kingman-Tajima n -coalescent, can be lumped into (ii) Kingman's labeled n -coalescent or the *unvintaged and labeled n -coalescent* $\{C^\uparrow(t)\}$ on \mathbb{C}_n , (iii) the *unvintaged and sized n -coalescent* on \mathbb{F}_n and (iv) the *vintaged and shaped n -coalescent* on \mathbb{G}_n . The latter two Markov lumpings are mediated via another Markov chain called (v) the *vintaged and sized n -coalescent* on \mathbb{D}_n . Finally, all these Markov chains are built from the coarsest resolution of

(vi) the *pure death process* on \mathbb{H}_n that gives the number of ancestral lineages of our sample. Figure 1 depicts the six state spaces and the Markov lumpings between them.

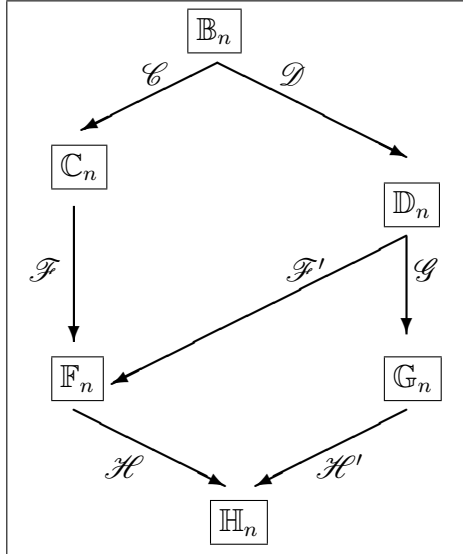


Figure 1: State spaces $\mathbb{B}_n, \mathbb{C}_n, \mathbb{D}_n, \mathbb{F}_n, \mathbb{G}_n, \mathbb{H}_n$ and the lumpings between them.

Here we focus on specific algebraic representations of these six Markov chains and derive their backward-transition, sequence-specific, state-specific and forward-transition probabilities. These derivations are novel for all but Kingman’s labeled n -coalescent and the pure death process [15, 14]. Our motivation for this study is two-fold. The first is historical and the second is statistical as outlined in the next two subsections, respectively.

1.1 Historical Motivation

Kingman and Tajima independently described the genealogical or evolutionary relationship of a sample of size n from a Wright-Fisher population in the early 1980s. The relation between the genealogical objects described by Kingman and Tajima has not been characterised before. We make the first formalization that relates the two distinct sample genealogical descriptions of Kingman and Tajima via the theory of lumped Markov chains. The vintaged and shaped n -coalescent $\{G^\uparrow(t)\}$ or Tajima’s n -coalescent is the first Markov description of Tajima’s evolutionary relationship in the spirit of Kingman’s n -coalescent. $\{G^\uparrow(t)\}$ requires temporal information about the extant sample lineages. Such temporal information is not required for Kingman’s

n -coalescent $\{C^\uparrow(t)\}$. The other resolutions of the sample genealogy studied in this paper make the Markov lumping relations among the n -coalescents, that are naturally spanned by the n -coalescents of Kingman and Tajima, explicit.

Phylogenetics and population genetics, despite being sub-fields of mathematical genetics, are studied by research communities that do not entirely overlap. This is partly driven by methodological preferences between inter-species and intra-species approaches to the study of genetic inter-relatedness. This paper attempts to use definitions and notions that are consistent across phylogenetic and population genetic literature. We show how different resolutions of coalescent sequences are in bijection with different kinds of phylogenetic trees. We show that various classical phylogenetic tree shape statistics can be directly obtained from coarser coalescent resolutions. We express the probability of obtaining coalescent sequences at the coarser resolutions in terms of phylogenetic tree shape statistics. We also show in this paper that classical phylogenetic tree shape statistics, such as, *Colless' index* [6], *Sackin's index* [28, 21], *number of cherries* [18], *Sequential Aldous shape statistics* [2] and *runs statistics* [8], can be obtained efficiently from Markov lumpings of the Kingman-Tajima n -coalescent.

1.2 Statistical Motivation

The n -coalescents provide the basic probability models underlying statistical experiments of interest in population genetics. They arise as prior mixtures over ${}^{\mathcal{C}_n}\mathbb{T}_n$, the partially observed genealogical space of binary coalescent trees with branch-lengths:

$${}^{\mathcal{C}_n}\mathbb{T}_n := \mathcal{C}_n \otimes \mathbb{T}_n := \{{}^{ct} := ({}^{c_n}t_n, {}^{c_{n-1}}t_{n-1}, \dots, {}^{c_2}t_2) : c \in \mathcal{C}_n, t \in \mathbb{T}_n\} ,$$

that one needs to integrate over, in order to obtain the likelihood of a parameter $\phi \in \Phi$ on the basis of some observed data x_{obs} :

$$P(x_{obs}|\phi) = \sum \int_{{}^{ct} \in {}^{\mathcal{C}_n}\mathbb{T}_n} P(x_{obs}|{}^{ct}, \phi) P({}^{ct}|\phi) \partial({}^{ct}) ,$$

where $\partial({}^{ct})$ is the dominating measure on ${}^{\mathcal{C}_n}\mathbb{T}_n$ (given as a product of counting measure on discrete-valued trees in \mathcal{C}_n and Lebesgue measure on the continuous-valued coalescent times in $\mathbb{R}_+^{n-1} =: \mathbb{T}_n$) and $P({}^{ct}|\phi)$ is an n -coalescent induced, possibly ϕ -specific, prior density over ${}^{\mathcal{C}_n}\mathbb{T}_n$. The integration space is depicted for $n = 3$ in Figure 2.

Computational feasibility of “full-likelihood” methods that conduct Monte Carlo integration over ${}^{\mathcal{C}_n}\mathbb{T}_n$, the partially observed genealogical space of binary coalescent

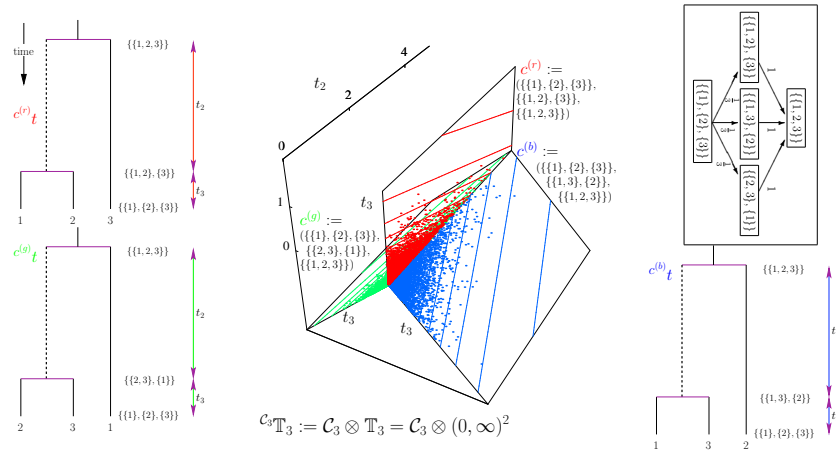


Figure 2: Realizations of 3-coalescent trees in the space of such trees is plotted on the three rectangles as colored points in middle panel. The lines on the rectangles are the contours of the independent exponentially distributed epoch times for each c -sequence. Each of the three coalescent trees, with two branch lengths (t_3, t_2) , representing a realization in the corresponding rectangle and the transition probability diagram of the embedded discrete time Markov chain $\{C^\dagger(k)\}_{k \in \{3,2,1\}}$ on \mathbb{C}_3 are shown counter clock-wise in the four corner panels, respectively. See Proposition 3.7 for details.

trees with branch-lengths, in order to compute the likelihood $P(x_{obs}|\phi)$ via importance samplers [9, 10, 3, 27, 26, 11] for instance, scales poorly with the resolution and size of modern population genomic data. Typical data sets contain DNA sequences of large homologous tracks of the genome for thousands of individuals in a population.

Given the massive scale of current genomic data, computational biologists are using “summary statistics” of the available data to reduce the computational burden of the inference procedure and make it “likelihood-free” on the basis of simulations from the finest genealogical resolutions [31, 19, 5, 17, 25, 4, 16]). However, these “approximate likelihood/Bayesian” computations do not take advantage of the appropriate and sufficient Markov lumpings of the hidden genealogy space for the “summary statistics” being used.

Markov lumping can be powerful in inference if the observed statistic of interest $T(x_{obs})$ only depends on the original chain through the lumping. This can reduce large summations over excessively fine state spaces as noted in [12, p. 124]. The reduction in state space can also be helpful in dynamic programming during integrations over some appropriate Markov lumping of the hidden genealogy space. The Markov lumpings of the Kingman-Tajima n -coalescent developed here can facilitate a computationally efficient and statistically sufficient approach to population-genetic inference based on various families of population-genetic statistics as done in [23] and [22].

Briefly, in [23], the sufficiency of the unvintaged and sized n -coalescent for the likelihood of a popular statistic called the *site frequency spectrum* or SFS is exploited. Computationally efficient inference based on SFS as well as its linear combinations [22], including, the *number of segregating sites* [1], *pair-wise heterozygosity*, and *Tajima’s D* [29], is possible due to the invariance of the sampling distribution of SFS up to the equivalence class induced in the hidden space \mathcal{C}_n by the sequence of states visited by $\{F^\uparrow(t)\}$, the unvintaged and sized n -coalescent. This Markov lumping $\mathcal{F} : \mathcal{C}_n \rightarrow \mathbb{F}_n$ allows us to efficiently integrate over f -sequences or sequential realisations of $\{F^\uparrow(t)\}$ in \mathcal{F}_n to compute the likelihood of the SFS, as opposed to the more conventional approach of integrating over (in importance sampling) or simulating from (in approximate Bayesian/likelihood computations) the unnecessarily finer resolution of c -sequences in \mathcal{C}_n . Importance sampling using a controlled Markov chain is developed in [23] from the forward-transition probabilities of the unvintaged and sized n -coalescent in order to produce f -sequences that are conditioned on the data. Similar inferential methods based on statistics that depend on the other lumped coalescents can be obtained from the coalescent probabilities developed here.

Thus, we formally describe lumped Markov processes at more resolutions of the hidden genealogy space. These descriptions, especially at the coarser resolutions, are

a prerequisite for subsequent computationally efficient inference in the spirit of [23] on the basis of other appropriate genetic statistics, including the sequential Aldous shape statistics, Colless' index, Sackin's index, number of cherries and runs statistics. Moreover, several non-classical statistics can be obtained from the genealogical resolutions studied here.

The backward-transition, sequence-specific, state-specific and forward-transition probabilities at each of our coalescent resolutions described in this paper constitute the applied probabilistic core of computationally efficient Monte Carlo algorithms for statistical inference in population genetics that can exploit the Markov lumping relations among the different coalescent resolutions. Several such algorithms are implemented in `lce`: a C++ class library for lumped coalescent experiments that is publicly available from <http://www.math.canterbury.ac.nz/~r.sainudiin/codes/lce/>.

1.3 Outline

The rest of this paper is organized as follows. In § 2, we review the conditions under which a lumped process is Markov, describe the basic population genetic model and the n -coalescent approximation of any sample genealogical Markov chain. In § 3 we introduce and discuss six n -coalescent resolutions of the genealogical space. Examples and applications are given in § 4.

2 Preliminaries

Let $\mathbb{N} := \{1, 2, 3, \dots\}$ denote the set of natural numbers. Let $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$ and $\mathbb{Z}_- := \{0, -1, -2, \dots\}$ denote the set of non-negative and non-positive integers, respectively. For any set A , let $|A|$ denote its cardinality or the number of elements in it. Let $[n : n']_- := \{n, n-1, \dots, n'+1, n'\}$ denote the linearly ordered descending index set from n to $n' \leq n$, where $n, n' \in \mathbb{Z}$ and let $[n]_- := [n : 1]_- = \{n, n-1, \dots, 2, 1\}$. Similarly, let $[n' : n]_+ := \{n', n'+1, \dots, n-1, n\}$ denote the linearly ordered ascending index set from n' to $n \geq n'$, where $n, n' \in \mathbb{Z}$ and let $[n]_+ := [1 : n]_+ = \{1, 2, \dots, n-1, n\}$.

The n -coalescent resolutions (with exception of the unvintaged and sized n -coalescent and the lineage death process) induce trees on n leaves. We will formally define the trees we will observe throughout this paper.

Definition 2.1. We define the following trees as in [24, § 2.4]. A *ranked, labeled tree* on n leaves is a rooted binary tree with unique leaf labels from the label set \mathcal{L} . The

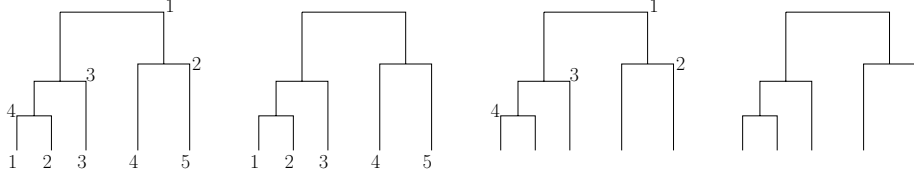


Figure 3: Example for a ranked, labeled tree with leaf label set $\mathfrak{L} = \{1, 2, 3, 4, 5\}$, a labeled tree with $\mathfrak{L} = \{1, 2, 3, 4, 5\}$, a ranked tree shape and a tree shape (from left to right).

interior vertices have a total order $<$ assigned, such that the root is the minimum in this order, and for any interior vertex v which is on the path from an interior vertex w to a leaf, we have $w < v$. We assign to the root of the tree the rank 1, to the second smallest element in this total order the rank 2, etc.

A *labeled tree* on n leaves is a ranked, labeled tree where the total order with the ranks are omitted.

A *ranked tree shape* on n leaves is a ranked, labeled tree where the leaf labels are omitted.

A *tree shape* on n leaves is a labeled tree where the leaf labels are omitted. For examples see Figure 3.

2.1 The lumped chain

In the following we define a *lumped chain* of a Markov chain as in [12, § 6.3, p. 123]. Assume we are given a discrete time Markov chain $\{S(n)\}_{n \in \mathbb{Z}_+}$ on a finite state space $\mathbb{S} = \{s_1, s_2, \dots, s_{|\mathbb{S}|}\}$. Suppose that for an initial distribution π , the values

$$\{P(s_j | s_i) := P(S(n+1) = s_j | S(n) = s_i)\}_{i, j \in [|\mathbb{S}|]}$$

are the time-homogeneous 1-step transition probabilities for our Markov chain.

Let the map

$$\mathcal{M} : \mathbb{S} \rightarrow \mathbb{M} = \{m_1, m_2, \dots, m_{|\mathbb{M}|}\}$$

induce a partition of \mathbb{S} into $|\mathbb{M}|$ non-empty elements via its inverse \mathcal{M}^{-1} . We denote this \mathcal{M} -partition or \mathcal{M} -lumping of \mathbb{S} by

$$\mathbb{S}^{\mathcal{M}} := \{\mathcal{M}^{-1}(m_1), \mathcal{M}^{-1}(m_2), \dots, \mathcal{M}^{-1}(m_{|\mathbb{M}|})\}.$$

Next, we define the *lumped chain* $\{S^{\mathcal{M}}(n)\}_{n \in \mathbb{Z}_+}$ on $\mathbb{S}^{\mathcal{M}}$ from the original Markov chain $\{S(n)\}_{n \in \mathbb{Z}_+}$ on \mathbb{S} . The state visited by the k -th step in the lumped chain is

the set that contains the state visited by the k -th step in the original chain, i.e. with $m_i := \mathcal{M}(s_{i'})$,

$$S(k) = s_{i'} \in \mathbb{S} \implies \mathbb{S}^{\mathcal{M}} \ni S^{\mathcal{M}}(k) = \mathcal{M}^{-1}(\mathcal{M}(s_{i'})) = \mathcal{M}^{-1}(m_i) \supset s_{i'}.$$

At the first level we assign, for the lumped chain,

$$P_{\pi}(S(0) \in \mathcal{M}^{-1}(m_i)).$$

The probability of the sequence $(\mathcal{M}^{-1}(m_k), \dots, \mathcal{M}^{-1}(m_i), \mathcal{M}^{-1}(m_j))$ of n states visited by the lumped chain is defined to be:

$$P_{\pi}(S(n-1) \in \mathcal{M}^{-1}(m_j), S(n-2) \in \mathcal{M}^{-1}(m_i), \dots, S(0) \in \mathcal{M}^{-1}(m_k)).$$

The lumped chain can replace a Markov chain on a vast state space with a chain on fewer states. Such coarser \mathcal{M} -lumpings when statistically sufficient for the considered problem can be advantageous, especially when the lumped chain is also Markov.

The proof of the next proposition is given in [12, Thm. 6.3.2, p. 124]. As this proposition is frequently applied throughout the paper, we give a proof using our terminology.

Proposition 2.2. *A necessary and sufficient condition for the lumped chain $\{S^{\mathcal{M}}(k)\}_{k \in \mathbb{Z}_+}$ to be a Markov chain on $\mathbb{S}^{\mathcal{M}}$ and not depending on the initial distribution π is:*

For every pair of sets $\mathcal{M}^{-1}(m_i)$ and $\mathcal{M}^{-1}(m_j)$ in $\mathbb{S}^{\mathcal{M}}$, the probability of moving from a state $s_{i'} \in \mathcal{M}^{-1}(m_i)$ to the set $\mathcal{M}^{-1}(m_j)$,

$$P(\mathcal{M}^{-1}(m_j)|s_{i'}) := \sum_{s_{j'} \in \mathcal{M}^{-1}(m_j)} P(s_{j'}|s_{i'}),$$

is identical for every $s_{i'} \in \mathcal{M}^{-1}(m_i)$, and thus depends on $s_{i'}$ only through $\mathcal{M}^{-1}(m_i)$.

We refer to these common probabilities by

$$P(\mathcal{M}^{-1}(m_j)|\mathcal{M}^{-1}(m_i)) = P(m_j|m_i),$$

and use them to define transition probabilities between sets of states $\mathcal{M}^{-1}(m_i)$, $\mathcal{M}^{-1}(m_j)$ for the lumped Markov chain $\{S^{\mathcal{M}}(k)\}_{k \in \mathbb{Z}_+}$ on $\mathbb{S}^{\mathcal{M}}$ or equivalently the transition probabilities $P(m_j|m_i)$ between states m_i, m_j for a Markov chain $\{M(k)\}_{k \in \mathbb{Z}_+}$ over \mathbb{M} .

Proof. As the transition probability of the lumped chain does not depend on the initial distribution π , we have,

$$P_\pi (S(1) \in \mathcal{M}^{-1}(m_j) \mid S(0) \in \mathcal{M}^{-1}(m_i))$$

being the same for all π . In particular, this also holds for π having a 1 in the i' -th component, for state $s_{i'} \in \mathcal{M}^{-1}(m_i)$, i.e. when the initial state in the original chain is $s_{i'}$. We denote this probability by

$$p_{ij} := P_{i'} (S(1) \in \mathcal{M}^{-1}(m_j)) = P(\mathcal{M}^{-1}(m_j) \mid s_{i'})$$

for every $s_{i'} \in \mathcal{M}^{-1}(m_i)$. So the condition given in the proposition is necessary.

To prove that it is sufficient, we must show that, if the condition is satisfied, the probability

$$P_\pi (S(n-1) \in \mathcal{M}^{-1}(m_j) \mid S(n-2) \in \mathcal{M}^{-1}(m_i), \dots, S(0) \in \mathcal{M}^{-1}(m_k)) \quad (1)$$

only depends on $\mathcal{M}^{-1}(m_i)$ and $\mathcal{M}^{-1}(m_j)$. We rewrite the probability (1) in the form

$$P_{\pi'} (S(1) \in \mathcal{M}^{-1}(m_j))$$

where π' is a vector with non-zero components only on the states contained in $\mathcal{M}^{-1}(m_i)$. The vector π' depends on π and the first $n-1$ outcomes. However, if $P_{i'} (S(1) \in \mathcal{M}^{-1}(m_j)) = p_{ij}$ for all $s_{i'} \in \mathcal{M}^{-1}(m_i)$, then it is also clear that $P_{\pi'} (S(1) \in \mathcal{M}^{-1}(m_j)) = p_{ij}$. Thus the probability in (1) only depends on $\mathcal{M}^{-1}(m_i)$ and $\mathcal{M}^{-1}(m_j)$. \square

Remark 2.3. The continuous-time Markov chains we encounter in this paper are constructed by composing independent and exponentially distributed waiting times with the discrete-time embedded Markov chain. We are primarily concerned with the lumped chains of the discrete-time Markov chains, since the independent waiting times can be composed with the lumped discrete-time Markov chains to obtain their continuous-time versions as we will see in § 2.2.3.

2.2 The Standard Neutral Wright-Fisher Model

In the Wright-Fisher model [7, 32] of selectively neutral reproduction within a finite population of constant size N , there are discrete, non-overlapping generations labeled by integers $\dots, -k, -k+1, -k+2, \dots, -2, -1, 0, +1, +2, \dots$ as we go forward in time. The current generation is labeled 0. Each individual in generation $-k+1$ is the child of exactly one individual in the previous generation $-k$ and the number of offspring

born to the 1st, 2nd, ..., i^{th} , ..., N^{th} individual of generation $-k$ is the symmetric multinomial random vector $V := (V_1, V_2, \dots, V_N)$, such that:

$$\begin{cases} \sum_{j=1}^N V_j = N, \\ P(V_1 = v_1, V_2 = v_2, \dots, V_N = v_N) = \frac{N!}{v_1! v_2! \dots v_N!} \left(\frac{1}{N}\right)^N. \end{cases} \quad (2)$$

This reproduction scheme is independently and identically enforced in each generation to obtain the standard neutral Wright-Fisher model as we go forward in time. This model has a simple structure as we go back in time. The forward-time offspring distribution of (2) is equivalent to the scheme where each individual in generation $-k + 1$ chooses its parent uniformly at random from among the N individuals in the previous generation $-k$. This simple scheme as we go back in time is at the foundation of the n -coalescent approximation to the Wright-Fisher model.

We can choose to track the sample genealogy (see Fig. 4), i.e. the sub-genealogy of our sample of size n with label set $\mathcal{L} = \{1, 2, \dots, n\}$, within the population genealogy of a Wright-Fisher population of constant size N , at some resolution of interest over an appropriate time-scale. In § 2.2.1 we simply track the number of lineages that are ancestral to our sample in the time-scale of the discrete Wright-Fisher model. The sample genealogical description of § 2.2.1 is the coarsest resolution and contrasts with the finest studied resolution of the sample genealogy in § 3.1. The finest resolution is depicted in Figure 4 for a small example.

2.2.1 Number of Ancestral Lineages of a Wright-Fisher Sample

In the simple Wright-Fisher discrete generation model with a constant population size N the offspring “choose” their parents uniformly and independently at random from the previous generation due to the symmetric multinomial sampling of N offspring from the N parents in the previous generation.

Let $S_i^{(j)}$ denote the Stirling number of the second kind, i.e. $S_i^{(j)}$ is the number of set partitions of a set of size i into j blocks. Let $N_{[j]} := N(N-1) \cdots (N-(j-1))$ and note that the following ratio can be approximated:

$$\begin{aligned} \frac{N_{[j]}}{N^j} &= \frac{N(N-1) \cdots (N-(j-1))}{N^j} = \frac{N}{N} \frac{N-1}{N} \cdots \frac{N-(j-1)}{N} \\ &= 1 \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{j-1}{N}\right) = \prod_{k=1}^{j-1} \left(1 - kN^{-1}\right) = 1 - N^{-1} \sum_{k=1}^{j-1} k + O(N^{-2}) \\ &= 1 - \binom{j}{2} N^{-1} + O(N^{-2}) . \end{aligned}$$

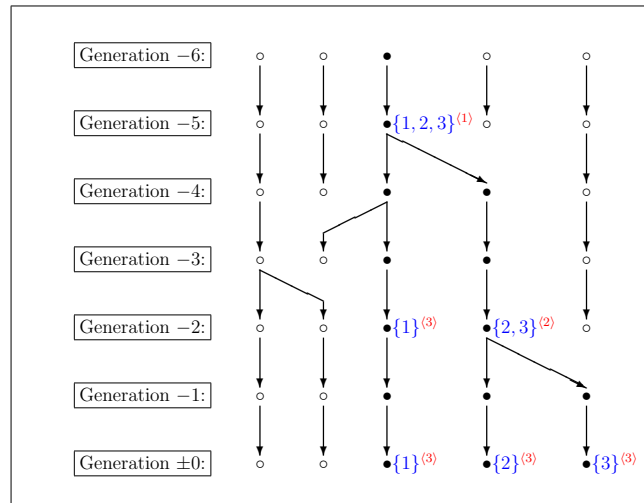


Figure 4: Genealogy of a sample of size $n = 3$ with label set $\mathcal{L} = \{1, 2, 3\}$ within a Wright-Fisher population of constant size $N = 5$. The vintage tags $\langle i \rangle$ in the labeled lineages are assigned as follows. The n sampled individuals at generation 0 have vintage tag $\langle n \rangle$ assigned. Now, going back in time, the first coalescent event has vintage tag $\langle n-1 \rangle$ assigned, the next coalescent event has $\langle n-2 \rangle$ assigned, and so on. Finally, the most recent common ancestor of the sample on n individuals has vintage tag $\langle 1 \rangle$ assigned (see § 3.1).

Thus, the N -specific probability of i extant sample lineages in the current generation becoming j extant lineages in the previous generation is:

$${}^N P_{i,j} = \begin{cases} S_i^{(i)} (N_{[i]} N^{-i}) = 1 (N_{[i]} N^{-i}) = \\ \quad 1 - \binom{i}{2} N^{-1} + O(N^{-2}) & \text{if } j = i, \\ \\ S_i^{(i-1)} (N_{[i-1]} N^{-i}) = \binom{i}{2} (N^{-1} N_{[i-1]} N^{-(i-1)}) = \\ \quad \binom{i}{2} N^{-1} (1 - N^{-1} \binom{i-1}{2} + O(N^{-2})) = \\ \quad \binom{i}{2} N^{-1} + O(N^{-2}) & \text{if } j = i - 1, \\ \\ S_i^{(i-\ell)} (N_{[i-\ell]} N^{-i}) = S_i^{(i-\ell)} (N^{-\ell} N_{[i-\ell]} N^{-(i-\ell)}) = \\ \quad S_i^{(i-\ell)} N^{-\ell} (1 - N^{-1} \binom{i-\ell}{2} + O(N^{-2})) = O(N^{-2}) & \text{if } j = i - \ell, \\ \\ 0 & \text{otherwise .} \end{cases} \quad (3)$$

where $1 < \ell < i - 1$. Let $\mathbb{Z}_- := \{0, -1, -2, \dots\}$ be the ordered and countably infinite discrete time index set. The discrete time Markov chain $\{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ over the state space $\mathbb{H}_n := \{n, n - 1, \dots, 1\}$ with 1-step transition probabilities (3) is termed *the death chain of the number of ancestral sample lineages within the Wright-Fisher population of constant size N* . The initial state and the final absorbing state of this chain are n and 1, respectively.

2.2.2 Death process of the number of lineages

Let us first obtain a coalescent approximation of $\{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$, the death chain of the number of ancestral sample lineages within the Wright-Fisher population of constant size N from § 2.2.1. This is the coarsest of the six coalescent resolutions we study here and forms the foundation for a continuous time approximation of *any* discrete time sample genealogical Markov chain $\{{}^N A^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ that has $\{{}^N A^{\uparrow \mathcal{H}}(k)\}_{k \in \mathbb{Z}_-} = \{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ as its lumped chain via the lumping $\mathcal{H} : \mathbb{A}_n \rightarrow \mathbb{H}_n$ that reports the number of ancestral lineages of our sample (see § 2.2.3).

Let us rescale time in the discrete time Markov chain $\{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ over the state space $\mathbb{H}_n := \{n, n - 1, \dots, 1\}$ with 1-step transition probabilities (3). Let the rescaled time t be g in units of N generations, i.e. $g = \lfloor Nt \rfloor$. In words, the probability that any specific pair of lineages, among the $\binom{i}{2}$ many pairs of the currently extant i

ancestors of the n sampled lineages, coalesces in one generation is $1/N$ and that this pair remains distinct for more than g generations is $(1 - 1/N)^g$. Then, the probability that a pair of lineages remain distinct for more than t units of the rescaled time is: $(1 - 1/N)^{\lfloor Nt \rfloor} \xrightarrow{N \rightarrow \infty} e^{-t}$. The $\lfloor Nt \rfloor$ -step transition probabilities, ${}^N P_{i,j}(\lfloor Nt \rfloor)$, of the discrete time death chain $\{{}^N H^\uparrow(\lfloor Nt \rfloor)\}_{\lfloor Nt \rfloor \in \mathbb{Z}_-}$ converge to the transition probabilities $P_{i,j}(t)$ of the *pure death process* $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$, in the rescaled time t , over the state space \mathbb{H}_n , as the population size N tends to infinity. The instantaneous transition rates for this pure death or epoch-time process is [14, (1.9)]:

$${}^N P_{i,j}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{i,j}(t) = \exp(Qt), \quad \text{where}$$

$$q_{i,j} = q(j|i) = \begin{cases} -\binom{i}{2} & \text{if } j = i \\ \binom{i}{2} & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The matrix Q is called the instantaneous rate matrix of the death process Markov chain $\{H(t)\}_{t \in \mathbb{R}_+}$ and its (i, j) -th entry is $q_{i,j} = q(j|i)$. Thus, the i -th holding time or epoch-time random variable T_i during which time there are i distinct ancestral lineages of our sample is approximately exponentially distributed with rate parameter $\binom{i}{2}$ and is independent of other epoch-times. In other words, for large N , the random vector $T = (T_2, T_3, \dots, T_n)$ of epoch-times, corresponding to the epoch times of the pure death process $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$ on the state space \mathbb{H}_n , has the product exponential density $\bigotimes_{i=2}^n \binom{i}{2} e^{-\binom{i}{2} t_i}$ over its support $\mathbb{T}_n := \mathbb{R}_+^{n-1}$. Note that the initial state of $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$ is n and the final absorbing state is 1.

Let $[n]_- := \{n, n-1, \dots, 2, 1\}$ denote the ordered discrete time index set of the jump chain. The embedded discrete time jump chain $\{H^\uparrow(k)\}_{k \in [n]_-}$ of this death process, termed *the embedded death chain*, moves from state i to state $i-1$ with probability 1, as follows:

$$\boxed{n} \xrightarrow{1} \boxed{n-1} \xrightarrow{1} \dots \xrightarrow{1} \boxed{i+1} \xrightarrow{1} \boxed{i} \xrightarrow{1} \boxed{i-1} \xrightarrow{1} \dots \xrightarrow{1} \boxed{2} \xrightarrow{1} \boxed{1}.$$

We keep track of the discrete time in terms of the extant number of lineages for convenience. The discrete time steps $k \in [n]_- := \{n, n-1, \dots, 2, 1\}$ of the embedded chain $\{H^\uparrow(k)\}_{k \in [n]_-}$ of the death process $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$ are referred to as *coalescent epochs* or *epochs* as they mark the beginning of a lineage death or coalescence event. Note however that the embedded discrete time jump chain of $\{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ denoted by $\{{}^N H^\uparrow(k)\}_{k \in [n:n']_-}$ can reach the absorbing state in merely $n - n'$ jumps where $1 \leq n' < n$.

$i' = i - \ell$, i.e. there are i and i' lineages in a_i and $a_{i'}$, respectively. Then,

$$\begin{aligned}
{}^N P_{a_i, a_{i'}} &:= P\left({}^N A(k+1) = a_{i'} \mid {}^N A(k) = a_i\right) \\
&= P\left({}^N A(k+1) = a_{i'}, {}^N A(k+1) \in \mathcal{H}^{-1}(\mathcal{H}(a_{i'})) \mid {}^N A(k) = a_i\right) \\
&= \frac{P\left({}^N A(k+1) = a_{i'}, {}^N A(k+1) \in \mathcal{H}^{-1}(i - \ell), {}^N A(k) = a_i\right)}{P\left({}^N A(k) = a_i\right)} \\
&= P\left({}^N A(k+1) = a_{i'} \mid {}^N A(k+1) \in \mathcal{H}^{-1}(i'), {}^N A(k) = a_i\right) \\
&\quad \times P\left({}^N A(k+1) \in \mathcal{H}^{-1}(i') \mid {}^N A(k) = a_i\right) \\
&=: {}^N P\left(a_{i'} \mid a_{i'} \in \mathcal{H}^{-1}(i'), a_i\right) {}^N P\left(\mathcal{H}^{-1}(i') \mid a_i\right) . \tag{5}
\end{aligned}$$

Let the conditional transition probability of the jump chain $\{{}^N A^\uparrow(k)\}_{k \in [n:n']_-}$ be

$$P(a_{i'} \mid a_i) := {}^N P\left(a_{i'} \mid a_{i'} \in \mathcal{H}^{-1}(\mathcal{H}(a_{i'})), a_i\right) . \tag{6}$$

If the lumped chain $\{{}^N A^{\uparrow \mathcal{H}}(k)\}_{k \in \mathbb{Z}_-}$ of $\{{}^N A^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ is the Markov chain $\{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ then the last probability term in (5) simplifies:

$$\begin{aligned}
&P\left({}^N A(k+1) \in \mathcal{H}^{-1}(i') \mid {}^N A(k) = a_i\right) \\
&= \sum_{a_{i'} \in \mathcal{H}^{-1}(i')} P\left({}^N A(k+1) = a_{i'} \mid {}^N A(k) = a_i\right) \\
&=: {}^N P\left(\mathcal{H}^{-1}(i') \mid a_i\right) \\
&= {}^N P\left(\mathcal{H}^{-1}(i') \mid \mathcal{H}^{-1}(a_i)\right) \quad \because \text{of Proposition 2.2} \\
&= {}^N P\left(i' \mid i\right) = S_i^{(i')} \left(N_{[i']} N^{-i}\right) \quad \because \text{of (3)} \tag{7}
\end{aligned}$$

Combining (5) with (6), (7) and (3) we get

$${}^N P_{a_i, a_{i'}} = \begin{cases} 1 S_i^{(i)} \left(N_{[i]} N^{-i}\right) = \\ \quad 1 - \binom{i}{2} N^{-1} + O(N^{-2}) & \text{if } a_i = a_{i'} \text{ and} \\ & \mathcal{H}(a_i) = \mathcal{H}(a_{i'}) = i, \\ P(a_{i'} \mid a_i) S_i^{(i-1)} \left(N_{[i-1]} N^{-i}\right) = \\ \quad P(a_{i'} \mid a_i) \binom{i}{2} N^{-1} + O(N^{-2}) & \text{if } a_{i'} \prec_a^N a_i \text{ and} \\ & i = \mathcal{H}(a_i) = \mathcal{H}(a_{i'}) + 1, \\ P(a_{i'} \mid a_i) S_i^{(i-\ell)} \left(N_{[i-\ell]} N^{-i}\right) = \\ \quad O(N^{-2}) & \text{if } a_{i'} \prec_a^N a_i \text{ and} \\ & i = \mathcal{H}(a_i) = \mathcal{H}(a_{i'}) + \ell, \\ 0 & \text{otherwise ,} \end{cases} \tag{8}$$

where $1 < \ell < i - 1$. Let us define the more restrictive immediate precedence relation on $\mathbb{A}_n \ni a_i, a_{i'}$:

$$a_{i'} \prec_a a_i \iff a_{i'} \prec_a^N a_i, i = \mathcal{H}(a_i) = \mathcal{H}(a_{i'}) + 1 .$$

The $\lfloor Nt \rfloor$ -step transition probabilities, ${}^N P_{a_i, a_{i'}}(\lfloor Nt \rfloor)$, of $\{{}^N A^\uparrow(\lfloor Nt \rfloor)\}_{\lfloor Nt \rfloor \in \mathbb{Z}_-}$ converge to the transition probabilities $P_{a_i, a_{i'}}(t)$ of the n -coalescent $\{A^\uparrow(t)\}_{t \in \mathbb{R}_+}$, in the rescaled time t , over the state space \mathbb{A}_n , as the population size N tends to infinity. The instantaneous transition rates for this continuous-time Markov chain, $\{A^\uparrow(t)\}_{t \in \mathbb{R}_+}$, generalizes [15, (2.10)] to the sample genealogical resolution of a -sequences in \mathcal{A}_n . More formally,

$${}^N P_{a_i, a_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{a_i, a_{i'}}(t) = \exp(Qt), Q := \{q_{a_i, a_{i'}}\}_{a_i, a_{i'} \in \mathbb{A}_n}, \text{ and}$$

$$q_{a_i, a_{i'}} = q(a_{i'} | a_i) = \begin{cases} -\binom{i}{2} & \text{if } a_{i'} = a_i \\ P(a_{i'} | a_i) \binom{i}{2} & \text{if } a_{i'} \prec_a a_i, i = \mathcal{H}(a_i) \\ 0 & \text{otherwise} \end{cases} . \quad (9)$$

This establishes the following proposition.

Proposition 2.4. *Let $\{{}^N A^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ be any discrete time sample genealogical Markov chain that has $\{{}^N A^{\uparrow \mathcal{H}}(k)\}_{k \in \mathbb{Z}_-} = \{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ — the Wright-Fisher death chain over $\mathbb{H}_n := \{n, n-1, \dots, 1\}$ with 1-step transition probabilities in (3) — as its lumped Markov chain, via the lumping map $\mathcal{H}(a_i) : \mathbb{A}_n \rightarrow \mathbb{H}_n$. Then $\{{}^N A^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ can be approximated by $\{A^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the n -coalescent of the desired resolution, in the sense of (9).*

Remark 2.5. Further, the Markov chain $\{A^\uparrow(t)\}_{t \in \mathbb{R}_+}$ has two independent components: (I) the death chain over $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with waiting times $\binom{i}{2}$, and (II) the simpler jump chain $\{A^\uparrow(k)\}_{k \in [n]_-}$ that only loses one lineage at each jump with transition probabilities $P(a_{i'} | a_i)$ (as opposed to complicated jump chain $\{{}^N A^\uparrow(k)\}_{k \in [n:n']_-}$, embedded in $\{{}^N A^\uparrow(k)\}_{k \in \mathbb{Z}_-}$). Note that $\{A^\uparrow(k)\}_{k \in [n]_-}$ is our refinement or delumping of the embedded death chain $\{H^\uparrow(k)\}_{k \in [n]_-}$.

3 Six coalescent resolutions

$\{H^\uparrow(k)\}_{k \in [n]_-}$, the embedded discrete time jump chain of the death chain introduced and discussed in § 2.2.2, is the coarsest of our coalescent resolutions. In this Section, we introduce n -coalescent approximations of five refined resolutions of the sample

genealogy. One of them (§ 3.2) was completely developed as Markov processes by Kingman. Another (§ 3.5) was pointed out by Kingman and yet another (§ 3.4) by Tajima without a full Markov description. The remaining two (§ 3.1, § 3.3), including the finest resolution of § 3.1, have not been studied before. Figure 1 depicts the six state spaces and the Markov lumpings between them.

As derived in the previous section, we can decompose the coalescent process: we have a continuous time process describing the time between a jump from k to $k - 1$ lineages and we have an embedded jump process describing the state of the genealogy at each jump at the given resolution. When the chain makes a jump we have a resolution-specific coalescence event. In the following, we will describe this embedded jump process at different resolutions.

3.1 Vintaged and labeled n -coalescent

We introduce the finest coalescent resolution in this study. At this resolution, in each epoch, we keep track of the descendants of each existing lineage as well as the epoch at which this lineage was created as we follow the genealogy of our sample back through continuous time. We will see that this genealogical process, $\{B^\dagger(t)\}_{t \in \mathbb{R}_+}$, called the vintaged and labeled n -coalescent, is a continuous time Markov chain and that each sequence of distinct states visited by $\{B^\dagger(k)\}_{k \in [n]_-}$, the jump Markov chain of $\{B^\dagger(t)\}_{t \in \mathbb{R}_+}$, induces a unique ranked, labeled tree, i.e. there is a bijection between the set of sequential realizations of the jump chain of the vintaged and labeled n -coalescent and the set of ranked, labeled trees. Furthermore, this process can be lumped to any other process we will introduce below.

Next we derive the state space, \mathbb{B}_n , of $\{B^\dagger(k)\}_{k \in [n]_-}$ and $\{B^\dagger(t)\}_{t \in \mathbb{R}_+}$. Let \mathbb{C}_n be the set of all set partitions of the label set $\mathcal{L} = \{1, 2, \dots, n\}$ of n samples. Let $|c_a|$ denote the number of elements in $c_a \in \mathbb{C}_n$. Denote by \mathbb{C}_n^i the set of all set partitions with i blocks, i.e., $\mathbb{C}_n = \bigcup_{i=1}^n \mathbb{C}_n^i$. Let $c_i := \{c_{i,1}, c_{i,2}, \dots, c_{i,i}\} \in \mathbb{C}_n^i$ denote the i elements of c_i . The partial ordering \prec_c on \mathbb{C}_n is based on the immediate precedence relation \prec_c :

$$c_{i'} \prec_c c_i \Leftrightarrow c_{i'} = c_i \setminus c_{i,j} \setminus c_{i,k} \cup (c_{i,j} \cup c_{i,k}), \quad j \neq k, j, k \in \{1, 2, \dots, |c_i|\}.$$

In words, $c_{i'} \prec_c c_i$, read as $c_{i'}$ immediately precedes c_i , means that $c_{i'}$ can be obtained from c_i by coalescing any distinct pair of elements in c_i . Thus, $c_{i'} \prec_c c_i$ implies $|c_{i'}| = |c_i| - 1$.

Let the coalescent epochs be labeled $n, n - 1, \dots, 1$ as we go back in time. Thus, there are k lineages during epoch k . We say that a lineage identified by $c_{i,j}$ in the i -th epoch, i.e. the lineage that subtends the sample labels in the set $c_{i,j}$, is of $m_{i,j}$ vintage

if $c_{i,j}$ originated in epoch $m_{i,j}$. We also say that $m_{i,j}$ is the coalescent-epoch vintage or simply the vintage of $c_{i,j}$. We notate such lineage-vintage pairs, $\text{lineage}^{(\text{vintage})}$, or *vintaged lineages* by $b_{i,j} := c_{i,j}^{\langle m_{i,j} \rangle}$ and let

$$b_i := \{b_{i,1}, b_{i,2}, \dots, b_{i,i}\} := \left\{ c_{i,1}^{\langle m_{i,1} \rangle}, c_{i,2}^{\langle m_{i,2} \rangle}, \dots, c_{i,i}^{\langle m_{i,i} \rangle} \right\},$$

denote the i vintaged lineages in epoch i formed by pairing the j -th element $c_{i,j} \in c_i \in \mathbb{C}_n^i$ with its respective vintage $m_{i,j} \in \{n, n-1, \dots, i\}$, for each $j \in \{1, 2, \dots, i\}$. Let the set of such b_i 's be \mathbb{B}_n^i and let $\mathbb{B}_n := \bigcup_{i=1}^n \mathbb{B}_n^i$. Thus, \mathbb{B}_n is a vintage augmentation of \mathbb{C}_n . The partial ordering \prec_b on \mathbb{B}_n is inherited from the immediate precedence relation \prec_b :

$$b_{i'} \prec_b b_i \Leftrightarrow b_{i'} = b_i \setminus c_{i,j}^{\langle m_{i,j} \rangle} \setminus c_{i,k}^{\langle m_{i,k} \rangle} \bigcup (c_{i,j} \cup c_{i,k})^{\langle |b_i| - 1 \rangle},$$

$$j \neq k, j, k \in \{1, 2, \dots, |b_i|\} .$$

In words, $b_{i'} \prec_b b_i$, read as $b_{i'}$ immediately precedes b_i , means that $b_{i'}$ can be obtained from b_i by coalescing any distinct pair of lineages in b_i and updating the coalesced lineage's vintage tag to that of the new epoch label. Let $b := (b_n, b_{n-1}, \dots, b_1)$ be a sequence of states in \mathbb{B}_n that consecutively satisfy the immediate precedence relation \prec_b and the set of such b -sequences be \mathcal{B}_n , i.e.,

$$b := (b_n, b_{n-1}, \dots, b_1)$$

$$\in \mathcal{B}_n := \{b : b_i \in \mathbb{B}_n^i, b_{i-1} \prec_b b_i, \forall i \in \{n, n-1, \dots, 3, 2\}\} .$$

The initial state and the final absorbing state of $\{B^\dagger(k)\}_{k \in [n]_-}$, the jump chain on \mathbb{B}_n , are $b_n = \{\{1\}^{\langle n \rangle}, \{2\}^{\langle n \rangle}, \dots, \{n\}^{\langle n \rangle}\}$ and $b_1 = \{\{1, 2, \dots, n\}^{\langle 1 \rangle}\}$, respectively. The three b -sequences when $n = 3$ are given in Table 3. Next we give the transition probabilities of $\{B^\dagger(k)\}_{k \in [n]_-}$ on \mathbb{B}_n .

Proposition 3.1 (Backward transition probabilities of a b -sequence). *The transition probabilities of the jump Markov chain $\{B^\dagger(k)\}_k$, with discrete time $k = n, n-1, \dots, 2, 1$ and finite state space \mathbb{B}_n are:*

$$P(b_{i-1}|b_i) = \begin{cases} \binom{i}{2}^{-1} & \text{if } b_{i-1} \prec_b b_i, b_i \in \mathbb{B}_n^i \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Proof. When there are i vintaged lineages in the i -th coalescent epoch, a coalescence event can reduce the number of lineages to $i-1$ by coalescing one of $\binom{i}{2}$ many pairs of vintaged lineages uniformly at random. Hence, the inverse $\binom{i}{2}^{-1}$ appears

in the transition probabilities. The conditions that $b_{i-1} \prec_b b_i$ and $b_i \in \mathbb{B}_n^i$ for each $i \in \{n, n-1, \dots, 3, 2\}$ ensure that our b -sequence $b = (b_n, \dots, b_1)$ remains in \mathcal{B}_n as we go backwards in time from the i -th coalescent epoch with i samples to the $(i-1)$ -th coalescent epoch. \square

Proposition 3.2 (Probabilities of a b -sequence). *The probability of a b -sequence $b := (b_n, b_{n-1}, \dots, b_1) \in \mathcal{B}_n$ is:*

$$P(b) = P(b_{n-1}|b_n)P(b_{n-2}|b_{n-1}) \cdots P(b_1|b_2) = \frac{2^{n-1}}{n!(n-1)!}. \quad (11)$$

Proof. The first equality in (11) is a consequence of Markov property and the second equality results from a telescoping cancellation when applying (10) to the product components of the second term in (11). \square

Therefore the probability of a b -sequence in (11) is constant for all b -sequences, i.e. it is uniformly distributed over \mathcal{B}_n with

$$P(b) = \frac{1}{|\mathcal{B}_n|} \implies |\mathcal{B}_n| = \frac{n!(n-1)!}{2^{n-1}}. \quad (12)$$

Proposition 3.3 (Bijection between ranked, labeled trees and b -sequences). *There is a bijection between the set of ranked, labeled trees on n leaves and \mathcal{B}_n , the set of b -sequences.*

Proof. It is easy to see that each ranked, labeled tree induces a distinct b -sequence and any two distinct b -sequences induce two distinct ranked, labeled trees. \square

The next proposition gives the probability of visiting a particular state b_i with i blocks.

Proposition 3.4 (Probability of $b_i \in \mathbb{B}_n^i$). *Without loss of generality, let us chronologically list $b_i = \{c_{i,1}^{(m_{i,1})}, c_{i,2}^{(m_{i,2})}, \dots, c_{i,i}^{(m_{i,i})}\}$, such that $m_{i,1} \leq m_{i,2} \leq \dots \leq m_{i,i}$. Let $c_{i,1:j} := c_{i,1} \cup c_{i,2} \cup \dots \cup c_{i,j}$, where $c_{i,j}$'s are the unvintaged blocks in the chronologically listed b_i . Then,*

$$P(b_i) = \frac{i!(i-1)!}{n!(n-1)!} \left(\frac{\prod_{j=1}^{i'} |c_{i,j}|! (|c_{i,j}| - 1) (|c_{i,1:j}| - j - 1 - m_{i,j} + i)!}{\prod_{j=1}^{i'-1} (|c_{i,1:j}| - j - m_{i,j+1} + i)!} \right), \quad (13)$$

Proof. For a b -sequence $b \in \mathbb{B}_n$ define $b_{n:i} := (b_n, b_{n-1}, \dots, b_{i+1}, b_i)$. Then by (11) and (10),

$$P(b_{n:i}) := P((b_n, \dots, b_i)) = P(b_{n-1}|b_n) \dots P(b_i|b_{i+1}) = \frac{2^{n-i} i! (i-1)!}{n!(n-1)!}.$$

In particular, each sequence $b_{n:i}$ is equally likely. Let N_i be the number of $b_{n:i}$ -sequences which lead to $b_i = \{c_{i,1}^{\langle m_{i,1} \rangle}, c_{i,2}^{\langle m_{i,2} \rangle}, \dots, c_{i,i}^{\langle m_{i,i} \rangle}\}$. Determining N_i establishes the probability for b_i , since each $b_{n:i}$ -sequence is equally likely, i.e.

$$P(b_i) = P(b_{n:i}) N_i. \quad (14)$$

Without loss of generality, we can assume for a given b_i , we have $m_{i,1} \leq m_{i,2} \leq \dots \leq m_{i,i}$. Thus, we assume that each b_i is chronologically listed. Let $i' = \max\{j : m_{i,j} < n\}$. Further define $c_{i,1} \cup c_{i,2} \cup \dots \cup c_{i,j-1} =: c_{i,1:j-1}$.

For a vintaged lineage $b_{i,j} = c_{i,j}^{\langle m_{i,j} \rangle}$ in epoch i , the number of possible b -sequences in $\mathcal{B}_{|c_{i,j}|}$ using (12) is:

$$|\mathcal{B}_{|c_{i,j}|}| = \frac{|c_{i,j}|! (|c_{i,j}| - 1)!}{2^{|c_{i,j}| - 1}}. \quad (15)$$

In order to calculate N_i , we need to define $N_{i,j}$. Let $N_{i,j}$ be the number of b -sequences on the label set $c_{i,1:j}$ stopped when all but j lineages coalesced, respecting (1) a fixed b -sequence on the label set $c_{i,1:j-1}$ stopped when all but $j-1$ lineages coalesced, and respecting (ii) a fixed b -sequence on the label set $c_{i,j}$, $j \leq i'$ stopped when all lineages coalesced. We have

$$N_i = |\mathcal{B}_{|c_{i,1}|}| \times |\mathcal{B}_{|c_{i,2}|}| \times N_{i,2} \times |\mathcal{B}_{|c_{i,3}|}| \times N_{i,3} \times \dots \times |\mathcal{B}_{|c_{i,i'}|}| \times N_{i,i'}. \quad (16)$$

We will now determine $N_{i,j}$. Note that there are $|c_{i,1:j-1}| - (j-1)$ coalescent events on $c_{i,1:j-1}$ up to epoch i . There are $|c_{i,j}| - 1$ coalescent events on $c_{i,j}$. The coalescent events in epoch $i, i+1, \dots, m_{i,j}-1$ happen on $c_{i,1:j-1}$, coalescent event $m_{i,j}$ happens on $c_{i,j}$. The remaining elements are shuffled together arbitrarily, the number of possible shuffles equals $N_{i,j}$, which is,

$$N_{i,j} = \binom{|c_{i,1:j-1}| - (j-1) - (m_{i,j} - i) + |c_{i,j}| - 2}{|c_{i,j}| - 2}. \quad (17)$$

So overall, using Equations (14 – 17), we obtain,

$$\begin{aligned}
P(b_i) &= P(b_{n:i})N_i \\
&= 2^{n-i} \frac{i!(i-1)!}{n!(n-1)!} \prod_{j=1}^{i'} \frac{|c_{i,j}|!(|c_{i,j}|-1)!}{2^{|c_{i,j}|-1}} \\
&\quad \prod_{j=2}^{i'} \binom{|c_{i,1:j-1}| - (j-1) - (m_{i,j} - i) + |c_{i,j}| - 2}{|c_{i,j}| - 2} \\
&= \frac{i!(i-1)!}{n!(n-1)!} \left(\prod_{j=1}^{i'} |c_{i,j}|! \right) (|c_{i,1}| - 1)! \left(\prod_{j=2}^{i'} \frac{(|c_{i,j}| - 1)(|c_{i,1:j}| - j - 1 - (m_{i,j} - i))!}{(|c_{i,1:j-1}| - j + 1 - (m_{i,j} - i))!} \right) \\
&= \frac{i!(i-1)!}{n!(n-1)!} \left(\prod_{j=1}^{i'} |c_{i,j}|! \right) (|c_{i,1}| - 1)! \\
&\quad \left(\prod_{j=2}^{i'} (|c_{i,j}| - 1) \right) \left(\frac{\prod_{j=2}^{i'} (|c_{i,1:j}| - j - 1 - (m_{i,j} - i))!}{\prod_{j=1}^{i'-1} (|c_{i,1:j}| - j - (m_{i,j+1} - i))!} \right) \\
&= \frac{i!(i-1)!}{n!(n-1)!} \left(\prod_{j=1}^{i'} |c_{i,j}|!(|c_{i,j}| - 1) \right) \left(\frac{\prod_{j=1}^{i'} (|c_{i,1:j}| - j - 1 - m_{i,j} + i)!}{\prod_{j=1}^{i'-1} (|c_{i,1:j}| - j - m_{i,j+1} + i)!} \right) \\
&= \frac{i!(i-1)!}{n!(n-1)!} \left(\frac{\prod_{j=1}^{i'} |c_{i,j}|!(|c_{i,j}| - 1)(|c_{i,1:j}| - j - 1 - m_{i,j} + i)!}{\prod_{j=1}^{i'-1} (|c_{i,1:j}| - j - m_{i,j+1} + i)!} \right)
\end{aligned}$$

which completes the proof. \square

Next we study the jump Markov chain on \mathbb{B}_n forward in time. This chain is denoted by $\{B^\downarrow(k)\}_{k \in [n]_+}$ over the ordered time index set $[n]_+ := \{1, 2, \dots, n\}$ that denotes the epochs.

Proposition 3.5 (Forward transition probabilities of a b -sequence). *The transition probability from state b_{i-1} to state b_i , such that $b_{i-1} \prec_b b_i \in \mathbb{B}_n^i$, in the forward jump chain $\{B^\downarrow(k)\}_{k \in [n]_+}$, denoted by $P(b_i|b_{i-1})$ is, by Bayes law,*

$$P(b_i|b_{i-1}) = \frac{P(b_{i-1}|b_i)P(b_i)}{P(b_{i-1})} = \frac{P(b_i)}{\binom{i}{2}P(b_{i-1})}$$

using the $P(b_i)$ from Equation 13.

Let $\{^N B^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ be the discrete time sample genealogical Markov chain of n vintaged sample lineages labeled by $\mathfrak{L} = \{1, 2, \dots, n\}$ and taken at random from the present generation of a Wright-Fisher population of constant size N over the state space \mathbb{B}_n . We derive an approximation of this chain in rescaled time next.

Proposition 3.6 (Vintaged and labeled n -coalescent). *The $\lfloor Nt \rfloor$ -step transition probabilities, ${}^N P_{b_i, b_{i'}}(\lfloor Nt \rfloor)$, of the chain $\{{}^N B^\uparrow(k)\}_{k \in \mathbb{Z}_-}$, converge to the transition probabilities of the continuous-time Markov chain $\{B^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with rate matrix Q , i.e.*

$${}^N P_{b_i, b_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{b_i, b_{i'}}(t) = \exp(Qt),$$

where the entries of Q , $q(b_{i'}|b_i)$, $b_{i'}, b_i \in \mathbb{B}_n$, specifying the transition rate from b_i to $b_{i'}$, are:

$$q(b_{i'}|b_i) = \begin{cases} -\binom{i}{2} & \text{if } b_{i'} = b_i, b_i \in \mathbb{B}_n^i \\ P(b_{i'}|b_i) \binom{i}{2} = \binom{i}{2}^{-1} \binom{i}{2} = 1 & \text{if } b_{i'} \prec_b b_i \\ 0 & \text{otherwise} \end{cases}. \quad (18)$$

We call this continuous-time Markov chain as the vintaged and labeled n -coalescent. The initial state is $b_n = \{\{1\}^{(n)}, \{2\}^{(n)}, \dots, \{n\}^{(n)}\}$ and the final absorbing state is $b_1 = \{\{1, 2, \dots, n\}^{(1)}\}$.

Proof. The proof is merely a consequence of substituting the backward transition probabilities at (10) in the general n -coalescent approximation of (9) since $\{{}^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ is a lumped Markov chain of $\{{}^N B^\uparrow(k)\}_{k \in \mathbb{Z}_-}$. \square

We call this vintaged and labeled n -coalescent as the Kingman-Tajima n -coalescent. We will see that Kingman's n -coalescent of § 3.2 as well Tajima's n -coalescent of § 3.4 are lumped Markov processes of the Kingman-Tajima n -coalescent.

3.2 Unvintaged and labeled n -coalescent

We will obtain $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the Markov chain called the unvintaged and labeled n -coalescent over \mathbb{C}_n , by a Markov lumping of $\{B^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and labeled n -coalescent over \mathbb{B}_n , that omits the epoch vintages from the states in \mathbb{B}_n . Each sequence of distinct states visited by the jump chain $\{C^\uparrow(k)\}_{k \in [n]_-}$ that is embedded in $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$ once again induces a ranked, labeled tree, i.e. there is a bijection between the set of sequential realizations of the jump chain of the unvintaged and labeled n -coalescent and the set of ranked, labeled trees. Note that we already established a bijection between the set of sequential realizations of the jump chain of the vintaged and labeled n -coalescent and the set of ranked, labeled trees. However, the state space of the unvintaged and labeled n -coalescent is significantly smaller than that of the vintaged and labeled n -coalescent. In our nomenclature, the unvintaged and labeled n -coalescent is Kingman's n -coalescent [15, 14]. The number of elements

in \mathbb{C}_n is the number of set partitions of a set of size n which is $\text{Bell}(n)$, the n -th Bell number

$$|\mathbb{C}_n| = \text{Bell}(n) := \sum_{j=0}^n S_n^{(j)}, \quad (19)$$

where $S_n^{(j)}$ is the Stirling number of the second kind.

Consider the jump Markov chain $\{C^\uparrow(k)\}_{k \in [n]_-}$ on \mathbb{C}_n with initial state $c_n = \{\{1\}, \{2\}, \dots, \{n\}\}$ and final absorbing state $c_1 = \{\{1, 2, \dots, n\}\}$, with the following transition probabilities [14, (2.2)]:

$$P(c_{i'}|c_i) = \begin{cases} \binom{i}{2}^{-1} & : \text{if } c_{i'} \prec_c c_i, c_i \in \mathbb{C}_n^i \\ 0 & : \text{otherwise} \end{cases}. \quad (20)$$

Now, let $c := (c_n, c_{n-1}, \dots, c_1)$ be a c -sequence or coalescent sequence obtained from the sequence of states visited by a sequential realization of $\{C^\uparrow(k)\}_{k \in [n]_-}$, and denote the set of such c -sequences by

$$\mathcal{C}_n := \{c := (c_n, c_{n-1}, \dots, c_1) : c_i \in \mathbb{C}_n^i, c_{i-1} \prec_c c_i, i \in \{n, n-1, \dots, 2\}\}$$

The probability that $c_i \in \mathbb{C}_n^i$ is visited by the chain [14, (2.3)] is:

$$P(c_i) = \frac{(n-i)! i! (i-1)!}{n! (n-1)!} \prod_{j=1}^i |c_{i,j}|!, \quad (21)$$

and the probability of a c -sequence is uniformly distributed over \mathcal{C}_n with

$$P(c) = \prod_{i=n}^2 P(c_{i-1}|c_i) = \frac{2^{n-1}}{n! (n-1)!} = \frac{1}{|\mathcal{C}_n|}. \quad (22)$$

Let $\{^N C^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ be the discrete time sample genealogical Markov chain of n samples labeled by $\mathcal{L} = \{1, 2, \dots, n\}$ and taken at random from the present generation of a Wright-Fisher population of constant size N over the state space \mathbb{C}_n . We derive a continuous-time Markov chain that approximates $\{^N C^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ next.

Proposition 3.7 (Unvintaged and labeled n -coalescent [15, (2.10)]). *The $\lfloor Nt \rfloor$ -step transition probabilities, ${}^N P_{c_i, c_{i'}}(\lfloor Nt \rfloor)$, of the chain $\{^N C^\uparrow(k)\}_{k \in \mathbb{Z}_-}$, converge to the transition probabilities of the continuous-time Markov chain $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with rate matrix Q , i.e.*

$${}^N P_{c_i, c_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{c_i, c_{i'}}(t) = \exp(Qt),$$

where the entries of Q , $q(c_{i'}|c_i)$, $c_{i'}, c_i \in \mathbb{C}_n$, specifying the transition rate from c_i to $c_{i'}$, are:

$$q(c_{i'}|c_i) = \begin{cases} -\binom{i}{2} & \text{if } c_{i'} = c_i, c_i \in \mathbb{C}_n^i \\ P(c_{i'}|c_i)\binom{i}{2} = \binom{i}{2}^{-1}\binom{i}{2} = 1 & \text{if } c_{i'} \prec_c c_i \\ 0 & \text{otherwise} \end{cases}. \quad (23)$$

We call this continuous-time Markov chain as the unvintaged and labeled n -coalescent. The initial state is $c_n = \{\{1\}, \{2\}, \dots, \{n\}\}$ and the final absorbing state is $c_1 = \{\{1, 2, \dots, n\}\}$.

Proof. The proof is merely a consequence of substituting the backward transition probabilities at (20) in the general n -coalescent approximation of (9) since $\{^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ is a lumped Markov chain of $\{^N C^\uparrow(k)\}_{k \in \mathbb{Z}_-}$. \square

The unvintaged and labeled n -coalescent is Kingman's n -coalescent [15, 14] specifically constructed in [14, (Sections 1, 2)]. We retain our nomenclature to emphasize the particularities of the sample genealogical resolution of Kingman's n -coalescent. Figure 2 depicts the coalescent tree space ${}^c_3\mathbb{T}_3 = \mathcal{C}_3 \times [0, \infty)^2$ for the label set $\mathfrak{L} = \{1, 2, 3\}$ with sample size $n = 3$. Thus, elements of ${}^c_3\mathbb{T}_3$ are the sequence of states and their waiting-times visited by the continuous time Markov chain $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$ on \mathbb{C}_n .

Remark 3.8. We can show that $P(c_i)$ can also be obtained from $P(b_i)$ in (13). Since we are not interested in the coalescent vintage of any of our lineages, Equation (17) becomes

$$\binom{|c_{i,1:j-1}| - (j-1) + |c_{i,j}| - 1}{|c_{i,j}| - 1}$$

as we allow any shuffle of the $|c_{i,1:j-1}| - (j-1)$ coalescent events with the $|c_{i,j}| - 1$ coalescent events. Let $i' = \max\{j : m_{i,j} < n\}$. We have,

$$\begin{aligned} P(c_i) &= 2^{n-i} \frac{i!(i-1)!}{n!(n-1)!} \prod_{j=1}^{i'} \frac{|c_{i,j}|!(|c_{i,j}|-1)!}{2^{|c_{i,j}|-1}} \prod_{j=2}^{i'} \binom{|c_{i,1:j-1}| - (j-1) + |c_{i,j}| - 1}{|c_{i,j}| - 1} \\ &= \frac{i!(i-1)!}{n!(n-1)!} \left(\prod_{j=1}^{i'} |c_{i,j}|! \right) (|c_{i,1}| - 1)! \left(\prod_{j=2}^{i'} \frac{(|c_{i,1:j}| - j)!}{(|c_{i,1:j-1}| - (j-1))!} \right) \\ &= \frac{i!(i-1)!}{n!(n-1)!} \left(\prod_{j=1}^{i'} |c_{i,j}|! \right) \left(\frac{\prod_{j=1}^{i'} (|c_{i,1:j}| - j)!}{\prod_{j=1}^{i'-1} (|c_{i,1:j}| - j)!} \right) \\ &= \frac{i!(i-1)!}{n!(n-1)!} \left(\prod_{j=1}^{i'} |c_{i,j}|! \right) (n-i)! \end{aligned}$$

since $(|c_{i,1:i'}| - i')! = (n - (i - i') - i')!$. Therefore, $P(c_i)$ can be obtained from $P(b_i)$, the probability that a vintaged and labeled n -coalescent visits a particular vintaged partition b_i in \mathbb{B}_n^i .

Proposition 3.9 (Bijection between ranked, labeled trees and c -sequences). *There is a bijection between ranked, labeled trees on n leaves and \mathcal{C}_n , the set of c -sequences.*

Proof. It is easy to see that each ranked, labeled tree induces a different c -sequence. Vice versa, any two different c -sequences induce two different ranked, labeled trees. \square

Proposition 3.10 (Forward transition probabilities of a c -sequence). *The transition probability of the forward jump chain $\{C^\downarrow(k)\}_{k \in [n]_+}$ from $c_{i-1} \in \mathcal{C}_n^{i-1}$ to $c_i \in \mathcal{C}_n^i$, where $c_{i-1} \prec_c c_i$, and j, j', j'' are indices such that $c_{i,j} \cup c_{i,j'} = c_{i-1,j''}$, is*

$$P(c_i | c_{i-1}) = \frac{2}{(n - i + 1) \binom{|c_{i,j}| + |c_{i,j'}|}{|c_{i,j}|}}. \quad (24)$$

Proof. For the forward jump chain $\{C^\downarrow(k)\}_{k \in [n]_+}$, first consider the case when $c_{i-1} \prec_c c_i$ with $c_i \in \mathcal{C}_n^i$ and j, j', j'' such that $c_{i,j} \cup c_{i,j'} = c_{i-1,j''} \in c_{i-1}$. Then with Bayes' rule,

$$\begin{aligned} P(c_i | c_{i-1}) &= \frac{P(c_{i-1} | c_i) P(c_i)}{P(c_{i-1})} \\ &= \frac{(n - i)! i! (i - 1)! \prod_{j=1}^i |c_{i,j}|! n! (n - 1)!}{\binom{i}{2} n! (n - 1)! (n - i + 1)! (i - 1)! (i - 2)! \prod_{j=1}^{i-1} |c_{i-1,j}|!} \\ &= \frac{2 \prod_{j=1}^i |c_{i,j}|!}{(n - i + 1) \prod_{j=1}^{i-1} |c_{i-1,j}|!} = \frac{2 |c_{i,j}|! |c_{i,j'}|!}{(n - i + 1) |c_{i-1,j''}|!} \\ &= \frac{2 |c_{i,j}|! |c_{i,j'}|!}{(n - i + 1) (|c_{i,j}| + |c_{i,j'}|)!} = \frac{2}{(n - i + 1) \binom{|c_{i,j}| + |c_{i,j'}|}{|c_{i,j}|}}. \end{aligned}$$

If we do not have $c_{i-1} \prec_c c_i$, then $P(c_i | c_{i-1}) = 0$. \square

Note that we can also obtain the relationship in (22) from the forward transition probabilities in (24):

$$P(c) = \prod_{i=1}^{n-1} P(c_{i+1} | c_i) = \frac{2^{n-1}}{(n - 1)!} \frac{1}{(|c_{2,j}| + |c_{2,j'}|)!} = \frac{2^{n-1}}{(n - 1)! n!}. \quad (25)$$

Remark 3.11. Our forward-time Markov chain $\{C^\downarrow(k)\}_{k \in [n]_+}$ on \mathbb{C}_n is different from Aldous' beta-splitting model [2]. The beta-splitting model also produces bipartitions of a label set forward in time as a *Markov branching* model. The distinguishing feature of the beta-splitting model is its recursive repetition of the same bipartitioning or splitting process anew on elements of a partition of the label set. Therefore the beta-splitting model only induces labeled trees, but no ranking. When the parameter $\beta = 0$, the beta-splitting model induces the same distribution on labeled trees (without ranking) as the vintaged/unvintaged and labeled n -coalescent. In § 4.3 we revisit Aldous' shape statistics that originated under the beta-splitting model from the lumped Markov chains of § 3.5; $\{F^\uparrow(k)\}_{k \in [n]_+}$ and $\{F^\downarrow(k)\}_{k \in [n]_+}$ on \mathbb{F}_n .

Proposition 3.12 (Markov lumping from \mathbb{B}_n to \mathbb{C}_n via \mathcal{C}). *Let the vintage-dropping map $\mathcal{C}(b_j) = c_j : \mathbb{B}_n \rightarrow \mathbb{C}_n$ be the following:*

$$\mathcal{C}(b_j) := \mathcal{C}(\{b_{j,1}, \dots, b_{j,j}\}) := \mathcal{C}(\{c_{j,1}^{\langle m_{j,1} \rangle}, \dots, c_{j,j}^{\langle m_{j,j} \rangle}\}) = \{c_{j,1}, \dots, c_{j,j}\} .$$

The lumped chain, $\{B^{\uparrow \mathcal{C}}(i)\}_{i \in [n]_-}$, of $\{B^\uparrow(i)\}_{i \in [n]_-}$, the jump Markov chain of the vintaged and labeled n -coalescent on \mathbb{B}_n , is Markov and equivalent to $\{C^\uparrow(i)\}_{i \in [n]_-}$, the jump Markov chain of the unvintaged and labeled n -coalescent on \mathbb{C}_n .

Proof. Let c_i, c_j be any two states in \mathbb{C}_n and $\mathcal{C}^{-1}(c_i), \mathcal{C}^{-1}(c_j)$ be their respective inverse images in \mathbb{B}_n . Then, the probability of moving from a state $b_k \in \mathcal{C}^{-1}(c_i)$ to the set $\mathcal{C}^{-1}(c_j)$:

$$P(\mathcal{C}^{-1}(c_j) | b_k) = \sum_{b_{j'} \in \mathcal{C}^{-1}(c_j)} P(b_{j'} | b_k) = \begin{cases} \binom{i}{2}^{-1} & \text{if } b_{j'} \prec_b b_k, b_k \in \mathbb{B}_n^i \\ 0 & \text{otherwise} \end{cases}$$

only depends on b_k through $\mathcal{C}^{-1}(c_i)$ and more specifically through $i = |c_i|$. The Proposition 3.12 follows from Proposition 2.2. \square

3.3 Vintaged and sized n -coalescent

Under $\{D^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and sized n -coalescent, in each state $d_i \in \mathbb{D}_n$, we keep track of the number of descendants of each lineage along with its vintage. Each sequence of visited states or sequential realization of the jump chain, $\{D^\uparrow(k)\}_{k \in [n]_-}$, embedded within $\{D^\uparrow(t)\}_{t \in \mathbb{R}_+}$, induces a ranked tree shape. We will see that there is a bijection between the set of sequential realizations of $\{D^\uparrow(k)\}_{k \in [n]_-}$ and the set of ranked tree shapes. Next, we develop the n -coalescent approximation of the sample genealogy at the resolution of ranked tree shapes.

Consider the coalescent epoch i during which there are i lineages. Let $d_{i,j}$ denote the number of leaves subtended by a lineage during the i -th epoch with coalescent vintage $j = 1, 2, \dots, n-1$. Let $d_{i,n}$ represent the number of leaf lineages (i.e. the number of lineages with coalescent vintage n) during the i -th epoch:

$$d_{i,n} = n - \sum_{j=1}^{n-1} d_{i,j} .$$

Let the number of leaves subtended by the non-leaf lineages during the i -th epoch be vintage-specifically represented by $d_i := (d_{i,1}, d_{i,2}, \dots, d_{i,n-1})$. The state space of such *vintaged* and *sized* ancestral sample lineages during the i -th epoch can be defined by the set,

$$\mathbb{D}_n^i := \left\{ d_i \in \mathbb{Z}_+^{n-1} : \begin{cases} \sum_{j=1}^{i-1} d_{i,j} = 0, \\ \sum_{j=1}^{n-1} \mathbf{1}_{\mathbb{N}}(d_{i,j}) + \left(n - \sum_{j=1}^{n-1} d_{i,j} \right) = i, \\ d_{i,1} \neq 1, d_{i,2} \neq 1, \dots, d_{i,n-1} \neq 1 \end{cases} \right\},$$

with $\mathbb{D}_n := \cup_{i=1}^n \mathbb{D}_n^i$.

Let e_i be the i -th unit vector of length n . The partial ordering \preceq_d of interest on \mathbb{D}_n is based on the immediate precedence relation \prec_d . We say $d_{i'} \prec_d d_i \in \mathbb{D}_n^i$ if and only if:

$$d_{i'} = \begin{cases} d_i + (d_{i,j} + d_{i,k})e_{i-1} - d_{i,j}e_j - d_{i,k}e_k & \text{if } i \leq j < k < n, d_{i,j} \neq 0, d_{i,k} \neq 0 \\ d_i + (d_{i,j} + 1)e_{i-1} - d_{i,j}e_j & \text{if } i \leq j < n, d_{i,j} \neq 0, d_{i,n} \geq 1 \\ d_i + 2e_{i-1} & \text{if } d_{i,n} \geq 2 \end{cases} .$$

A d -sequence $d := (d_n, d_{n-1}, \dots, d_1)$ is an $n \times (n-1)$ matrix:

$$d := \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix} := \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n-1} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n-1,1} & d_{n-1,2} & \cdots & d_{n-1,n-1} \\ d_{n,1} & d_{n,2} & \cdots & d_{n,n-1} \end{pmatrix},$$

that is obtained from a sequence of immediately preceding states in \mathbb{D}_n . Let \mathcal{D}_n be the set of such d -sequences,

$$d \in \mathcal{D}_n := \{d := (d_n, d_{n-1}, \dots, d_1) : d_i \in \mathbb{D}_n^i, d_{i-1} \prec_d d_i, i \in \{2, 3, \dots, n\}\}.$$

The initial state and the final states of the jump chain $\{D^\dagger(k)\}_{k \in [n]_-}$ are

$$d_n = (d_{n,1}, d_{n,2}, \dots, d_{n,n-1}) = (0, 0, \dots, 0) \in \mathbb{D}_n^n \text{ and} \\ d_1 = (d_{1,1}, d_{1,2}, \dots, d_{1,n-1}) = (n, 0, 0, \dots, 0) \in \mathbb{D}_n^1 ,$$

respectively, and \mathcal{D}_n is the set of d -sequences or sequential realizations of this chain on \mathbb{D}_n .

Proposition 3.13 (Backward transition probabilities of a d -sequence). *The transition probabilities of the jump Markov chain $\{D^\dagger(k)\}_{k \in [n]_-}$ on \mathbb{D}_n is:*

$$P(d_{i'}|d_i) = \begin{cases} \binom{d_{i,n}}{d_{i,n}-d_{i',n}} \binom{i}{2}^{-1} & \text{if } d_{i'} \prec_d d_i \in \mathbb{D}_n^i \\ 0 & \text{otherwise} \end{cases} . \quad (26)$$

Proof. The number of leaf lineages that coalesced at the end of epoch i is $d_{i,n} - d_{i',n}$, where $\mathbb{D}_n^{i-1} \ni d_{i'} \prec_d d_i \in \mathbb{D}_n^i$. Note that $(d_{i,n} - d_{i',n}) \in \{0, 1, 2\}$, for any $i \in \{2, 3, \dots, n\}$. Therefore, three type of coalescent events need to be discriminated among the $\binom{i}{2}$ many pairs from i distinct lineages during epoch i . First, when $(d_{i,n} - d_{i',n}) = 0$ we have a coalescent event between two specific non-leaf lineages, each with coalescent vintage smaller than n . Thus, there is exactly $\binom{d_{i,n}}{0} = 1$ such event among $\binom{i}{2}$ possibilities. Second, when $(d_{i,n} - d_{i',n}) = 1$ we have a coalescent event between one specific non-leaf lineage and any one of $d_{i,n}$ many leaf lineages. Thus, there are exactly $\binom{d_{i,n}}{1} = d_{i,n}$ many events among $\binom{i}{2}$ possibilities of the second type. Third, when $(d_{i,n} - d_{i',n}) = 2$ we have a coalescent event between any two of $d_{i,n}$ many leaf lineages. Thus, there are exactly $\binom{d_{i,n}}{2}$ many events among $\binom{i}{2}$ possibilities of the third type. All three types of events are accounted for in (26). \square

Proposition 3.14 (Probability of a d -sequence). *The probability of a d -sequence can be obtained as follows:*

$$P(d) = \frac{2^{n-\mathfrak{J}(d)-1}}{(n-1)!} , \quad (27)$$

where $\mathfrak{J}(g)$ is the number of cherries in d , i.e. the number of times that we have $d_{i,n} - d_{i-1,n} = 2$ as i varies from n to 2. More formally,

$$\mathfrak{J}(d) := \sum_{i=2}^n \mathbf{1}_{\{2\}}(d_{i,n} - d_{i-1,n}) .$$

Note that $P(d)$ has been established in [28, Eqn. 1].

Proof.

$$\begin{aligned}
P(d) &= \prod_{i=n}^2 P(d_{i-1}|d_i) = \prod_{i=n}^2 \binom{d_{i,n}}{d_{i,n} - d_{i-1,n}} \binom{i}{2}^{-1} \\
&= \prod_{i=n}^2 \frac{d_{i,n}!}{d_{i-1,n}!(d_{i,n} - d_{i-1,n})!} \binom{i}{2}^{-1} = d_{n,n}! \prod_{i=n}^2 \frac{1}{(d_{i,n} - d_{i-1,n})!} \binom{i}{2}^{-1} \\
&= n! \left(\prod_{i=n}^2 ((d_{i,n} - d_{i-1,n})!)^{-1} \right) \left(\prod_{i=n}^2 \binom{i}{2}^{-1} \right) \\
&= n! \left(\prod_{j=0,1,2} (j!)^{-\sum_{i=2}^n \mathbf{1}_{\{j\}}(d_{i,n} - d_{i-1,n})} \right) \prod_{i=n}^2 \binom{i}{2}^{-1} \\
&= n! \left(1 \times 1 \times 2^{-\sum_{i=2}^n \mathbf{1}_{\{2\}}(d_{i,n} - d_{i-1,n})} \right) \prod_{i=n}^2 \binom{i}{2}^{-1} \\
&= \frac{n!}{2^{\mathfrak{I}(d)}} \prod_{i=2}^n \binom{i}{2}^{-1} = \frac{2^{n-\mathfrak{I}(d)-1}}{(n-1)!} .
\end{aligned}$$

□

Proposition 3.15 (Bijection between ranked tree shapes and d -sequences). *There is a bijection between ranked tree shapes on n leaves and \mathcal{D}_n , the set of d -sequences.*

Proof. It is easy to see that each ranked tree shape induces a different d -sequence. Vice versa, any two different d -sequences induce two different ranked tree shapes. □

Let $\{^N D^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ be the discrete time sample genealogical Markov chain of n vintaged and unlabeled samples taken at random from the present generation of a Wright-Fisher population of constant size N over the state space \mathbb{D}_n . We derive a continuous-time Markov chain that approximates $\{^N D^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ on \mathbb{D}_n next.

Proposition 3.16 (Vintaged and sized n -coalescent). *The $\lfloor Nt \rfloor$ -step transition probabilities, ${}^N P_{d_i, d_{i'}}(\lfloor Nt \rfloor)$, of the chain $\{^N D^\uparrow(k)\}_{k \in \mathbb{Z}_-}$, converge to the transition probabilities of the continuous-time Markov chain $\{D^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with rate matrix Q , i.e.*

$${}^N P_{d_i, d_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{d_i, d_{i'}}(t) = \exp(Qt),$$

where the entries of Q , $q(d_{i'}|d_i)$, $d_{i'}, d_i \in \mathbb{D}_n$, specifying the transition rate from d_i to $d_{i'}$, are:

$$q(d_{i'}|d_i) = \begin{cases} -\binom{i}{2} & \text{if } d_{i'} = d_i \in \mathbb{D}_n^i \\ \binom{d_{i,n}}{d_{i,n}-d_{i',n}} & \text{if } d_{i'} \prec_g d_i \in \mathbb{D}_n^i \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

The initial state of the chain is $d_n = (0, 0, \dots, 0) \in \mathbb{D}_n^n$ and the final absorbing state is $d_1 = (1, 0, 0, \dots, 0) \in \mathbb{D}_n^1$. This continuous time Markov chain $\{D^\dagger(t)\}_{t \in \mathbb{R}_+}$ on \mathbb{D}_n is called the *vintaged and sized n -coalescent*.

Proof. The proof is merely a consequence of substituting the backward transition probabilities at (26) in the general n -coalescent approximation of (9) since $\{^N H^\dagger(k)\}_{k \in \mathbb{Z}_-}$ is a lumped Markov chain of $\{^N D^\dagger(k)\}_{k \in \mathbb{Z}_-}$. \square

The vintaged and sized n -coalescent gives a novel n -coalescent resolution. Our nomenclature emphasizes the particularities of the sample genealogical resolution of this n -coalescent. In subsequent sections we will see that the vintaged and sized n -coalescent can be lumped into the vintaged and shaped n -coalescent of Tajima as well as to the unvintaged and sized n -coalescent of Kingman. Next we show that the lumping \mathcal{D} from \mathbb{B}_n to \mathbb{D}_n is Markov.

Proposition 3.17 (Markov lumping from \mathbb{B}_n to \mathbb{D}_n via \mathcal{D}). *We define the lumping map $\mathcal{D}(b_k) = d_i : \mathbb{B}_n \rightarrow \mathbb{D}_n$ by*

$$\begin{aligned} \mathcal{D}(b_k) &:= \mathcal{D} \left(\left\{ c_{k,1}^{\langle m_{k,1} \rangle}, \dots, c_{k,k}^{\langle m_{k,k} \rangle} \right\} \right) \\ &= \left(\sum_{j=1}^k |c_{k,j}| \mathbf{1}_{\{1\}}(m_{k,j}), \dots, \sum_{j=1}^k |c_{k,j}| \mathbf{1}_{\{n-1\}}(m_{k,j}) \right). \end{aligned}$$

The lumped chain, $\{B^{\dagger \mathcal{D}}(i)\}_{i \in [n]_-}$, of $\{B^\dagger(i)\}_{i \in [n]_-}$, the jump Markov chain embedded in $\{B^\dagger(t)\}_{t \in \mathbb{R}_+}$, the n -coalescent on \mathbb{B}_n , is Markov and equivalent to $\{D^\dagger(i)\}_{i \in [n]_-}$, the jump Markov chain embedded in $\{D^\dagger(t)\}_{t \in \mathbb{R}_+}$, the vintaged and sized n -coalescent on \mathbb{D}_n .

Proof. Let d_i, d_j be any two states in \mathbb{D}_n and $\mathcal{D}^{-1}(d_i), \mathcal{D}^{-1}(d_j)$ be their respective inverse images in \mathbb{B}_n . Then, the probability of moving from a state $b_{i'} \in \mathcal{D}^{-1}(d_i)$ to

the set $\mathcal{D}^{-1}(d_j)$:

$$\begin{aligned} P(\mathcal{D}^{-1}(d_j)|b_{i'}) &= \sum_{b_{j'} \in \mathcal{D}^{-1}(d_j)} P(b_{j'}|b_{i'}) \\ &= \begin{cases} \binom{d_{i,n}}{d_{i,n}-d_{j,n}} \binom{i}{2}^{-1} & \text{if } d_j \prec_d d_i \in \mathbb{D}_n^i \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

only depends on $b_{i'}$ through $\mathcal{D}^{-1}(d_i)$ and specifically through $d_{i,n}$. Proposition 2.2 completes the proof. \square

Proposition 3.18 (Probability of $d_i \in \mathbb{D}_n^i$). *The probability that the Markov chain $\{D^\dagger(k)\}_{k \in [n]_-}$ visits a state $d_i \in \mathbb{D}_n^i$ is*

$$P(d_i) = \frac{i!(i-1)!}{(n-1)!} \left(\frac{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,j} - 1)(d_{i,1:j} - k_{i,j} - j - 1 + i)!}{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,1:j} - k_{i,j} - m'_{i,j} + i)!} \right), \quad (29)$$

where $d_{i,1:j} := \sum_{k=1}^j d_{i,k}$ and $m'_{i,j} := \min\{k > j : d_{i,k} > 0\}$ and $k_{i,j} := |\{m \leq j : d_{i,m} > 0\}|$.

Proof. We exploit the Markov lumping from \mathbb{B}_n to \mathbb{D}_n (Proposition 3.17) and derive $P(d_i)$ from $P(b_i)$ (Proposition 13), where $d_i \in \mathbb{D}_n^i$ and $b_i \in \mathbb{B}_n^i$ such that dropping the labels in each subset of b_i (but retaining the size and vintage) yields $d_i = \mathcal{D}(b_i)$. We count the number of possible labelings of an element d_i . This is $\frac{n!}{\prod_{j=1}^{n-1} d_{i,j}!}$. We have, with Proposition 13 by multiplying over all epochs ($j = 1 \dots n-1$),

$$\begin{aligned} P(d_i) &= P(b_i) \frac{n!}{\prod_{j=1, d_{i,j} > 0}^{n-1} d_{i,j}!} \\ &= \frac{i!(i-1)!}{n!(n-1)!} \left(\frac{\prod_{j=1, d_{i,j} > 0}^{n-1} d_{i,j}! (d_{i,j} - 1)(d_{i,1:j} - k_{i,j} - 1 - j + i)!}{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,1:j} - k_{i,j} - m'_{i,j} + i)!} \right) \frac{n!}{\prod_{j=1, d_{i,j} > 0}^i d_{i,j}!} \\ &= \frac{i!(i-1)!}{(n-1)!} \left(\frac{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,j} - 1)(d_{i,1:j} - k_{i,j} - j - 1 + i)!}{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,1:j} - k_{i,j} - m'_{i,j} + i)!} \right). \end{aligned}$$

\square

Proposition 3.19 (Forward transition probabilities of a d -sequence). *The transition*

probability of the forward jump chain $\{D^\downarrow(k)\}_{k \in [n]_+}$ from d_{i-1} to d_i is:

$$P(d_i|d_{i-1}) = \begin{cases} \frac{2 \binom{d_{i,n}}{d_{i,n}-d_{i-1,n}} \left(\frac{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,j-1})(d_{i,1:j-k_{i,j}-j-1+i})!}{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,1:j-k_{i,j}-m'_{i,j}+i})!} \right)}{\left(\frac{\prod_{j=1, d_{i-1,j} > 0}^{n-1} (d_{i-1,j-1})(d_{i-1,1:j-k_{i-1,j}-j-2+i})!}{\prod_{j=1, d_{i-1,j} > 0}^{n-1} (d_{i-1,1:j-k_{i-1,j}-m'_{i-1,j}+i-1})!} \right)} & \text{if } d_{i-1} \prec_d d_i \in \mathbb{D}_n^i \\ 0 & \text{otherwise} \end{cases} \quad (30)$$

where $d_{i,1:j} := \sum_{k=1}^j d_{i,k}$ and $m'_{i,j} = \min\{k > j : d_{i,k} > 0\}$.

Proof. $P(d_i|d_{i-1})$, the probability of transition from $d_{i-1} \in \mathbb{D}_n^{i-1}$ to $d_i \in \mathbb{D}_n^i$, where $d_{i-1} \prec_d d_i$, is obtained as follows from (26) and (29) using Bayes' rule,

$$\begin{aligned} P(d_i|d_{i-1}) &= P(d_{i-1}|d_i) \frac{P(d_i)}{P(d_{i-1})} \\ &= \frac{\binom{d_{i,n}}{d_{i,n}-d_{i-1,n}} \binom{i}{2}^{-1} \frac{i!(i-1)!}{(n-1)!} \left(\frac{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,j-1})(d_{i,1:j-k_{i,j}-j-1+i})!}{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,1:j-k_{i,j}-m'_{i,j}+i})!} \right)}{\frac{(i-1)!(i-2)!}{(n-1)!} \left(\frac{\prod_{j=1, d_{i-1,j} > 0}^{n-1} (d_{i-1,j-1})(d_{i-1,1:j-k_{i-1,j}-j-2+i})!}{\prod_{j=1, d_{i-1,j} > 0}^{n-1} (d_{i-1,1:j-k_{i-1,j}-m'_{i-1,j}+i-1})!} \right)} \\ &= \frac{2 \binom{d_{i,n}}{d_{i,n}-d_{i-1,n}} \left(\frac{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,j-1})(d_{i,1:j-k_{i,j}-j-1+i})!}{\prod_{j=1, d_{i,j} > 0}^{n-1} (d_{i,1:j-k_{i,j}-m'_{i,j}+i})!} \right)}{\left(\frac{\prod_{j=1, d_{i-1,j} > 0}^{n-1} (d_{i-1,j-1})(d_{i-1,1:j-k_{i-1,j}-j-2+i})!}{\prod_{j=1, d_{i-1,j} > 0}^{n-1} (d_{i-1,1:j-k_{i-1,j}-m'_{i-1,j}+i-1})!} \right)}. \end{aligned}$$

And if $d_{i-1} \not\prec_d d_i$ then $P(d_i|d_{i-1}) = 0$. \square

3.4 Vintaged and shaped n -coalescent

We have seen that there is a bijection between the set of c -sequences and the set of ranked, labeled trees. Another set of interest is that of the *evolutionary relationships* of Tajima [28, Figures 1-3], which are ranked tree shapes in our terms. In this section, we develop $\{G^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and shaped n -coalescent of Tajima via $\{G^\uparrow(k)\}_{k \in [n]_-}$, its embedded jump chain. We will see that there is a bijection between \mathcal{G}_n , the set of sequential realizations of $\{G^\uparrow(k)\}_{k \in [n]_-}$, and the set of ranked tree shapes or Tajima's evolutionary relationships. Note that we already established a bijection from the set of sequential realizations of $\{D^\uparrow(k)\}_{k \in [n]_-}$, the jump chain of the vintaged and sized n -coalescent, to the set of ranked tree shapes. However, \mathbb{G}_n , the state space of $\{G^\uparrow(k)\}_{k \in [n]_-}$, is significantly smaller than \mathbb{D}_n , the state space of

$\{D^\dagger(k)\}_{k \in [n]_-}$. Therefore, it is preferable to use the vintaged and shaped n -coalescent for inference if it adequately describes the hidden genealogical space up to equivalence classes of ranked tree shapes.

Consider the coalescent epoch i during which there are i lineages. Let $g_{i,j}$ denote the presence ($g_{i,j} = 1$) or absence ($g_{i,j} = 0$) of a lineage during the i -th epoch with coalescent vintage $j = 1, 2, \dots, n-1$. Define the set of such vintaged and shaped ancestral lineages of our unlabeled sample of size n , during the i -th coalescent epoch by,

$$\mathbb{G}_n^i := \left\{ g_i \in \{0, 1\}^{n-1} : g_{i,i} = 1, \sum_{j=1}^{i-1} g_{i,j} = 0, \sum_{j=1}^{n-1} g_{i,j} \leq i \right\},$$

with $\mathbb{G}_n := \cup_{i=1}^n \mathbb{G}_n^i$. We interpret the vector $g_i \in \mathbb{G}_n^i$ in the i -th epoch as follows. The component $g_{i,i} = 1$ represents the lineage that just arose at the beginning of the i -th epoch. The component with $g_{i,j} = 1$, for $i < j < n$, represents the presence of the lineage with coalescent vintage j . The vertices of the unit $(n-1)$ -dimensional hypercube contain \mathbb{G}_n . We count the elements in \mathbb{G}_n^i and \mathbb{G}_n next.

Proposition 3.20. *The number of elements in \mathbb{G}_n^i is, for $i < n$,*

$$|\mathbb{G}_n^i| = \sum_{k=0}^{i-1} \binom{n-i-1}{k}. \quad (31)$$

For $i = n$, we have $|\mathbb{G}_n^n| = 1$.

Proof. For $i = n$, we only have one element, a sequence of only 0s, i.e. $|\mathbb{G}_n^n| = 1$. Now let $i < n$. Let $g_i \in \mathbb{G}_n^i$. Since we have i lineages in epoch i , we have at the most i non-zero entries in g_i . In g_i , we have $g_{i,j} = 0$ for $j = 1, \dots, i-1$. Further, $g_{i,i} = 1$. The remaining $n-1-i$ elements are 0 or 1. For k non-zero entries in the remaining elements, we have $\binom{n-1-i}{k}$ possibilities to assign the 0s and 1s. Summing over all possible k -values yields (31). \square

Proposition 3.21. *The number of elements in \mathbb{G}_n is*

$$|\mathbb{G}_n| = \text{Fibo}(n+1), \quad (32)$$

where $\text{Fibo}(n)$ is the n -th Fibonacci number.

Proof. From Proposition 3.20 we have, by summing over all i ,

$$|\mathbb{G}_n| = \sum_{i=1}^n |\mathbb{G}_n^i| = \sum_{i=1}^{n-1} \sum_{k=0}^{i-1} \binom{n-i-1}{k} + 1.$$

By basic properties of the binomial coefficient, we get,

$$\begin{aligned} \sum_{i=1}^{n-1} \sum_{k=0}^{i-1} \binom{n-i-1}{k} &= \sum_{k=0}^{n-2} \sum_{j=k}^{n-2-k} \binom{j}{k} = \sum_{k=0}^{n-2} \binom{n-k-1}{k+1} \\ &= \sum_{k=1}^{n-1} \binom{n-k}{k} = \text{Fibo}(n+1) - 1 \end{aligned} \quad (33)$$

which proves the proposition. \square

Lemma 3.22. *Let the jump Markov chain $\{G^\uparrow(k)\}_{k \in [n]_-}$ be at state g_i . Then the number of leaves not having coalesced by epoch i is*

$$g_{i,n} = i - \sum_{j=1}^{n-1} g_{i,j} . \quad (34)$$

Proof. The proof is by induction on i . In epoch $i = n - 1$, two leaves are coalescing, i.e. we have $n - 2$ remaining leaves. Due to (34), we get, $g_{n-1,n} = n - 1 - \sum_{j=1}^{n-1} g_{n-1,j} = n - 1 - 1 = n - 2$.

Now assume that (34) holds for all $i > k$. Then, for $g_{k,n}$ we have to consider three cases:

(i) g_k is the result of the coalescence of two leaves in g_{k+1} . By the induction assumption, we have $g_{k+1,n} = k + 1 - \sum_{j=1}^{n-1} g_{k+1,j}$. Since two leaves are coalescing, we have $g_{k,n} = g_{k+1,n} - 2$. Further, $g_{k,k} = 1$, $g_{k+1,k} = 0$ and $g_{k,j} = g_{k+1,j}$ for $k < j < n$. So,

$$g_{k,n} = g_{k+1,n} - 2 = k + 1 - \sum_{j=1}^{n-1} g_{k+1,j} - 2 = k - 1 - \sum_{j=1}^{n-1} g_{k,j} + 1 = k - \sum_{j=1}^{n-1} g_{k,j}.$$

(ii) g_k is the result of the coalescence of one leaf and a non-leaf component in g_{k+1} . By the induction assumption, we have $g_{k+1,n} = k + 1 - \sum_{j=1}^{n-1} g_{k+1,j}$. Since one leaf is coalescing, we have $g_{k,n} = g_{k+1,n} - 1$. Further, $g_{k,k} = 1$, $g_{k+1,k} = 0$. Assume that component which evolved in epoch j^* is coalescing with the leaf. Then $g_{k,j} = g_{k+1,j}$ for $k < j < n$, $j \neq j^*$ and $g_{k,j^*} = 0$. So,

$$g_{k,n} = g_{k+1,n} - 1 = k + 1 - \sum_{j=1}^{n-1} g_{k+1,j} - 1 = k - \sum_{j=1}^{n-1} g_{k,j}.$$

(iii) g_k is the result of the coalescence of two non-leaf component in g_{k+1} . By the induction assumption, we have $g_{k+1,n} = k + 1 - \sum_{j=1}^{n-1} g_{k+1,j}$. Since no

leaf is coalescing, we have $g_{k,n} = g_{k+1,n}$. Further, $g_{k,k} = 1$, $g_{k+1,k} = 0$. Further, $\sum_{j=k+1}^{n-1} g_{k+1,j} = \sum_{j=k+1}^{n-1} g_{k,j} + 1$. So,

$$g_{k,n} = g_{k+1,n} - 2 = k + 1 - \sum_{j=1}^{n-1} g_{k+1,j} = k - \sum_{j=1}^{n-1} g_{k,j}.$$

□

Let e_i be the i -th unit vector of length n . The partial ordering $\ddot{\prec}_g$ of interest on \mathbb{G}_n is based on the immediate precedence \prec_g . We say $g_{i'} \prec_g g_i \in \mathbb{G}_n^i$ if and only if

$$g_{i'} = \begin{cases} g_i + e_{i-1} - e_j - e_k & \text{if } i \leq j < k < n, g_{i,j} = g_{i,k} = 1 \\ g_i + e_{i-1} - e_j & \text{if } i \leq j < n, g_{i,j} = 1, g_{i,n} \geq 1 \\ g_i + e_{i-1} & \text{if } g_{i,n} \geq 2 \end{cases}$$

A g -sequence $g := (g_n, g_{n-1}, \dots, g_1)$ is an $n \times (n-1)$ matrix:

$$g := \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{n-1} \\ g_n \end{pmatrix} := \begin{pmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,n-1} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n-1,1} & g_{n-1,2} & \cdots & g_{n-1,n-1} \\ g_{n,1} & g_{n,2} & \cdots & g_{n,n-1} \end{pmatrix}$$

that is obtained from a sequence of immediately preceding states in \mathbb{G}_n . Examples of g -sequences when $n = 3$ and 4 are depicted in Table 2. Let \mathcal{G}_n be the set of such g -sequences:

$$g \in \mathcal{G} := \{g := (g_n, g_{n-1}, \dots, g_1) : g_i \in \mathbb{G}_n^i, g_{i-1} \prec_g g_i, i \in \{2, 3, \dots, n\}\}.$$

Proposition 3.23 (Backward transition probabilities of a g -sequence). *The transition probability of the jump Markov chain $\{G^\uparrow(k)\}_{k \in [n]_-}$ on \mathbb{G}_n is*

$$P(g_{i'} | g_i) = \begin{cases} \binom{g_{i,n}}{g_{i,n} - g_{i',n}} \binom{i}{2}^{-1} & \text{if } g_{i'} \prec_g g_i \in \mathbb{G}_n^i, \\ 0 & \text{otherwise} \end{cases}, \quad (35)$$

where $g_{i,n}$ is the number of leaves that have not coalesced by epoch i , as derived in (34) of Lemma 3.22. The initial state of the chain is $g_n = (0, 0, \dots, 0) \in \mathbb{G}_n^n$ and the final absorbing state is $g_1 = (1, 0, 0, \dots, 0) \in \mathbb{G}_n^1$.

Proof. It is identical to that of d -sequence transition probabilities in (26). \square

Proposition 3.24 (Probability of a g -sequence). *The probability of a g -sequence can be obtained as follows:*

$$\begin{aligned} P(g) &= \prod_{i=n}^2 P(g_{i-1}|g_i) = \prod_{i=n}^2 \binom{g_{i,n}}{g_{i,n} - g_{i-1,n}} \binom{i}{2}^{-1} \\ &= \frac{n!}{2^{\mathfrak{J}(g)}} \prod_{i=2}^n \binom{i}{2}^{-1} = \frac{2^{n-\mathfrak{J}(g)-1}}{(n-1)!} , \end{aligned} \quad (36)$$

where $\mathfrak{J}(g)$ is the number of cherries in g , i.e. the number of times that we have $g_{i,n} - g_{i-1,n} = 2$ as i varies from n to 2. More formally,

$$\mathfrak{J}(g) := \sum_{i=2}^n \mathbf{1}_{\{2\}}(g_{i,n} - g_{i-1,n}) .$$

Note that $P(g)$ has been established in [28, Eqn. 1].

Proof. The proof is similar to that of Proposition 3.14. \square

Proposition 3.25 (Bijection between ranked, labeled trees and g -sequences). *There is a bijection between the set of ranked tree shapes on n leaves and \mathcal{G}_n , the set of g -sequences.*

Proof. It is easy to see that each ranked tree shape induces a different g -sequence. Vice versa, any two different g -sequences induce two different ranked tree shapes. \square

Let $\{^N G^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ be the discrete time sample genealogical Markov chain of n vintaged and unlabeled samples taken at random from the present generation of a Wright-Fisher population of constant size N over the state space \mathbb{G}_n . We derive a continuous-time Markov chain that approximates $\{^N G^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ on \mathbb{G}_n next.

Proposition 3.26 (Vintaged and shaped n -coalescent). *The $\lfloor Nt \rfloor$ -step transition probabilities, $^N P_{g_i, g_{i'}}(\lfloor Nt \rfloor)$, of the chain $\{^N G^\uparrow(k)\}_{k \in \mathbb{Z}_-}$, converge to the transition probabilities of the continuous-time Markov chain $\{G^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with rate matrix Q , i.e.*

$$^N P_{g_i, g_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{g_i, g_{i'}}(t) = \exp(Qt),$$

where the entries of Q , $q(g_{i'}|g_i)$, $g_{i'}, g_i \in \mathbb{G}_n$, specifying the transition rate from g_i to $g_{i'}$, are:

$$q(g_{i'}|g_i) = \begin{cases} -\binom{i}{2} & \text{if } g_{i'} = g_i \in \mathbb{G}_n^i \\ \binom{g_{i,n}}{g_{i,n}-g_{i',n}} & \text{if } g_{i'} \prec_g g_i \in \mathbb{G}_n^i \\ 0 & \text{otherwise} \end{cases} \quad (37)$$

The initial state of the chain is $g_n = (0, 0, \dots, 0) \in \mathbb{G}_n^n$ and the final absorbing state is $g_1 = (1, 0, 0, \dots, 0) \in \mathbb{G}_n^1$. This continuous time Markov chain $\{G^\uparrow(t)\}_{t \in \mathbb{R}_+}$ on \mathbb{G}_n is called the *vintaged and shaped n -coalescent*.

Proof. The proof is merely a consequence of substituting the backward transition probabilities at (35) in the general n -coalescent approximation of (9) since $\{^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ is a lumped Markov chain of $\{^N G^\uparrow(k)\}_{k \in \mathbb{Z}_-}$. \square

The genealogical resolution of the vintaged and shaped n -coalescent is Tajima's evolutionary relationships. We sometimes call the vintaged and shaped n -coalescent as Tajima's n -coalescent. Next we show that the lumping \mathcal{G} from \mathbb{D}_n to \mathbb{G}_n is Markov.

Proposition 3.27 (Markov lumping from \mathbb{D}_n to \mathbb{G}_n via \mathcal{G}). *Consider the following size-dropping map $\mathcal{G}(d_k) = g_h : \mathbb{D}_n \rightarrow \mathbb{G}_n$:*

$$\mathcal{G}(d_k) := \mathcal{G}((d_{k,1}, \dots, d_{k,n})) = (\mathbf{1}_N(d_{k,1}), \dots, \mathbf{1}_N(d_{k,n-1})) = (g_{h,1}, \dots, g_{h,n-1})$$

The lumped chain, $\{D^{\uparrow \mathcal{G}}(i)\}_{i \in [n]_-}$, of $\{D^\uparrow(i)\}_{i \in [n]_-}$, the jump Markov chain embedded in $\{D^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and sized n -coalescent on \mathbb{D}_n , is Markov and equivalent to $\{G^\uparrow(i)\}_{i \in [n]_-}$, the jump Markov chain embedded in $\{G^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and shaped n -coalescent on \mathbb{G}_n .

Proof. Let g_i, g_j be any two states in \mathbb{G}_n and $\mathcal{G}^{-1}(g_i), \mathcal{G}^{-1}(g_j)$ be their respective inverse images in \mathbb{D}_n . Then, the probability of moving from a state $d_{i'} \in \mathcal{G}^{-1}(g_i)$ to the set $\mathcal{G}^{-1}(g_j)$:

$$\begin{aligned} P(\mathcal{G}^{-1}(g_j)|d_{i'}) &= \sum_{d_{j'} \in \mathcal{G}^{-1}(g_j)} P(d_{j'}|d_{i'}) = \begin{cases} \binom{g_{i,n}}{g_{i,n}-g_{j,n}} \binom{i}{2}^{-1} & \text{if } g_j \prec_g g_i, g_i \in \mathbb{G}_n^i \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

only depends on $d_{i'}$ through g_i and specifically through $g_{i,n}$. Proposition 2.2 completes the proof. \square

The probability that $g_i \in \mathbb{G}_n^i$ is visited by the chain is obtained by considering the inverse images, $\mathcal{G}^{-1}(g_i)$:

$$P(g_i) = P(\mathcal{G}^{-1}(g_i)) = \sum_{d_j \in \mathcal{G}^{-1}(g_i)} P(d_j).$$

with $P(d_j)$ from Proposition 29. The probability $P(g_i) = P(\mathcal{G}^{-1}(g_i))$ can be written explicitly as follows. Let $L = i - g_{i,n}$, which is the number of non-leaf lineages in epoch i . Let $\mathfrak{f}(g_i, j_1, \dots, j_L) = d_i \in \mathbb{D}_i^n$ where $d_{i,j} = 0$ if and only if $g_{i,j} = 0$, $d_{i,n} = g_{i,n}$ and $d_{i,j} = j_k$ if and only if $g_{i,j}$ is the k -th entry which is bigger than zero. The probability $P(g_i)$ is,

$$P(g_i) = \sum_{j_1=2}^{n-g_{i,n}-2(L-1)} \sum_{j_2=2}^{n-g_{i,n}-2(L-2)-j_1} \dots \sum_{j_{L-1}=2}^{n-g_{i,n}-2-\sum_{i=1}^{L-2} j_i} P(\mathfrak{f}(g_i, j_1, \dots, j_{L-1}, n - g_{i,n} - \sum_{i=1}^{L-2} j_i)) . \quad (38)$$

Finally, the transition probabilities of the forward jump chain $\{G^\downarrow(k)\}_{k \in [n]_+}$ can be obtained from Bayes' rule as follows:

$$P(g_i | g_{i-1}) = \begin{cases} P(g_{i-1} | g_i) \frac{P(g_i)}{P(g_{i-1})} & \text{if } g_{i-1} \prec_g g_i \in \mathbb{G}_n^i \\ 0 & \text{otherwise.} \end{cases} .$$

3.5 Unvintaged and sized n -coalescent

The unvintaged and sized n -coalescent is mentioned as a lumped Markov chain of the unvintaged and labeled n -coalescent and termed the ‘label-killed’ process by Kingman [15, 5.2]. Tavaré [30, p. 136-137] terms the unvintaged and sized n -coalescent as the ‘family-size process’ as part of the nomenclature of a more general birth-death-immigration process [13]. The transition probabilities of this Markov process are not explicitly developed in [15] or [30]. They have been developed in [23] into $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the unvintaged and sized n -coalescent. It is shown in [23, 22] that $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$ resolves the hidden genealogy space just enough to prescribe the likelihood of site frequency spectrum and its linear summaries. We briefly retrace $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$ and its embedded jump chain $\{F^\uparrow(k)\}_{k \in [n]_-}$ and show that they can provide the sampling distribution of a large family of shape statistics including several classical ones. The significantly smaller state space of $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$ allows for a computationally efficient and statistically sufficient inference based on these statistics.

Consider the coalescent epoch at which there are i lineages. Let $f_{i,j}$ denote the number of lineages subtending j leaves, i.e. the frequency of lineages that are ancestral to j samples, at this epoch. Let us summarize these frequencies from the i lineages as j varies over its support by $f_i := (f_{i,1}, f_{i,2}, \dots, f_{i,n})$. Then the space of f_i 's is defined by,

$$\mathbb{F}_n^i := \left\{ f_i := (f_{i,1}, f_{i,2}, \dots, f_{i,n}) \in \mathbb{Z}_+^n : \sum_{j=1}^n j f_{i,j} = n, \sum_{j=1}^n f_{i,j} = i \right\} .$$

Let the set of such frequencies over all epochs be $\mathbb{F}_n := \bigcup_{i=1}^n \mathbb{F}_n^i$. Note that \mathbb{F}_n contains the frequency of the cardinalities of sets belonging to every element of \mathbb{C}_n , the state space of $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the unvintaged and labeled n -coalescent. Thus, \mathbb{F}_n is the frequency representation of the integer partitions of n , i.e. the solutions to the Diophantine equation $\{(p_1, p_2, \dots, p_n) \in \mathbb{Z}_+^n : \sum_{i=1}^n i p_i = n\}$, and \mathbb{F}_n^i are those integer partitions composed of i positive integers. Thus, the cardinality of \mathbb{F}_n is the number of integer partitions of n :

$$|\mathbb{F}_n| = 1 + \sum_{k=1}^{\lfloor n/2 \rfloor} \mathfrak{p}(k, n-k), \quad \text{where}$$

$$\mathfrak{p}(k, n) = \begin{cases} 0 & \text{if } k > n \\ 1 & \text{if } k = n \\ \mathfrak{p}(k+1, n) + \mathfrak{p}(k, n-k) & \text{otherwise} \end{cases} . \quad (39)$$

Let us define an f -sequence f as follows:

$$f := (f_n, f_{n-1}, \dots, f_1) \in \mathcal{F}_n := \{f : f_i \in \mathbb{F}_n^i, f_{i-1} \prec_f f_i, \forall i \in \{2, \dots, n\}\},$$

where \prec_f is the immediate precedence relation that induces the partial ordering $\ddot{\prec}_f$ on \mathbb{F}_n . It is defined by denoting the j -th unit vector of length n by e_j , as follows:

$$f_{i'} \prec_f f_i \Leftrightarrow f_{i'} = f_i - e_j - e_k + e_{j+k} .$$

Thus, \mathcal{F}_n is the set of f -sequences with n samples. One can see \mathcal{F}_n as the set of the frequencies of the cardinalities of c -sequences in \mathbb{C}_n . Recall the c -sequence $c = (c_n, c_{n-1}, \dots, c_1)$, where $c_{i-1} \prec_c c_i$, $c_{i-1} \in \mathbb{C}_n^{i-1}$, $c_i \in \mathbb{C}_n^i$, and $c_i := (c_{i,1}, c_{i,2}, \dots, c_{i,i})$

contains its canonically ordered i subsets. Then the corresponding state space lumping map $\mathcal{F}(c_i) = f_i : \mathbb{C}_n \rightarrow \mathbb{F}_n$ and the sequence map $\underline{\mathcal{F}}(c) = f : \mathbb{C}_n \rightarrow \mathcal{F}_n$ are:

$$\mathcal{F}(c_i) := \left(\sum_{h=1}^i \mathbf{1}_{\{1\}}(|c_{i,h}|), \dots, \sum_{h=1}^i \mathbf{1}_{\{n\}}(|c_{i,h}|) \right),$$

$$\underline{\mathcal{F}}(c) := (\mathcal{F}(c_n), \dots, \mathcal{F}(c_1)) . \quad (40)$$

An f -sequence f written as $(f_n, f_{n-1}, \dots, f_1)$ is an $n \times n$ matrix:

$$f := \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} := \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n-1} & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n-1} & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{n-1,1} & f_{n-1,2} & \cdots & f_{n-1,n-1} & f_{n-1,n} \\ f_{n,1} & f_{n,2} & \cdots & f_{n,n-1} & f_{n,n} \end{pmatrix}$$

Note that \mathcal{F}_n indexes an equivalence class in \mathbb{C}_n via the inverse map $\underline{\mathcal{F}}^{-1}$ at (40).

Having defined f -sequences and their associated sets, we are ready to define $\{F^\uparrow(k)\}_{k \in [n]_-}$, the jump Markov chain of the unvintaged and sized n -coalescent on \mathbb{F}_n . Equations (41), (42), (43), (44) and (45) have been derived in [23]. The transition probability of $\{F^\uparrow(k)\}_{k \in [n]_-}$ from $f_i \in \mathbb{F}_n^i$ to $f_{i-1} \in \mathbb{F}_n^{i-1}$ is:

$$P(f_{i-1}|f_i) = \begin{cases} f_{i,j} f_{i,k} \binom{i}{2}^{-1} & \text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k}, j \neq k \\ \binom{f_{i,j}}{2} \binom{i}{2}^{-1} & \text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k}, j = k \\ 0 & \text{otherwise} \end{cases} . \quad (41)$$

The initial state and the final absorbing state of $\{F^\uparrow(k)\}_{k \in [n]_-}$ on \mathbb{F}_n are $f_n = (n, 0, \dots, 0)$ and $f_1 = (0, 0, \dots, 1)$, respectively. The probability of an f -sequence, $f := (f_n, f_{n-1}, \dots, f_1) \in \mathcal{F}_n$, is given by the product:

$$P(f) = \prod_{i=n}^2 P(f_{i-1}|f_i), \quad (42)$$

and the probability that $\{F^\uparrow(k)\}_{k \in [n]_-}$ visits a particular $f_i \in \mathbb{F}_n^i$ at the i -th epoch [30, Equation (7.11)] is:

$$P(f_i) = \frac{i!}{\prod_{j=1}^i f_{i,j}!} \binom{n-1}{i-1}^{-1} \quad (43)$$

Let us consider the forward time jump chain $\{F^\downarrow(k)\}_{k \in [n]_+}$ on \mathbb{F}_n . The transition probability of $\{F^\downarrow(k)\}_{k \in [n]_+}$ from $f_{i-1} \in \mathbb{F}_n^{i-1}$ to $f_i \in \mathbb{F}_n^i$ is:

$$P(f_i | f_{i-1}) = \begin{cases} 2f_{i-1, j+k}(n-i+1)^{-1} & \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k}, j \neq k, \\ & j+k > 1, f_i \in \mathbb{F}_n^i, f_{i-1} \in \mathbb{F}_n^{i-1} \\ f_{i-1, j+k}(n-i+1)^{-1} & \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k}, j = k, \\ & j+k > 1, f_i \in \mathbb{F}_n^i, f_{i-1} \in \mathbb{F}_n^{i-1} \\ 0 & \text{otherwise} \end{cases} \quad (44)$$

The final absorbing state and the initial state of $\{F^\uparrow(k)\}_{k \in [n]_-}$ on \mathbb{F}_n are $f_n = (n, 0, \dots, 0)$ and $f_1 = (0, 0, \dots, 1)$, respectively. The probability of an f -sequence, $f := (f_n, f_{n-1}, \dots, f_1) \in \mathcal{F}_n$, is given by the product:

$$P(f) = \prod_{i=2}^n P(f_i | f_{i-1}), \quad (45)$$

Let $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ be the discrete time sample genealogical Markov chain of n unvintaged and unlabeled samples taken at random from the present generation of a Wright-Fisher population of constant size N over the state space \mathbb{F}_n . We derive a continuous-time Markov chain that approximates $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ on \mathbb{G}_n next.

Proposition 3.28 (Unvintaged and sized n -coalescent). *The $\lfloor Nt \rfloor$ -step transition probabilities, $^N P_{f_i, f_{i'}}(\lfloor Nt \rfloor)$, of the chain $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$, converge to the transition probabilities of the continuous-time Markov chain $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with rate matrix Q , i.e.*

$$^N P_{f_i, f_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{f_i, f_{i'}}(t) = \exp(Qt),$$

where the entries of Q , $q(f_{i'} | f_i)$, $f_{i'}, f_i \in \mathbb{F}_n$, specifying the transition rate from $f_i \in \mathbb{F}_n^i$ to $f_{i'}$, are:

$$q(f_{i'} | f_i) = \begin{cases} -i(i-1)/2 & \text{if } f_i = f_{i'}, f_i \in \mathbb{F}_n^i \\ f_{i,j} f_{i,k} & \text{if } f_{i'} = f_i - e_j - e_k + e_{j+k}, j \neq k, f_i \in \mathbb{F}_n^i, f_{i'} \in \mathbb{F}_n^{i-1} \\ (f_{i,j})(f_{i,j} - 1)/2 & \text{if } f_{i'} = f_i - e_j - e_k + e_{j+k}, j = k, f_i \in \mathbb{F}_n^i, f_{i'} \in \mathbb{F}_n^{i-1} \\ 0 & \text{otherwise} \end{cases} \quad (46)$$

The initial state is $f_n = (n, 0, 0, \dots, 0)$ and the final absorbing state is $f_1 = (0, 0, \dots, 1)$. This continuous time Markov chain $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$ on \mathbb{F}_n is called the unvintaged and sized n -coalescent.

Proof. The proof is merely a consequence of substituting the backward transition probabilities at (41) in the general n -coalescent approximation of (9) since $\{^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ is a lumped Markov chain of $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$. \square

Next we show that the lumping \mathcal{F} from \mathbb{C}_n to \mathbb{F}_n as well as the lumping \mathcal{F}' from \mathbb{D}_n to \mathbb{F}_n are Markov.

Proposition 3.29 (Markov lumping from \mathbb{C}_n to \mathbb{F}_n via \mathcal{F}). *Our lumping of the unvintaged and labeled n -coalescent over \mathbb{C}_n to the unvintaged and sized n -coalescent over \mathbb{F}_n , via the mapping $\mathcal{F}(c_i) = f_i : \mathbb{C}_n \rightarrow \mathbb{F}_n$ in (40), is Markov as pointed out by Kingman [15, (5.1),(5.2)] using the arguments in [20, § IIIId].*

Proof. Let f_i, f_j be any two states in \mathbb{F}_n and $\mathcal{F}^{-1}(f_i), \mathcal{F}^{-1}(f_j)$ be their respective inverse images in \mathbb{C}_n . Then, the probability of moving from a state $c_{j'} \in \mathcal{F}^{-1}(f_j)$ to the set $\mathcal{F}^{-1}(f_i)$:

$$\sum_{c_{j'} \in \mathcal{F}^{-1}(f_j)} P(c_{j'} | c_{i'}) = \begin{cases} f_{i,\ell} f_{i,k} \binom{i}{2}^{-1} & \text{if } f_j = f_i - e_\ell - e_k + e_{\ell+k}, \ell \neq k \\ \binom{f_{i,\ell}}{2} \binom{i}{2}^{-1} & \text{if } f_j = f_i - e_\ell - e_k + e_{\ell+k}, \ell = k \\ 0 & \text{otherwise} \end{cases},$$

depends on $c_{j'}$ only through $f_i = \mathcal{F}(c_{i'})$. For any given $f_i, f_j \in \mathbb{F}_n$, this condition is satisfied by construction, since the above sum equals $P(f_j | f_i)$ at (41), a quantity that only depends on f_i . \square

Proposition 3.30 (Markov lumping from \mathbb{D}_n to \mathbb{F}_n via \mathcal{F}'). *Consider the following vintage-dropping map $\mathcal{F}'(d_k) = f_i : \mathbb{D}_n \rightarrow \mathbb{F}_n$:*

$$\begin{aligned} \mathcal{F}'(d_k) &:= \mathcal{F}'((d_{k,1}, \dots, d_{k,n})) \\ &= \left(n - \sum_{j=1}^{n-1} d_{k,j}, \sum_{j=1}^{n-1} \mathbf{1}_{\{2\}}(d_{k,i}), \dots, \sum_{j=1}^{n-1} \mathbf{1}_{\{n\}}(d_{k,i}) \right) = (f_{i,1}, f_{i,2}, \dots, f_{i,n}) . \end{aligned}$$

The lumped chain, $\{D^{\uparrow \mathcal{F}'}(i)\}_{i \in [n]_-}$, of $\{D^\uparrow(i)\}_{i \in [n]_-}$, the jump Markov chain embedded in $\{D^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and sized n -coalescent on \mathbb{D}_n , is Markov and equivalent to $\{F^\uparrow(i)\}_{i \in [n]_-}$, the jump Markov chain embedded in $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$, the vintaged and shaped n -coalescent on \mathbb{F}_n .

Proof. Let f_i, f_j be any two states in \mathbb{F}_n and $\mathcal{F}'^{-1}(f_i), \mathcal{F}'^{-1}(f_j)$ be their respective inverse images in \mathbb{D}_n . Then, the probability of moving from a state $d_{j'} \in \mathcal{F}'^{-1}(f_j)$ to

the set $\mathcal{F}'^{-1}(f_j)$:

$$\begin{aligned} P(\mathcal{F}'^{-1}(f_j)|d_{i'}) &= \sum_{d_{j'} \in \mathcal{F}'^{-1}(f_j)} P(d_{j'}|d_{i'}) \\ &= \begin{cases} f_{i,\ell} f_{i,k} \binom{i}{2}^{-1} & \text{if } f_j = f_i - e_\ell - e_k + e_{\ell+k}, \ell \neq k \\ \binom{f_{i,\ell}}{2} \binom{i}{2}^{-1} & \text{if } f_j = f_i - e_\ell - e_k + e_{\ell+k}, \ell = k \\ 0 & \text{otherwise} \end{cases} . \end{aligned}$$

only depends on $d_{i'}$ through f_i for any given f_j . Proposition 2.2 completes the proof. \square

Next we define a shape statistic triple of any $f \in \mathcal{F}_n$. Let us denote the entry-wise maximum or minimum of a vector x by $\max\langle x \rangle$ and $\min\langle x \rangle$, respectively. There are $n - 1$ coalescence events in any f . Define $\mathfrak{J}(f)$ as the number of events resulting from the coalescence of a pair of leaves or samples. Such an event is also said to be a cherry. Next define $\mathfrak{T}(f)$ as the number of events that arise from coalescing two sets of distinct sizes. Let the number of the remaining events in f be defined as $\widehat{\mathfrak{J}}(f)$. Thus, $\widehat{\mathfrak{J}}(f)$ is the number of events resulting from the coalescence of two sets of equal size that are not cherries. A distinctly-sized split of a lineage subtending i leaves gives rise to two lineages subtending i_1 and i_2 leaves, such that $i_1 \neq i_2$ and $i = i_1 + i_2$. In formulae, the above is,

$$\mathfrak{J}(f) := \sum_{i=2}^n \mathbf{1}_{\{1\}}(f_{i-1,2} - f_{i,2}) \quad (47)$$

$$\mathfrak{T}(f) := \sum_{i=2}^n \mathbf{1}_{\{1\}}(\max\langle f_i - f_{i-1} \rangle) \quad (48)$$

$$\widehat{\mathfrak{J}}(f) := n - 1 - \mathfrak{T}(f) - \mathfrak{J}(f) \quad (49)$$

Denoting the entry-wise or Hadamard product by \boxtimes , let us define \ddot{f}_i as the frequency of lineages that subtend the same number of leaves as the lineage that was split at the beginning of the i -th epoch (forward in time) and the corresponding *split frequency vector* $\ddot{\Lambda}(f) = \ddot{f} := (\ddot{f}_2, \ddot{f}_3, \dots, \ddot{f}_n)$ for a given f -sequence f by

$$\ddot{\Lambda}(f) = \ddot{f} := (\ddot{f}_2, \ddot{f}_3, \dots, \ddot{f}_n) : \mathcal{F}_n \rightarrow \ddot{\mathcal{F}}_n, \quad \ddot{f}_i := f_{i-1, -\min\langle (f_i - f_{i-1}) \boxtimes (1, 2, \dots, n) \rangle}. \quad (50)$$

For example, if there were four lineages that subtend three leaves each and one of these four lineages split at the beginning of the i -th epoch, then $\ddot{f}_i = 4$.

Proposition 3.31 (Probability of an f -sequence in terms of its shape statistics).

$$P(f) = \frac{2^{\mathfrak{T}(f)}}{(n-1)!} \prod_{i=2}^n \ddot{f}_i . \quad (51)$$

Proof. For any $f \in \mathcal{F}_n$, we can simplify $P(f)$ given by (45) and (44), as follows:

$$\begin{aligned} P(f) &= \prod_{i=2}^n P(f_i | f_{i-1}) = \prod_{i=2}^n \left(2^{\mathbf{1}_{\{1\}} \max\langle f_i - f_{i-1} \rangle} \ddot{f}_i (n-i+1)^{-1} \right) \\ &= \frac{2^{\sum_{i=2}^n \mathbf{1}_{\{1\}} \max\langle f_i - f_{i-1} \rangle}}{(n-1)!} \prod_{i=2}^n \ddot{f}_i = \frac{2^{\mathfrak{T}(f)}}{(n-1)!} \prod_{i=2}^n \ddot{f}_i . \end{aligned}$$

We get (51) from the definition of $\mathfrak{T}(f)$ at (48) as the number of distinctly-sized lineage splits in f . \square

4 Applications of lumped n -coalescents

Next we introduce the formalities to frame a partially ordered graph of lumped n -coalescents. We identify any n -coalescent $\{A^\uparrow(t)\}_{t \in \mathbb{R}_+}$ with its constitutive ordered triple $\mathfrak{C}_a := (\mathbb{A}_n, \{A^\uparrow(k)\}_{k \in [n]_-}, \mathcal{A}_n)$. The three components are $\mathfrak{C}_a(1) := \mathbb{A}_n$, its state space, $\mathfrak{C}_a(2) := \{A^\uparrow(k)\}_{k \in [n]_-}$, its embedded jump Markov chain, and $\mathfrak{C}_a(3) := \mathcal{A}_n$, the set of its sequential realizations. We index the n -coalescent triple \mathfrak{C}_a by a generic a -sequence $a \in \mathfrak{C}_a(3) := \mathcal{A}_n$. Let \mathfrak{C}_α and \mathfrak{C}_β be two n -coalescent triples with a Markov lumping $\mathcal{M}_{\alpha,\beta} : \mathfrak{C}_\alpha(1) \rightarrow \mathfrak{C}_\beta(1)$. We can apply this lumping to each component of any α -sequence $\alpha = (\alpha_n, \alpha_{n-1}, \dots, \alpha_1) \in \mathfrak{C}_\alpha(3)$ to obtain the lumped β -sequence according to the sequential lumping:

$$\begin{aligned} \underline{\mathcal{M}}_{\alpha,\beta}(\alpha) &= \beta : \mathfrak{C}_\alpha(3) \rightarrow \mathfrak{C}_\beta(3), \\ \underline{\mathcal{M}}_{\alpha,\beta}(\alpha) &= (\mathcal{M}_{\alpha,\beta}(\alpha_n), \dots, \mathcal{M}_{\alpha,\beta}(\alpha_1)) = (\beta_n, \dots, \beta_1) = \beta \in \mathfrak{C}_\beta(3) . \end{aligned}$$

Definition 4.1 (The lumped n -coalescents graph). Consider an \mathfrak{Y} -indexed set of n -coalescent triples $\{\mathfrak{C}_\alpha, \alpha \in \mathfrak{Y}\}$. Let, $\mathcal{M}_{\alpha,\beta} : \mathfrak{C}_\alpha(1) \rightarrow \mathfrak{C}_\beta(1)$, for some $\alpha, \beta \in \mathfrak{Y}$ be a Markov lumping. Let \mathfrak{E} be a set of such maps as well as the identity map. Then, the directed graph $\mathfrak{G}_{\mathfrak{Y},\mathfrak{E}}$ with vertices in $\{\mathfrak{C}_\alpha, \alpha \in \mathfrak{Y}\}$ and directed edges from a vertex \mathfrak{C}_α to a vertex \mathfrak{C}_β , provided there exists an $\mathcal{M}_{\alpha,\beta} \in \mathfrak{E}$, is the lumped n -coalescents graph. The immediate succedence relation: $\mathfrak{C}_\alpha \succ_{\mathfrak{E}} \mathfrak{C}_\beta \iff \exists \mathcal{M}_{\alpha,\beta} \in \mathfrak{E}$, induces the partial ordering $\succ_{\mathfrak{E}}$ on $\{\mathfrak{C}_\alpha, \alpha \in \mathfrak{Y}\}$, the vertices of $\mathfrak{G}_{\mathfrak{Y},\mathfrak{E}}$.

We introduced six different resolutions of the n -coalescent and the Markov lumpings between their state spaces (Figure 1). Suppose \mathbb{A}_n is the state space of another n -coalescent with a Markov lumping $\mathcal{B}(a_i) = b_i : \mathbb{A}_n \rightarrow \mathbb{B}_n$. Although there are several ways to augment \mathbb{B}_n to \mathbb{A}_n , depending on the statistical problem and data at hand, we abstract \mathbb{A}_n here to emphasize that \mathbb{B}_n is not the finest possible n -coalescent resolution. Our lumped n -coalescents graph is $\mathfrak{G}_{\mathfrak{V}, \mathfrak{E}}$ with

$$\begin{aligned} \mathfrak{V} &= \{a, b, c, d, f, g, h\} \text{ and} \\ \mathfrak{E} &= \{\mathcal{M}_{a,b}, \mathcal{M}_{b,c}, \mathcal{M}_{b,d}, \mathcal{M}_{c,f}, \mathcal{M}_{d,f}, \mathcal{M}_{d,g}, \mathcal{M}_{f,h}, \mathcal{M}_{g,h}\}, \text{ where} \\ \mathcal{M}_{a,b} &:= \mathcal{B} : \mathbb{A}_n \rightarrow \mathbb{B}_n, & \mathcal{M}_{b,c} &:= \mathcal{C} : \mathbb{B}_n \rightarrow \mathbb{C}_n, & \mathcal{M}_{b,d} &:= \mathcal{D} : \mathbb{B}_n \rightarrow \mathbb{D}_n, \\ \mathcal{M}_{c,f} &:= \mathcal{F} : \mathbb{C}_n \rightarrow \mathbb{F}_n, & \mathcal{M}_{d,f} &:= \mathcal{F}' : \mathbb{D}_n \rightarrow \mathbb{F}_n, & \mathcal{M}_{d,g} &:= \mathcal{G} : \mathbb{D}_n \rightarrow \mathbb{G}_n, \\ & & \mathcal{M}_{f,h} &:= \mathcal{H} : \mathbb{F}_n \rightarrow \mathbb{H}_n, & \mathcal{M}_{g,h} &:= \mathcal{H}' : \mathbb{G}_n \rightarrow \mathbb{H}_n . \end{aligned}$$

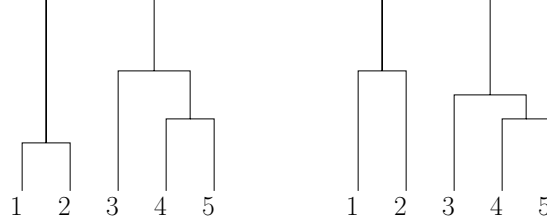
The lumped n -coalescents graph is the companion structure of the *n -coalescent experiments graph* defined in [23]. The lumped n -coalescents graph formalizes equivalence classes in the hidden space of genealogical sequences that one has to integrate over in order to compute the likelihood of the observed statistics at each node of the n -coalescent experiments graph. We can achieve maximal computational efficiency during likelihood evaluation if we conduct our integrations over the coarsest possible n -coalescent resolution in $\mathfrak{G}_{\mathfrak{V}, \mathfrak{E}}$ that will yield the exact likelihood of the desired statistics. We can measure this efficiency by the extent of various Markov lumpings and the size of the state spaces at different resolutions of $\mathfrak{G}_{\mathfrak{V}, \mathfrak{E}}$.

4.1 Nature and extent of Markov Lumpings

Here we study the nature and extent of the Markov lumpings between our six concrete state spaces in the lumped n -coalescents graph $\mathfrak{G}_{\mathfrak{V}', \mathfrak{E}'}$ with the sequence-specific index set $\mathfrak{V}' = \{b, c, d, f, g, h\}$ and $\mathfrak{E}' = \{\mathcal{C}, \mathcal{D}, \mathcal{F}, \mathcal{F}', \mathcal{G}, \mathcal{H}, \mathcal{H}'\}$ (Figure 1). We have seen that there is a bijection from \mathcal{B}_n , the set of b -sequences, as well as from \mathcal{C}_n , the set of c -sequences, to the set of ranked, labeled trees. We introduced b -sequences since there are Markov lumpings from b -sequences to all other resolutions. Since the state space of c -sequences is much smaller — there are no vintage tags — we will only consider c -sequences when the object of interest in inference is a ranked, labeled tree. The next two propositions state the impossibility of Markov lumpings between some state spaces in our lumped n -coalescents graph.

Proposition 4.2. *There is no Markov lumping from the state space of c -sequences to that of g -sequences and vice versa.*

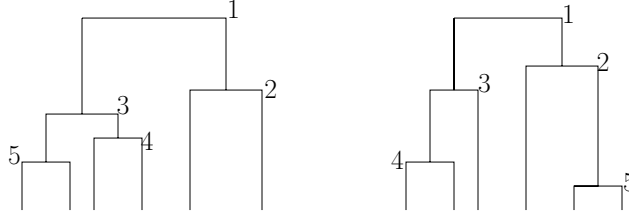
Proof. Since $|\mathbb{G}_n| < |\mathbb{C}_n|$, due to (19) and (32), there is clearly no lumping from \mathbb{G}_n to \mathbb{C}_n . For the other direction, consider the trees below. In the left tree, for $c_l \in \mathbb{C}_5^2$ we have $c_l = \{\{1, 2\}, \{3, 4, 5\}\}$. Also in the right tree, for $c_r \in \mathbb{C}_5^2$ we have $c_r = \{\{1, 2\}, \{3, 4, 5\}\}$, i.e. $c_l = c_r$. In the left tree, for $g_l \in \mathbb{G}_5^2$ we have $g_l = (0, 1, 0, 1)$. However, in the right tree, for $g_r \in \mathbb{G}_5^2$ we have $g_r = (0, 1, 1, 0)$.



So for two different states in \mathbb{G}_5^2 we have the same state in \mathbb{C}_5^2 . Thus there is no lumping from \mathbb{C}_n to \mathbb{G}_n . \square

Proposition 4.3. *There is no Markov lumping from the state space of g -sequences to that of f -sequences and vice versa.*

Proof. Since $|\mathbb{F}_n| < |\mathbb{G}_n|$, due to (32) and (39), there is clearly no lumping from \mathbb{F}_n to \mathbb{G}_n . For the other direction, consider the trees below. In the left tree, for $g_l \in \mathbb{G}_6^2$ we have $g_l = (0, 1, 1, 0, 0)$. Also in the right tree, for $g_r \in \mathbb{G}_6^2$ we have $g_r = (0, 1, 1, 0, 0)$, i.e. $g_l = g_r$. In the left tree, for $f_l \in \mathbb{F}_6^2$ we have $f_l = (0, 1, 0, 1, 0, 0)$. However, in the right tree, for $f_r \in \mathbb{F}_6^2$ we have $f_r = (0, 0, 2, 0, 0)$.



So for two different states in \mathbb{F}_6^2 we have the same state in \mathbb{G}_6^2 . Thus there is no lumping from \mathbb{G}_n to \mathbb{F}_n . \square

Let us now gain some insight on the extent of lumpings between \mathbb{C}_n , \mathbb{G}_n and \mathbb{F}_n . Note that the cardinality of \mathbb{C}_n , $|\mathbb{C}_n|$, is the n -th Bell number in (19). Further, the cardinality of \mathbb{G}_n , $|\mathbb{G}_n|$, is the $(n+1)$ -th Fibonacci number in (32). The cardinality of \mathbb{F}_n , $|\mathbb{F}_n|$, is the number of integer partitions of n in (39). The approximate values of $|\mathbb{C}_n|$, $|\mathbb{G}_n|$ and $|\mathbb{F}_n|$ are shown in Table 1 for typical samples sizes of interest to us. In fact, $|\mathbb{F}_n|/|\mathbb{G}_n| \rightarrow 0$ and $|\mathbb{G}_n|/|\mathbb{C}_n| \rightarrow 0$ as $n \rightarrow \infty$. This can be advantageous during integrations, involving dynamic programming, over paths of the Markov chain on \mathbb{G}_n

or \mathbb{F}_n instead of \mathbb{C}_n or over paths on \mathbb{F}_n instead of \mathbb{G}_n , provided the coarser resolution preserves the likelihood of the statistic of interest, i.e. the sampling distribution of the statistic of interest only depends on c , the hidden c -sequence, up to equivalence classes specified by $\underline{\mathcal{F}}(c) = f$ or $\underline{\mathcal{G}}(c) = g$, the corresponding f - or g -sequences, via their inverse sequential images in \mathbb{C}_n given by $\underline{\mathcal{F}}^{-1}(f)$ or $\underline{\mathcal{G}}^{-1}(g)$, respectively.

Table 1: Cardinalities of the state spaces \mathbb{C}_n , \mathbb{G}_n and \mathbb{F}_n .

$n = \mathbb{H}_n $	4	10	30	60	90	120
$ \mathbb{C}_n $	15	1.2×10^5	8.5×10^{23}	9.8×10^{59}	1.4×10^{101}	5.1×10^{145}
$ \mathbb{G}_n $	5	88	1.3×10^6	2.5×10^{12}	4.7×10^{18}	8.7×10^{24}
$ \mathbb{F}_n $	5	42	5.6×10^3	9.7×10^5	5.7×10^7	1.8×10^9

In the following, we will investigate how much information is lost when lumping the c -sequences to g -sequences or f -sequences. The next two propositions precisely describe the number of c -sequences or b -sequences or ranked, labeled trees that are coarsened by any specific f - or g -sequence.

Proposition 4.4 (The ranked, labeled trees of an f -sequence). *Let $f \in \mathcal{F}_n$ be any given f -sequence and let $c \in \underline{\mathcal{F}}^{-1}(f)$ be a corresponding c -sequence. Then the number of c -sequences (which is the number of ranked, labeled trees) corresponding to the given f is*

$$|\underline{\mathcal{F}}^{-1}(f)| = 2^{1-n} n! (n-1)! P(f) = n! 2^{\lceil(f)+1-n} \prod_{i=2}^n \ddot{f}_i, \quad (52)$$

and the conditional probability of c given f is

$$P(c|f) = 2^{\lceil(f)+n-1} (n!)^{-1} \prod_{i=2}^n \ddot{f}_i^{-1}. \quad (53)$$

Proof. The uniform probability on \mathbb{C}_n given by $2^{n-1} (n!(n-1)!)^{-1}$ invokes the probability on f -sequences in \mathcal{F}_n via the inverse image of $\underline{\mathcal{F}}^{-1}$, i.e.,

$$P(f) = P(\underline{\mathcal{F}}^{-1}(f)) = |\underline{\mathcal{F}}^{-1}(f)| 2^{n-1} (n!(n-1)!)^{-1}$$

and we have the first equality at (52). The second equality at (52) follows from substituting $P(f)$ at (51). The probability $P(c|f)$ at (53) follows from

$$P(c|f) = \frac{P(c, f)}{P(f)} = \frac{P(c)}{P(f)}.$$

□

Proposition 4.5 (The ranked, labeled trees of an g -sequence). *Let $g \in \mathcal{G}_n$ be any given g -sequence and let $b \in (\mathcal{D} \circ \mathcal{G})^{-1}(g) := \{\mathcal{D}^{-1}(d) : d \in \mathcal{G}^{-1}(g)\}$ be a corresponding b -sequence. Then the number of b -sequences (which is the number of ranked, labeled trees) corresponding to the given g is*

$$|(\mathcal{D} \circ \mathcal{G})^{-1}(g)| = |\mathcal{D}^{-1}(\mathcal{G}^{-1}(g))| = 2^{1-n} n! (n-1)! P(g) = n! 2^{-\mathfrak{J}(g)} , \quad (54)$$

where $\mathfrak{J}(g)$ is the number of cherries of the ranked tree shape induced by g . The conditional probability of b or c given g is

$$P(b|g) = P(c|g) = 2^{\mathfrak{J}(g)} / n! . \quad (55)$$

Proof. The bijection $\mathcal{G} : \mathcal{D}_n \rightarrow \mathcal{G}_n$, yields the first equality in (54) as follows:

$$(\mathcal{D} \circ \mathcal{G})^{-1}(g) := \{\mathcal{D}^{-1}(d) : d \in \mathcal{G}^{-1}(g)\} = \mathcal{D}^{-1}(\mathcal{G}^{-1}(g)) .$$

We derived $P(g)$, the probability of a g -sequence, at (36) in Proposition 3.24. Since each b -sequence $b \in \mathcal{B}_n = (\mathcal{D} \circ \mathcal{G})(\mathcal{G}_n)$, that is bijectively mapped to a ranked, labeled tree, has probability $2^{n-1}(n!(n-1)!)^{-1}$, we obtain,

$$|(\mathcal{D} \circ \mathcal{G})^{-1}(g)| = 2^{1-n} n! (n-1)! P(g) = 2^{1-n} n! (n-1)! \frac{2^{n-\mathfrak{J}(g)-1}}{(n-1)!} = n! 2^{-\mathfrak{J}(g)} .$$

Thus, (54) gives us the number of ranked, labeled trees that map to any given g -sequence g based on $\mathfrak{J}(g)$, the number of cherries of g . Due to the bijection from \mathcal{B}_n to \mathcal{C}_n and the uniform distribution on \mathcal{B}_n and \mathcal{C}_n , the probability $P(c|g) = P(b|g)$

$$P(c|g) = \frac{P(c, g)}{P(g)} = \frac{P(c)}{P(g)} = \frac{P(\mathcal{B}^{-1}(c))}{P(g)} = \frac{P(b)}{P(g)} = \frac{P(b, g)}{P(g)} = P(b|g) .$$

Now,

$$P(b|g) = P(c|g) = \frac{P(c)}{P(g)} = \frac{2^{n-1}(n!(n-1)!)^{-1}}{2^{n-\mathfrak{J}(g)-1}((n-1)!)^{-1}} = 2^{\mathfrak{J}(g)} / n! .$$

□

There is a bijection from \mathcal{D}_n , the set of d -sequences, as well as from \mathcal{G}_n , the set of g -sequences, to the set of ranked tree shapes. Again, since the state space of g -sequences is much smaller — as we do not track the size of components — we will only consider g -sequences when the object of interest in inference is a ranked tree shape. We introduced d -sequences since there are lumpings from d -sequences to f -sequences. For various shape statistics of ranked tree shapes, whose likelihood only

depends on the hidden f -sequence (described in § 4.3), it is preferable to study the lumped Markov chain on \mathbb{F}_n as opposed to that on \mathbb{G}_n . The next proposition gives the number of g -sequences or d -sequences or ranked tree shapes that are coarsened by any specific f -sequence.

Proposition 4.6 (The ranked tree shapes of an f -sequence). *Let $f \in \mathcal{F}_n$ be any given f -sequence and let $d \in \underline{\mathcal{F}}'^{-1}(f)$ and $g \in \underline{\mathcal{G}}(\underline{\mathcal{F}}'^{-1}(f)) := \{\underline{\mathcal{G}}(d) : d \in \underline{\mathcal{F}}'^{-1}(f)\}$ be a corresponding d - and g -sequence, respectively. The number of ranked tree shapes corresponding to the given f is*

$$|\underline{\mathcal{F}}'^{-1}(f)| = |\underline{\mathcal{G}}(\underline{\mathcal{F}}'^{-1}(f))| = 2^{-\hat{\mathfrak{A}}(f)} \prod_{i=2}^n \ddot{f}_i, \quad (56)$$

and the conditional probability of g given f is

$$P(g|f) = 2^{\hat{\mathfrak{A}}(f)} \left(\prod_{i=2}^n \ddot{f}_i \right)^{-1}. \quad (57)$$

Proof. The first equality in (56) is due to the bijection between \mathcal{D}_n and \mathcal{G}_n . For the second equality in (56), we establish $|\underline{\mathcal{F}}'^{-1}(f)| = 2^{-\hat{\mathfrak{A}}(f)} \prod_{i=2}^n \ddot{f}_i$ next. Recall that out of the $n-1$ splits in an f , $\mathfrak{J}(f)$ many of them are cherries and directly lead to leaves while $\mathfrak{T}(f)$ many of them lead to distinctly-sized splits. Let the number of remaining splits in f be defined as $\hat{\mathfrak{A}}(f) := n-1 - \mathfrak{T}(f) - \mathfrak{J}(f)$. Thus, $\hat{\mathfrak{A}}(f)$ is the number of balanced or equal-sized splits that are not cherries.

Let us highlight the following two facts: (1) for any $b, b' \in \underline{\mathcal{D}}^{-1}(\underline{\mathcal{F}}'^{-1}(f)) = \underline{\mathcal{C}}^{-1}(\underline{\mathcal{F}}^{-1}(f)) \subseteq \mathcal{B}_n$, $P(b) = P(b') = 2^{n-1}(n!(n-1)!)^{-1}$, and (2) for any $d, d' \in \underline{\mathcal{F}}'^{-1}(f)$ and any $g, g' \in \underline{\mathcal{G}}(\underline{\mathcal{F}}'^{-1}(f))$, $P(d) = P(d') = P(g) = P(g') = 2^{n-1-\mathfrak{J}(f)}/(n-1)!$, since $\mathfrak{J}(f) = \sum_{i=2}^n f_{i,2} = \mathfrak{J}(d) = \mathfrak{J}(d') = \mathfrak{J}(g) = \mathfrak{J}(g')$. Therefore, the number of ranked tree shapes mapped by a given f -sequence is the number of ranked labeled trees of an f -sequence divided by the number of ranked labeled trees of a g - or d -sequence with the same number of cherries as the f -sequence:

$$\begin{aligned} |\underline{\mathcal{F}}'^{-1}(f)| &= |\underline{\mathcal{G}}(\underline{\mathcal{F}}'^{-1}(f))| = \frac{|\underline{\mathcal{C}}^{-1}(\underline{\mathcal{F}}^{-1}(f))|}{|(\underline{\mathcal{D}} \circ \underline{\mathcal{G}}^{-1})(g)|} = \frac{|\underline{\mathcal{F}}^{-1}(f)|}{n! 2^{-\mathfrak{J}(g)}} = \frac{|\underline{\mathcal{F}}^{-1}(f)|}{n! 2^{-\mathfrak{J}(f)}} \\ &= 2^{\mathfrak{T}(f) + \mathfrak{J}(f) + 1 - n} \prod_{i=2}^n \ddot{f}_i, \end{aligned}$$

where we use (54) for the third-last equality and (52) for the last equality. Finally, (56) follows from the definition of $\hat{\mathfrak{A}}(f) := n-1 - \mathfrak{T}(f) - \mathfrak{J}(f)$. We get (57) from

$P(g)$ at (36), $P(f)$ at (51) and the definition of $\hat{\mathfrak{A}}(f)$ as follows:

$$\begin{aligned} P(g|f) &= \frac{P(g, f)}{P(f)} = \frac{P(g)}{P(f)} = \frac{2^{n-\mathfrak{I}(g)-1}((n-1)!)^{-1}}{2^{\mathfrak{I}(f)}((n-1)!)^{-1} \prod_{i=2}^n \ddot{f}_i} = \frac{2^{n-1-\mathfrak{I}(f)-\mathfrak{I}(f)}}{\prod_{i=2}^n \ddot{f}_i} \\ &= 2^{\hat{\mathfrak{A}}(f)} \left(\prod_{i=2}^n \ddot{f}_i \right)^{-1}. \end{aligned}$$

□

4.2 Examples

Next we provide some concrete examples of α -sequences for small n where $\alpha \in \mathfrak{V}' = \{b, c, d, f, g, h\}$ and calculate $P(f)$, $|\underline{\mathcal{F}}^{-1}(f)|$, $P(g)$, $|(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g)|$ and $|\underline{\mathcal{F}}'^{-1}(f)|$ based on (51), (52) (36), (54) and (56), respectively.

Example 4.7 (2 Samples). When there are 2 samples, we have exactly one b -, c -, d -, g - and f -sequence. We provide the d -, g - and f -sequences in Table 2. The only c -sequence in \mathcal{C}_2 is $(\{\{1\}, \{2\}\}, \{\{1, 2\}\})$ and the only b -sequence in \mathcal{B}_2 is $(\{\{1\}^{(2)}, \{2\}^{(2)}\}, \{\{1, 2\}^{(1)}\})$.

In Example 4.7 with $n = 2$ (see first row of Table 2), there is only one f -sequence whose $\mathfrak{I}(f) = 0$ and $\ddot{\Lambda}(f) = \ddot{f} = (1)$ and $\prod_2^2 \ddot{f}_i = 1$. Thus, $P(f) = (2^0/(2-1)!) 1 = 1$. We confirm the solitary c -sequence in \mathcal{C}_2 since $|\underline{\mathcal{F}}^{-1}(f)| = 2! 2^{0+1-2} 1 = 1$. Also, there is only one f - and g -sequence with $\mathfrak{I}(f) = \mathfrak{I}(g) = 1$, and thus $P(g) = 2^{2-1-1}/(2-1)! = 1$, $|(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g)| = 2! 2^{-1} = 1$. Since there are no equal sized splits that are not cherries, $\hat{\mathfrak{A}}(f) := n - 1 - \mathfrak{I}(f) - \mathfrak{I}(f) = 2 - 1 - 0 - 1 = 0$, and thus $|\underline{\mathcal{F}}'^{-1}(f)| = 2^{-0} 1 = 1$.

Example 4.8 (3 Samples). When there are 3 samples, we have 3 b -sequences, 3 c -sequences, 1 d -sequence, 1 g -sequence and 1 f -sequence. In Table 3, we tabulate the state-space, (backward) transition diagram, sequences and the corresponding probabilities at each of the six n -coalescent resolutions in \mathfrak{V}' .

There is only one f -sequence whose $\mathfrak{I}(f) = 1$, $\ddot{\Lambda}(f) = \ddot{f} = (1, 1)$ and $\prod_{i=2}^3 \ddot{f}_i = 1$. Thus, $P(f) = (2^1/(3-1)!) 1 = 1$ and $|\underline{\mathcal{F}}^{-1}(f)| = 3! 2^{1+1-3} 1 = 3$. Again, there is only one f - and g -sequence with one cherry, i.e. $\mathfrak{I}(f) = \mathfrak{I}(g) = 1$, and $\hat{\mathfrak{A}}(f) := n - 1 - \mathfrak{I}(f) - \mathfrak{I}(f) = 3 - 1 - 1 - 1 = 0$. Thus, $P(g) = 2^{3-1-1}/(3-1)! = 1$, $|(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g)| = 3! 2^{-1} = 3$ and $|\underline{\mathcal{F}}'^{-1}(f)| = 2^{-0} 1 = 1$.

n	ranked tree shape	d -sequence	g -sequence	f -sequence	\mathfrak{J}	\mathfrak{T}	$\widehat{\mathfrak{J}}$	\check{f}
2		$d = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$	$g = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$f = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$	1	0	0	(1)
3		$d = \begin{pmatrix} 3 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix}$	$g = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$	$f = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 3 & 0 & 0 \end{pmatrix}$	1	1	0	(1,1)
4		$d^\lambda = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$	$g^\lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$	$f^\lambda = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$	1	2	0	(1,1,1)
4		$d^\wedge = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$	$g^\wedge = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$	$f^\wedge = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$	2	0	1	(1,2,1)

Table 2: The d -, g - and f -sequences when n is 2, 3, and 4 are shown along with the corresponding ranked tree shape and the four shape statistics, namely, $\mathfrak{J} = \mathfrak{J}(f)$, $\mathfrak{T} = \mathfrak{T}(f)$, $\widehat{\mathfrak{J}} = \widehat{\mathfrak{J}}(f)$ and $\check{f} = \check{\Lambda}(f)$.

Example 4.9 (4 Samples). In the case of four samples, there are 18 b -sequences, 18 c -sequences, 2 d -sequence, 2 g -sequence and 2 f -sequence. We provide the d -, g - and f -sequences in Table 2. Out of the 18 c -sequences in \mathcal{C}_4 , it is possible to apply (40) and find that 12 c -sequences map to f^λ and 6 map to f^\wedge . Note that the ranked tree shapes corresponding to all the c -sequences $\underline{\mathcal{F}}^{-1}(f^\lambda)$ is the completely unbalanced g -sequence g^λ and that corresponding to all the c -sequences $\underline{\mathcal{F}}^{-1}(f^\wedge)$ is the completely balanced g -sequence g^\wedge . Finally, the shape statistic triple for the two f -sequences are:

$$(\mathfrak{J}(f^\lambda), \mathfrak{T}(f^\lambda), \widehat{\mathfrak{J}}(f^\lambda)) = (1, 2, 0) \quad \text{and} \quad (\mathfrak{J}(f^\wedge), \mathfrak{T}(f^\wedge), \widehat{\mathfrak{J}}(f^\wedge)) = (2, 0, 1) .$$

Let us examine the two f -sequences closely. For f^\wedge with $\mathfrak{T}(f^\wedge) = 0$, $\check{\Lambda}(f^\wedge) = \check{f}^\wedge = (1, 2, 1)$ and $\prod_{i=2}^4 \check{f}_i^\wedge = 2$ we obtain $P(f^\wedge) = (2^0/(4-1)!) 2 = 1/3$, $|\underline{\mathcal{F}}^{-1}(f^\wedge)| = 4! 2^{0+1-4} 2 = 6$ and $|\underline{\mathcal{F}}'^{-1}(f^\wedge)| = 2^{-1} 2 = 1$. Similarly, for f^λ with $\mathfrak{T}(f^\lambda) = 2$, $\check{\Lambda}(f^\lambda) = \check{f}^\lambda = (1, 1, 1)$ and $\prod_{i=2}^4 \check{f}_i^\lambda = 1$, we obtain $P(f^\lambda) = (2^2/(4-1)!) 1 = 2/3$, $|\underline{\mathcal{F}}^{-1}(f^\lambda)| = 4! 2^{2+1-4} 1 = 12$ and $|\underline{\mathcal{F}}'^{-1}(f^\lambda)| = 2^{-0} 1 = 1$.

Let us examine the two g -sequences closely. For g^\wedge with $\mathfrak{J}(g^\wedge) = 2$, $P(g^\wedge) = 2^{4-1-2}/(4-1)! = 1/3$ and $|(\underline{\mathcal{D}} \circ \underline{\mathcal{L}})^{-1}(g^\wedge)| = 4! 2^{-2} = 6$ and for g^λ with $\mathfrak{J}(g^\lambda) = 1$, $P(g^\lambda) = 2^{4-1-1}/(4-1)! = 2/3$ and $|(\underline{\mathcal{D}} \circ \underline{\mathcal{L}})^{-1}(g^\lambda)| = 4! 2^{-1} = 12$.

Example 4.10 (5 Samples). In the case of five samples, there are 180 b -sequences, 180 c -sequences, 5 d -sequence, 5 g -sequence and 4 f -sequence. As shown in Ta-

State Space & Transition Diagram	Sequences	P(sequence)
<p style="text-align: center;">E_3</p>	$b^{(1)} := (\{ \{1\}^{(3)}, \{2\}^{(3)}, \{3\}^{(3)} \}, \{ \{1,2\}^{(2)}, \{3\}^{(3)} \}, \{ \{1,2,3\}^{(1)} \})$ $b^{(2)} := (\{ \{1\}^{(3)}, \{2\}^{(3)}, \{3\}^{(3)} \}, \{ \{1,3\}^{(2)}, \{2\}^{(3)} \}, \{ \{1,2,3\}^{(1)} \})$ $b^{(3)} := (\{ \{1\}^{(3)}, \{2\}^{(3)}, \{3\}^{(3)} \}, \{ \{2,3\}^{(2)}, \{1\}^{(3)} \}, \{ \{1,2,3\}^{(1)} \})$	<p>1/3</p> <p>1/3</p> <p>1/3</p>
<p style="text-align: center;">C_3</p>	$\underline{c}^{(1)} = (\{ \{1\}, \{2\}, \{3\} \}, \{ \{1,2\}, \{3\} \}, \{ \{1,2,3\} \})$ $\underline{c}^{(2)} = (\{ \{1\}, \{2\}, \{3\} \}, \{ \{1,3\}, \{2\} \}, \{ \{1,2,3\} \})$ $\underline{c}^{(3)} = (\{ \{1\}, \{2\}, \{3\} \}, \{ \{2,3\}, \{1\} \}, \{ \{1,2,3\} \})$	<p>1/3</p> <p>1/3</p> <p>1/3</p>
<p style="text-align: center;">D_3</p>	$\underline{d}^{(1)} = \underline{d}^{(2)} = \underline{d}^{(3)} = d = ((0,0), (0,2), (3,0))$	<p>1</p>
<p style="text-align: center;">G_3</p>	$\underline{g}^{(d)} = g = ((0,0), (0,1), (1,0))$	<p>1</p>
<p style="text-align: center;">F_3</p>	$\underline{f}^{(c^{(1)})} = \underline{f}^{(c^{(2)})} = \underline{f}^{(c^{(3)})} = \underline{f}^{(d)} = f = ((3,0,0), (1,1,0), (0,0,1))$	<p>1</p>
<p style="text-align: center;">H_3</p>	$\underline{h}^{(f)} = \underline{h}^{(g)} = h = (3,2,1)$	<p>1</p>

Table 3: When $n = 3$ we tabulate the state spaces, (backward) transition diagrams, the sequences and their probabilities at six resolutions of the n -coalescent.

ble 4, we denote the 5 g -sequences as g^a, g^b, g^c, g^d, g^e and the five d -sequences as d^a, d^b, d^c, d^d, d^e along with their corresponding f -sequences as f^a, f^b, f^{ad}, f^e . Note that g^c and g^d as well as d^c and d^d map to the same f -sequence f^{ad} . Finally, the shape statistic triples for the four f -sequences are:

$$\begin{aligned} (\mathfrak{J}(f^a), \mathfrak{T}(f^a), \widehat{\mathfrak{J}}(f^a)) &= (1, 3, 0), & (\mathfrak{J}(f^b), \mathfrak{T}(f^b), \widehat{\mathfrak{J}}(f^b)) &= (2, 1, 1), \\ (\mathfrak{J}(f^{ad}), \mathfrak{T}(f^{ad}), \widehat{\mathfrak{J}}(f^{ad})) &= (2, 2, 0), & (\mathfrak{J}(f^e), \mathfrak{T}(f^e), \widehat{\mathfrak{J}}(f^e)) &= (2, 2, 0) . \end{aligned}$$

For the four f -sequences: f^a, f^b, f^{ad} and f^e , and the five g -sequences: g^a, g^b, g^c, g^d and g^e , we apply their shape statistics:

$$\begin{aligned} \mathfrak{T}(f^a) &= 3 & \mathfrak{T}(f^b) &= 1 & \mathfrak{T}(f^{ad}) &= \mathfrak{T}(f^e) &= 2 \\ \mathfrak{J}(g^a) &= 1 & \mathfrak{J}(g^b) &= \mathfrak{J}(g^c) &= \mathfrak{J}(g^d) &= \mathfrak{J}(g^e) &= 2 \\ \prod_{i=2}^5 \ddot{f}_i^a &= \prod_{i=2}^5 \ddot{f}_i^e &= 1^4 &= 1 & \prod_{i=2}^5 \ddot{f}_i^b &= \prod_{i=2}^5 \ddot{f}_i^{ad} &= 1 \ 1 \ 2 \ 1 = 2, \end{aligned}$$

to obtain the probabilities and cardinalities, based on (51), (52) (36), (54) and (56), as follows:

$$\begin{aligned} P(f^a) &= (2^3/(5-1)!) \ 1 = P(f^{ad}) = (2^2/(5-1)!) \ 2 = 1/3 \\ P(f^b) &= (2^1/(5-1)!) \ 2 = P(f^e) = (2^2/(5-1)!) \ 1 = 1/6 \\ |\underline{\mathcal{F}}^{-1}(f^a)| &= 5! \ 2^{3+1-5} \ 1 = |\underline{\mathcal{F}}^{-1}(f^{ad})| = 5! \ 2^{2+1-5} \ 2 = 60 \\ |\underline{\mathcal{F}}^{-1}(f^b)| &= 5! \ 2^{1+1-5} \ 2 = |\underline{\mathcal{F}}^{-1}(f^e)| = 5! \ 2^{2+1-5} \ 1 = 30 \\ P(g^a) &= 2^{5-1-1}/(5-1)! = 1/3 \\ P(g^b) &= P(g^c) = P(g^d) = P(g^e) = 2^{5-1-2}/(5-1)! = 1/6 \\ |(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g^a)| &= 5! \ 2^{-1} = 60 \\ |(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g^b)| &= |(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g^c)| = 5! \ 2^{-2} = 30 \\ |(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g^d)| &= |(\underline{\mathcal{D}} \circ \underline{\mathcal{G}})^{-1}(g^e)| = 5! \ 2^{-2} = 30 \\ |\underline{\mathcal{F}}'^{-1}(f^a)| &= |\underline{\mathcal{F}}'^{-1}(f^e)| = 2^{-0} \ 1 = 1 \\ |\underline{\mathcal{F}}'^{-1}(f^b)| &= 2^{-1} \ 2 = 1 \\ |\underline{\mathcal{F}}'^{-1}(f^{ad})| &= 2^{-0} \ 2 = 2 . \end{aligned}$$

Applications of (36) and (54) to the g -sequences of Examples 4.7, 4.8, 4.9 and 4.10 above are consistent with those of Tajima's topological relationships [28, Figures 1-3].

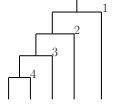
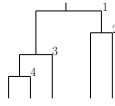
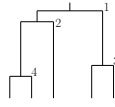
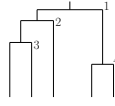
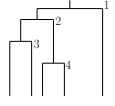
ranked tree shape	d -sequence	g -sequence	f -sequence	\mathfrak{J}	\mathfrak{T}	$\hat{\mathfrak{Q}}$	\check{f}
	$d^a = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$g^a = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$f^a = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$	1	3	0	(1, 1, 1, 1)
	$d^b = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 2 & 3 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$g^b = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$f^b = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$	2	1	1	(1, 1, 2, 1)
	$d^c = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 2 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$g^c = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$f^{cd} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$	2	2	0	(1, 1, 2, 1)
	$d^d = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 0 & 2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$g^d = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$f^{cd} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$	2	2	0	(1, 1, 2, 1)
	$d^e = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$g^e = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	$f^e = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{pmatrix}$	2	2	0	(1, 1, 1, 1)

Table 4: The d -, g - and f -sequences when $n = 5$ are shown along with the corresponding ranked tree shape and the four shape statistics, namely, $\mathfrak{J} = \mathfrak{J}(f)$, $\mathfrak{T} = \mathfrak{T}(f)$, $\hat{\mathfrak{Q}} = \hat{\mathfrak{Q}}(f)$ and $\check{f} = \check{f}(f)$. Note that the third and fourth row have the same f -sequence.

4.3 Shape Statistics where f -sequences are sufficient

Next we will show that any f -sequence f realized under the unvintaged and sized n -coalescent captures a considerable amount of information about the ranked tree shapes in the equivalence class of c -sequences $\mathcal{F}^{-1}(f)$ or in $\mathcal{G}(\mathcal{F}^{-1}(f))$. For instance, various tree shape statistics are further summaries of the f -sequence. We will make the former sentence precise by showing that several tree-shape statistics in the literature are functions of a sequence of $n - 1$ ordered pairs obtained from f -sequences. For a given c -sequence $c := (c_n, c_{n-1}, \dots, c_1)$, the corresponding shape statistic sequence or \tilde{s} -sequence is $\tilde{s} := (\tilde{s}_n, \tilde{s}_{n-1}, \dots, \tilde{s}_1)$, where $\tilde{s}_i := (\tilde{s}_{i,1}, \tilde{s}_{i,2})$. The i -th ordered pair $(\tilde{s}_{i,1}, \tilde{s}_{i,2})$ of the \tilde{s} -sequence is the size of the set $c_{i-1,j}$ that just coalesced and the size of the smaller of the two sets that just coalesced at the end of the i -th coalescent epoch. Here, we map the \tilde{s} -sequences directly from the set of f -sequences. The \tilde{s} -sequence or the *sequential Aldous shape statistic* [2] $\tilde{S}(f) = \tilde{s} : \mathcal{F}_n \rightarrow \tilde{\mathcal{S}}_n$ is obtained from an f -sequence f as follows:

$$\begin{aligned} \tilde{S}(f_n, f_{n-1}, \dots, f_1) &= \tilde{s} := (\tilde{s}_n, \tilde{s}_{n-1}, \dots, \tilde{s}_2), \\ \tilde{s}_i &:= (\tilde{s}_{i,1}, \tilde{s}_{i,2}) := (\max(\|f\|_i), \min(\|f\|_i) 2^{-\mathbf{1}_{\{0\}}(\max(\|f\|_i) - \min(\|f\|_i))}), \\ \|f\|_i &:= \{j | f_{i,j} - f_{i-1,j} \in \mathbb{N} : j \in \{1, 2, \dots, n\}\}. \end{aligned} \quad (58)$$

Therefore, f -sequences contain the information in \tilde{s} -sequences. Aldous [2] constructs the \tilde{s} -sequence forward in time using a tree-splitting model. This is partly motivated by a description of tree-shape imbalance via median-regression over a scatter-plot of the ordered pairs $(\tilde{s}_{i,1}, \tilde{s}_{i,2})$'s obtained from phylogenetic trees that were estimated from DNA sequences of extant taxa [2]. Next we show that several classical scalar-valued tree shape statistics are functions of $\tilde{s} = \tilde{S}(f)$. First consider the following family of scalar-valued tree shape statistics indexed by the non-empty elements of the power set of $\{2, 3, \dots, n\}$.

$$\mathfrak{Q}_n := \{Q_I(\tilde{s}) = q_I := \sum_{i=n}^2 \tilde{s}_{i,1} \mathbf{1}_I(\tilde{s}_{i,1}) : \tilde{\mathcal{S}}_n \rightarrow \mathcal{Q}_{I_n}, I \in \mathbf{2}^{\{2,3,\dots,n\}} \setminus \emptyset\}$$

Then, $Q_{\{2,3,\dots,n\}}(\tilde{s}) = q_{\{2,3,\dots,n\}} = \sum_{i=n}^2 \tilde{s}_{i,1}$ is the *Sackin's index* which is the sum of the number of leaves subtended by each internal node [28, 21]. $Q_{\{2\}}/2 = q_{\{2\}}/2$ is the *number of cherries*, i.e., the number of internal nodes that subtend exactly 2 leaves [18]. There are $2^{n-1} - 3$ other scalar-valued shape statistics in the family \mathfrak{Q}_n for the n -coalescent. Another scalar-valued statistic that needs more information than the number of leaves subtended by the set of internal nodes is the Colless' index [6].

It is the sum of the absolute difference between the number of leaves subtended by the two branches bifurcating from each internal node up to a constant factor. The Colless' index of an f -sequence f only depends on its Aldous shape statistic sequence $\tilde{S}(f) = \tilde{s}$ and is given by $(n^2 - 3n + 2)^{-1} \sum_{i=n}^2 (\tilde{s}_{i,1} - 2\tilde{s}_{i,d})$. Thus, we have shown that any f -sequence f captures a lot of information about the ranked tree shapes in $\mathcal{G}(\mathcal{F}^{-1}(f))$. However, some information is lost about the ranked tree shapes in the coarsening as one f -sequence may encode several distinct g -sequences — recall that 2 distinct g -sequences mapped to the same f -sequence in Example 4.10.

4.4 Shape Statistics where g -sequences are sufficient

In the last section, we showed that sampling distributions of f -sequences are sufficient to obtain that of several tree shape statistics. However, there are statistics based on ranked tree shapes for which the n -coalescent resolution of f -sequences is not sufficient. In [8], the *runs statistic* was proposed for detecting lineage-specific bursts within a population or between species.

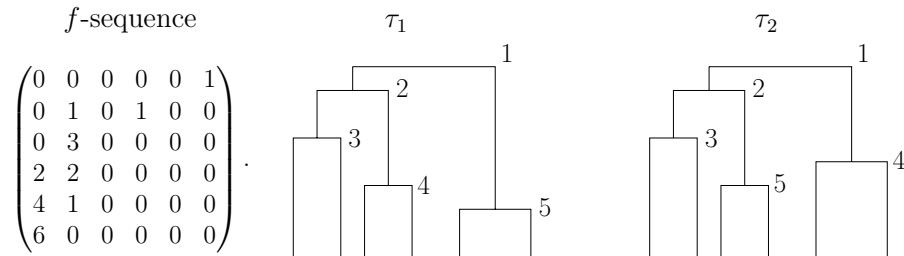


Figure 5: Two ranked tree shapes on six leaves. Note that τ_1 , the ranked tree shape in the middle panel, has run statistic 4 while τ_2 on the right has run statistic 5. However, both ranked tree shapes have the same f -sequence on the left.

The runs statistic is calculated recursively from a ranked tree shape τ . Note that the ranking on a tree shape is simply a total order of the interior vertices of the tree shape. By deleting the root of τ , we obtain two ranked tree shapes τ_1 and τ_2 . The ranked tree shape τ is induced by these two ranked tree shapes τ_1 and τ_2 together with a *shuffle* on the interior vertices of τ_1 and τ_2 . A shuffle puts the n_1 interior vertices in τ_1 and the n_2 interior vertices in τ_2 in order, e.g. 112122 means that first we have two bifurcations in τ_1 , then a bifurcation in τ_2 , followed by one bifurcation in τ_1 , then two bifurcations in τ_2 . The number of runs of a shuffle is the number of times we switch from i to j ($i \neq j$) plus one. Our shuffle 112122 has four runs. The

number of runs of a ranked tree shape τ is defined recursively,

$$R(\tau) = R(\tau_1) + R(\tau_2) + s(\tau) \text{ ,}$$

where $s(\tau)$ is the number of runs in the shuffle on the interior vertices of τ_1 and τ_2 . For details see [8].

As g -sequences can be mapped to ranked tree shapes via a bijection (Proposition 3.25), the g -sequences are sufficient for determining the runs statistic. Runs statistic cannot be obtained from f -sequences. For example, let us consider τ_1 and τ_2 , the two ranked tree shapes in Figure 5. There are 4 runs in τ_1 whereas τ_2 has 5 runs. However, both τ_1 and τ_2 have the same f -sequence.

5 Summary

We investigated the n -coalescent approximation of sample genealogical Markov chains of the simplest Wright-Fisher model. We showed that Kingman's n -coalescent approximation can be applied to any genealogical Markov chain that has the death chain as its lumped Markov chain. We described the combinatorial structures, forward and backward transition probabilities, sequence-specific and state-specific probabilities of the n -coalescent at six concrete genealogical resolutions. They include the genealogical resolutions of α -sequences, where $\alpha \in \mathfrak{V}' = \{b, c, d, g, f, h\}$.

Tajima's evolutionary relationships have been formalized into Tajima's n -coalescent or the vintaged and shaped n -coalescent. Its realizations are g -sequences that are in bijection with ranked tree shapes over a state space that is contained in $\{0, 1\}^{n-1}$. Kingman's unlabeled n -coalescent or the unvintaged and sized n -coalescent has been given a complete Markov description to produce f -sequences over the state space of integer partitions of n . The augmentation of the set of all set partitions of $\mathfrak{L} = \{1, 2, \dots, n\}$, the state space of Kingman's labeled n -coalescent or the unvintaged and labeled n -coalescent, by coalescent vintage tags, led to the state space of the Kingman-Tajima n -coalescent or the vintaged and labeled n -coalescent. Kingman's n -coalescent as well Tajima's n -coalescent are lumped Markov processes of the Kingman-Tajima n -coalescent. The b - and c -sequences that are realized sequentially under the Kingman-Tajima and the Kingman's labeled n -coalescents, respectively, are in bijection with ranked, labeled trees. The vintaged and sized n -coalescent over the state space of ordered integer partitions of n has d -sequences as its realizations. Both d - and g -sequences are in bijection with ranked tree shapes. Our second coarsest resolution of f -sequences preserves considerable information about the genealogies although it is not in bijection with any of the familiar definitions of phylogenetic

trees. The f -sequences are sufficient for site frequency spectrum and its linear combinations as shown in [23, 22] as well as for several tree shape statistics as shown here. Finally, the coarsest resolution is the pure death chain with only one h -sequence $(n, n - 1, \dots, 2, 1)$.

Using the theory of lumped Markov chains we formalized several Markov lumpings between $\mathfrak{V}' = \{b, c, d, g, f, h\}$, the six n -coalescent resolutions we pursued here. There is a partial order on \mathfrak{V}' induced by \mathfrak{E}' , the set of Markov lumpings between the six resolutions. We formalized this structure by $\mathfrak{G}_{\mathfrak{V}', \mathfrak{E}'}$, the lumped n -coalescents graph, and noted its implications for computational efficiency during likelihood evaluations in n -coalescent experiments. For likelihood evaluations during inference, we want the state space of the hidden genealogical Markov chains to be as small as possible. For instance, if the likelihood of our statistics requires integration at the resolution of ranked, labeled trees, we use c -sequences. If it requires integration over ranked tree shapes, we use g -sequences, and if it only requires integration over block sizes, we use f -sequences. The lumped n -coalescents graph allows us to consistently move between different n -coalescent resolutions as needed.

The lumped n -coalescents graph $\mathfrak{G}_{\mathfrak{V}', \mathfrak{E}'}$ is a formal and constructive embodiment of the *unified multi-resolution n -coalescent*. The lumped Markov chain projections of the underlying sample genealogical process through the Kingman-Tajima n -coalescent at the maximal vertex in $\mathfrak{G}_{\mathfrak{V}', \mathfrak{E}'}$ simultaneously at all other vertices of $\mathfrak{G}_{\mathfrak{V}', \mathfrak{E}'}$ gives us our unified multi-resolution n -coalescent. The basic properties of the n -coalescent, including (i) the robustness to variations in the underlying discrete population genetic models and (ii) the consistent embedding of the n -coalescent in the $(n + 1)$ -coalescent to obtain the coalescent, naturally apply to the unified multi-resolution coalescent. One can also obtain a unified multi-resolution coalescent of other more general coalescent processes.

Kemeny & Snell [12, p. 124] observe the following about a lumped process:

It is also often the case in applications that we are only interested in questions which relate to this coarser analysis of the possibilities. Thus it is important to be able to determine whether the new process can be treated by Markov chain methods.

It is exactly this observation about a lumped Markov process in the coalescent context that led to this paper and we have taken the necessary applied probabilistic steps towards realizing the potential for computationally efficient and statistically sufficient inference from population genetic statistics of today's massive genomic data.

6 Acknowledgements

We are most grateful to Robert C. Griffiths for his insights, comments and guidance on this project, to Amandine Véber for her comments on an earlier version of this manuscript, to Mike Steel for [12, def. 6.3.1] and his comments on a later version and to an anonymous referee for helpful comments and fixing thirty one typos. R.S. was supported by a research fellowship from the Royal Commission for the Exhibition of 1851 under the sponsorship of Peter Donnelly during the initial course of this study. T.S. was supported by a PhD scholarship of the German Science Foundation and a summer studentship of the Allan Wilson Centre.

References

- [1] W. G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7:256–276, 1975.
- [2] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.
- [3] M. Bahlo and R. Griffiths. Inference from gene trees in a subdivided population. *Theoret. Pop. Biol.*, 57:79–95, 1996.
- [4] M. Beaumont, C. Robert, J.-M. Marin, and J. Cornuet. Adaptivity for abc algorithms: the abc-pmc scheme. *Biometrika*, (to appear), 2009.
- [5] M. Beaumont, W. Zhang, and D. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.
- [6] D. H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31:100–104, 1982.
- [7] R. Fisher. *The Genetical Theory of Natural Selection*. Clarendon, Oxford, 1930.
- [8] D. Ford, T. Gernhard, and E. Matsen. A method for investigating relative timing information on phylogenetic trees. *Syst. Biol.*, 58(2):167–183, 2009.
- [9] R. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Stat. Sci.*, 9:307–319, 1994.
- [10] R. Griffiths and S. Tavaré. Markov chain inference methods in population genetics. *Math. Comput. Modelling*, 23:141–158, 1996.

- [11] M. Iorio and R. Griffiths. Importance sampling on coalescent histories. I. *Adv. Appl. Prob.*, 36:417–433, 2004.
- [12] J. Kemeny and J. Snell. *Finite Markov chains*. D. van Nostrand Company, Inc., Princeton, 1960.
- [13] D. G. Kendall. Some problems in mathematical genealogy. In J. Gani, editor, *Perspectives in Probability and Statistics*, pages 325–345. Academic Press, 1975.
- [14] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.
- [15] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.
- [16] C. Leuenberger and D. Wegmann. Bayesian computation and model selection without likelihoods. *Genetics*, 10.1534/genetics.109.109058:Published Articles Ahead of Print, 2009.
- [17] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100:15324–15328, 2003.
- [18] A. McKenzie and M. Steel. Distribution of cherries for two models of trees. *Math. Biosci.*, 164:81–92, 2000.
- [19] J. Pritchard, M. Seielstad, A. Perez-Lezaun, and M. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Mol. Biol. Evol.*, 16:1791–1798, 1999.
- [20] M. Rosenblatt. *Random Processes*. Springer-Verlag, 1974.
- [21] M. J. Sackin. “good” and “bad” phenograms. *Systematic Zoology*, 21:225–226, 1975.
- [22] R. Sainudiin, K. Thornton, J. Booth, M. Stillman, and R. Yoshida. Coalescent experiments II: Markov bases of classical population genetic statistics. *Bulletin of Mathematical Biology: Special Issue in Algebraic Biology*, pages Submitted. See UCDMS Research Report 2009/8, May 19, 2009. <http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/CoalExpsII.pdf>, 2009.

- [23] R. Sainudiin, K. Thornton, R. Griffiths, G. McVean, and P. Donnelly. Coalescent experiments I: Unlabeled n -coalescent and the site frequency spectrum. *Bulletin of Mathematical Biology: Special Issue in Algebraic Biology*, pages Submitted. See UCDMS Research Report 2009/7, May 19, 2009. <http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/CoalExpsI.pdf>, 2009.
- [24] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.
- [25] S. Sisson, Y. Fan, and M. Tanaka. Sequential monte carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104:1760–1765, 2007.
- [26] M. Slatkin. A vectorized method of importance sampling with applications to models of mutation and migration. *Theoret. Pop. Biol.*, 62:339–348, 2002.
- [27] M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. R. Statist. Soc. B*, 62:605–655, 2000.
- [28] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.
- [29] F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123:585–595, 1989.
- [30] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26:119–164, 1983.
- [31] G. Weiss and A. von Haeseler. Inference of population history using a likelihood approach. *Genetics*, 149:1539–1546, 1998.
- [32] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.