Electron. J. Probab. **23** (2018), no. 94, 1–27.
ISSN: 1083-6489 https://doi.org/10.1214/18-EJP219

Trees within trees: simple nested coalescents

Airam Blancas[*] Jean-Jil Duchamps[†‡] Amaury Lambert[†‡]
Arno Siri-Jégousse[§]

Abstract

We consider the compact space of pairs of nested partitions of $\mathbb{N}$, where by analogy with models used in molecular evolution, we call "gene partition" the finer partition and "species partition" the coarser one. We introduce the class of nondecreasing processes valued in nested partitions, assumed Markovian and with exchangeable semigroup. These processes are said simple when each partition only undergoes one coalescence event at a time (but possibly the same time). Simple nested exchangeable coalescent (SNEC) processes can be seen as the extension of Λ-coalescents to nested partitions. We characterize the law of SNEC processes as follows. In the absence of gene coalescences, species blocks undergo Λ-coalescent type events and in the absence of species coalescences, gene blocks lying in the same species block undergo i.i.d. Λ-coalescents. Simultaneous coalescence of the gene and species partitions are governed by an intensity measure ν_s on $(0,1] \times \mathcal{M}_1([0,1])$ providing the frequency of species merging and the law in which are drawn (independently) the frequencies of genes merging in each coalescing species block. As an application, we also study the conditions under which a SNEC process comes down from infinity.

Keywords: lambda-coalescent; exchangeable; partition; coming down from infinity; random tree; gene tree; population genetics; species tree; phylogenetics; evolution.
AMS MSC 2010: 60G09; 60G57; 60J35; 60J75; 92D10; 92D15.
Submitted to EJP on March 6, 2018, final version accepted on September 3, 2018.
Supersedes arXiv:1803.02133.
Supersedes HAL:hal-01725123.

[*]Institut für Mathematik, Goethe-Universität, Frankfurt, Germany.
E-mail: airam.blancas@gmail.com
[†]Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, CNRS UMR 8001, Paris, France.
[‡]Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS UMR 7241, INSERM U1050, PSL Research University, Paris, France.
E-mail: jean-jil.duchamps@normalesup.org
E-mail: amaury.lambert@upmc.fr
[§]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico.
E-mail: arno@sigma.iimas.unam.mx

1 Introduction

In the framework of population biology, one can see asexual organisms, but also DNA sequences or even species, as replicating particles. The genealogical ascendance of co-existing replicating particles can always be represented by a tree whose tips are labelled by the names of these particles [25, 26, 42]. Even if species are not strictly speaking replicating particles, ancestral relationships between species are also usually represented by a tree whose nodes are interpreted as *speciation* events, i.e., the emergence of two or more species from one single species. The inference of the so-called *gene tree* of contemporary DNA sequences from their comparison has a decade-long history. It is considered as a field in its own right, called *molecular phylogenetics* [19, 32], which relies heavily on the theory of Markov processes. (This can be misleading, but the *species tree*, much more often than the gene tree, is called a *phylogeny*.)

When one type of replicating particle is physically embedded in another type of particle, like a virus in its host, their common history can be depicted as a *tree within a tree* [14, 29, 34]: tree of dividing parasites inside the tree of dividing hosts, tree of paralogous genes (i.e. distinct DNA segments resulting from gene duplication and coding for similar functions) inside the gene family tree, gene tree inside the species tree. In many such cases, biologists are more interested in the coarser tree rather than in the finer tree. Typically, the finer tree is a gene tree and is inferred thanks to methods developed in molecular phylogenetics. One of the current methodological challenges in quantitative biology is to devise fast statistical algorithms able to also infer the coarser tree. When the genes are sampled from infecting pathogens of the same species (Influenza, HIV...), the coarser tree is the epidemic transmission process [21, 45]. When the genes are sampled from (any kind of) different species, the coarser tree is the *species tree* [33, 22, 43]. It is often required to use several gene trees nested in the same species tree to infer the latter.

In terms of stochastic modeling, the standard strategy is to define the two nested trees in a hierarchical model referred to as the *multispecies coalescent model* [36, 12] (see also [18, 3] for recent surveys on general coalescent theory and applications to population genetics). First, the species tree is fixed or drawn from some classic probability distribution (e.g., pure-birth process stopped at some fixed time, viewed as present time). Second, each gene sequence is assigned to the contemporary species it is (supposed to be) sampled from. Recall that each contemporary species is in correspondence with a tip of the species tree. Third, conditional on the species tree, each gene lineage can then be traced backwards in time inside the species tree, starting from the tip species harboring it and traveling through its ancestral species successively. In addition, gene lineages are assumed to coalesce according to the *censored Kingman coalescent* [24], i.e., each pair of lineages *lying in the same species* independently coalesces at constant rate.

In the case when the species tree is also distributed as a Kingman coalescent, the former two-type coalescent process is a Markov process as time runs backward, that we call the *nested Kingman coalescent* (or 'Kingman-in-Kingman') [28, 7, 11]. Our goal here is to display a much richer class of Markov models for trees within trees, called *simple nested exchangeable coalescent* (SNEC) processes, where multiple species lineages can merge into one single species lineage, and where simultaneously, within those merging species, multiple gene lineages can merge into one single gene lineage. To make this more precise, we show in the next display some valid and invalid coalescence events from an initial state where six genes, labeled from 1 to 6, are grouped by pairs in three species lineages. We represent this situation in the next display by a pair of partitions $\binom{\pi^s}{\pi^g}$, as in the left-hand side of the display. Event (A) is valid because the first two species merge and simultaneously, *within* these species, genes labeled 1, 2 and 3 coalesce. On

the contrary, event (B) is not a valid transition because there are two distinct gene coalescences (1 with 2, and 3 with 4), which is proscribed, and event (C) is not valid because the gene coalescence (5 with 6) is outside the species coalescence.

$$\begin{pmatrix} \{\,1,\quad 2\,\}\{\,3,\quad 4\,\}\{\,5,\quad 6\,\} \\ \{\,1\,\}\{\,2\,\}\{\,3\,\}\{\,4\,\}\{\,5\,\}\{\,6\,\} \end{pmatrix} \to \begin{pmatrix} \{\,1,\quad 2,\quad 3,\quad 4\,\}\{\,5,\quad 6\,\} \\ \{\,1,\quad 2,\quad 3\,\}\{\,4\,\}\{\,5\,\}\{\,6\,\} \end{pmatrix} \quad \text{(A)}$$

$$\not\to \begin{pmatrix} \{\,1,\quad 2,\quad 3,\quad 4\,\}\{\,5,\quad 6\,\} \\ \{\,1,\quad 2\,\}\{\,3,\quad 4\,\}\{\,5\,\}\{\,6\,\} \end{pmatrix} \quad \text{(B)}$$

$$\not\to \begin{pmatrix} \{\,1,\quad 2,\quad 3,\quad 4\,\}\{\,5,\quad 6\,\} \\ \{\,1\,\}\{\,2\,\}\{\,3\,\}\{\,4\,\}\{\,5,\quad 6\,\} \end{pmatrix} \quad \text{(C)}$$

In brief, SNEC processes are the generalization of Λ-*coalescents* to processes valued, not in partitions of $\mathbb{N}$, but in pairs of nested partitions of $\mathbb{N}$. The class of Λ-coalescents [37, 35], for which only one coalescence event can occur at a time, is a subclass of Markov, exchangeable processes with possibly non-binary nodes, called Ξ-*coalescents*, where several coalescence events can be simultaneous [4, 38].

Non-binary nodes in species trees can be interpreted as *unresolved nodes* (a sequence of binary nodes following each other too closely in time for their order to be inferred correctly) or *radiation* events (periods of frequent speciations due to the opening of new ecological opportunities that can be exploited by different, new species). In gene trees, non-binary nodes are increasingly recognized as a conspicuous sign of natural selection both by biologists [44, 30] and by mathematicians and physicists [2, 16, 10, 31, 13, 41] ; it is also well understood that non-binary nodes could be consequences of bottlenecks as well as large variance in offspring distributions [17, 40]. The class of SNEC processes includes all these features. They can distinguish unresolved nodes (sequence of stochastically close, binary coalescences) from radiations (multiple merger in the species tree). Under the interpretation of non-binary nodes as a result of natural selection, SNEC processes can model the appearance of alleles responsible for positive selection (multiple merger in the gene tree) or for divergent adaptation (multiple merger simultaneously in the gene tree and in the species tree).

From a mathematical point of view as well, SNEC processes open up the door to many possible new investigations. For example some of us are currently studying the speed of coming down from infinity of SNEC processes [28, 7] as well as similar extensions [15] to fragmentation processes [4]. It will be interesting to investigate how the nested trees generated by SNEC processes can be cast in the frameworks of multilevel measure-valued processes [8, 11] and flows of bridges [5, 6] as well as of exchangeable combs [20, 27]. It would also be natural to study the extension of Ξ-coalescents to nested partitions.

Organization of the article In Section 2, we introduce some notation, and give examples of nested coalescent processes whose distributions are characterized by four parameters. Section 3 formally defines our object of study, the SNEC processes. We prove our main result in Section 4, and show in Section 5 how SNEC processes can be constructed from a collection of Poisson point processes. Finally, Section 6 gives a necessary and sufficient condition under which SNEC processes come down from infinity.

2 Statement of results and notation

2.1 Statement of results and examples

An exchangeable partition is a random partition of $\mathbb{N}$ whose law is invariant by permutations of $\mathbb{N}$ (with finite support). A Λ-coalescent is a Markov process valued

in the exchangeable partitions of $\mathbb{N}$ typically starting from the partition $\mathbf{0}_\infty$ of $\mathbb{N}$ into singletons, and such that only one coalescence event can occur at a time. The generator of a Λ-coalescent $\mathcal{R} = (\mathcal{R}(t), t \geq 0)$ is characterized by a σ-finite measure ν on $(0,1]$ called the coagulation measure and a non-negative real number a called the Kingman coefficient. Then $\mathcal{R}$ can be constructed from a Poisson point process as follows.

For $x \in (0,1]$, let P_x denote the law of a sequence of i.i.d. Bernoulli(x) r.v.'s and define

$$P := \int_{(0,1]} \nu(dx) P_x$$

Also define $\mathtt{K}_{i,i'}$ the (Dirac) law of the sequence with only zero entries except a 1 at positions i and i' and set

$$\mathtt{K} := \sum_{1 \leq i < i'} \mathtt{K}_{i,i'}$$

Finally, let M be a Poisson point process with intensity measure $dt \otimes (P + a\mathtt{K})$. Roughly speaking, at each atom $(t, (X_i, i \geq 1))$ of M, $\mathcal{R}(t)$ is obtained from $\mathcal{R}(t-)$ by merging exactly the i-th block of $\mathcal{R}(t-)$ together, for all i such that $X_i = 1$. The rigorous description is given through restrictions of $\mathcal{R}$ to $[n] := \{1, \ldots, n\}$ and by applying Kolmogorov extension theorem. See [4] for details. Note that for this description to apply (i.e., for restrictions of $\mathcal{R}$ to $[n]$ to have positive holding times), one needs the coagulation measure to satisfy

$$\int_{(0,1]} x^2 \, \nu(dx) < \infty. \tag{2.1}$$

The finite measure $x^2 \, \nu(dx)$ is usually denoted $\Lambda(dx)$, hence the name Λ-coalescent.

We can now draw the parallel with the results obtained in this paper. We want to define a Markov process $\mathcal{R} = ((\mathcal{R}^s(t), \mathcal{R}^g(t)), \ t \geq 0)$ valued in exchangeable bivariate, nested partitions of $\mathbb{N}$, in the sense that the *gene partition* $\mathcal{R}^g(t)$ is finer than the *species partition* $\mathcal{R}^s(t)$ for all t a.s.

We now have to allow for coalescences in both the gene partition and the species partition. To this aim, we will consider a doubly indexed array of 0's and 1's $\mathbf{Z} = (\mathbf{X}, (\mathbf{Y}_i, \, i \geq 1)) = (X_i, Y_{ij}, \, i, j \geq 1)$. The goal is to give a characterization and a Poissonian construction of $\mathcal{R}$ under the assumptions that the semigroup of $\mathcal{R}$ is exchangeable and that both $\mathcal{R}^s$ and $\mathcal{R}^g$ undergo only one coalescence at a time (but possibly the same time), as detailed in forthcoming Definition 3.1. Roughly speaking, and similarly as previously, X_i will determine whether the i-th species block participates in the coalescence in the species partition $\mathcal{R}^s$, and Y_{ij} whether the j-th gene block of the i-th species block participates in the coalescence in the gene partition $\mathcal{R}^g$.

Let us start with the Kingman-type coalescences. Let $\mathtt{K}^s_{i,i'}$ be the (Dirac) law of the array $\mathbf{Z}$ with only zero entries except $X_i = X_{i'} = 1$ and let $\mathtt{K}^g_{i;j,j'}$ be the (Dirac) law of the array $\mathbf{Z}$ with only zero entries except $X_i = Y_{ij} = Y_{ij'} = 1$. Finally, define

$$\mathtt{K}^s = \sum_{1 \leq i < i'} \mathtt{K}^s_{i,i'} \quad \text{and} \quad \mathtt{K}^g = \sum_{1 \leq i} \sum_{1 \leq j < j'} \mathtt{K}^g_{i;j,j'}$$

Let us carry on with multiple gene mergers without simultaneous species coalescences. Let $x \in (0,1]$ and $i \in \mathbb{N}$. Let $P^g_{i,x}$ be the distribution of the array $\mathbf{Z}$ with only zero entries except at row i, where $X_i = 1$ and the $(Y_{ij}, \, j \geq 1)$ are i.i.d. Bernoulli(x) r.v.'s. Let us define

$$P^g_x := \sum_{i \geq 1} P^g_{i,x}$$

Finally, let us consider multiple species mergers, with possible simultaneous gene mergers. Let $x \in (0,1]$ and $\mu \in \mathcal{M}_1([0,1])$. Let $(X_i, \, i \geq 1)$ be a sequence of i.i.d.

Bernoulli(x) r.v.'s and let $(Q_i,\, i \geq 1)$ be an independent sequence of i.i.d. r.v.'s of $[0,1]$ with distribution μ. Then for each $i \geq 1$, conditional on X_i and Q_i, let $(Y_{ij},\, j \geq 1)$ be an independent sequence of i.i.d. Bernoulli(Q_i) r.v.'s. if $X_i = 1$ and the null array otherwise. Let us write $P^s_{x,\mu}$ for the distribution of the array $\mathbf{Z}$ thus defined.

Our main result is that for any simple nested exchangeable coalescent (SNEC) process $\mathcal{R}$, there are

- two non-negative real numbers a_s and a_g;
- a σ-finite measure ν_g on $(0,1]$;
- a σ-finite measure ν_s on $(0,1] \times \mathcal{M}_1([0,1])$,

such that $\mathcal{R}$ can be constructed from a Poisson point process M with intensity $dt \otimes \nu(d\mathbf{Z})$ where

$$\nu := a_s \mathtt{K}_s + a_g \mathtt{K}_g + \int_{(0,1]} \nu_g(dx)\, P^g_x + \int_{(0,1]\times\mathcal{M}_1([0,1])} \nu_s(dx, d\mu)\, P^s_{x,\mu}.$$

Similarly as explained previously, at each atom $(t, \mathbf{Z})$ of M, the double array $\mathbf{Z}$ prescribes which blocks have to merge at time t. For the finite restrictions of $\mathcal{R}$ to have positive holding times, the measures ν_s and ν_g are required to satisfy the forthcoming conditions (3.6) and (3.7) respectively, which are the analogs to (2.1).

Note that coagulations of the Kingman type cannot occur simultaneously in the species partition and in the gene partition.

We now give a couple of examples of SNEC processes.

If $\nu_s(dp, d\mu) = \nu'_s(dp)\, \delta_{\delta_0}(d\mu)$, species and genes never coalesce simultaneously and the nested coalescent is a multispecies coalescent (see Introduction), where the species tree is given by the Λ-coalescent with coagulation measure ν'_s and Kingman coefficient a_s, while the genes in the same species block undergo independent Λ-coalescents with coagulation measure ν_g and Kingman coefficient a_g. In particular, when ν'_s and ν_g are zero, the SNEC process is a nested Kingman coalescent (Kingman-in-Kingman).

Whenever ν_s is not under the form $\nu_s(dp, d\mu) = \nu'_s(dp)\, \delta_{\delta_0}(d\mu)$, species blocks and gene blocks can coalesce simultaneously. For example if $\nu_s(dp, d\mu) = \nu'_s(dp)\, \delta_{\delta_x}(d\mu)$ for $x \in (0,1]$, at each species coalescence event, a proportion x of gene blocks contained in the species blocks participating in the coalescence event, are simultaneously merged together. In particular, if $x = 1$, the gene tree coincides with the species tree on lineages situated after a species coalescence event. Recall that there are conditions (see (3.6)) for ν_s to be a correct SNEC measure, which in this case translate to

$$\int_{(0,1]} \nu'_s(dp)\, p^2 < \infty \quad \text{and} \quad \int_{(0,1]} \nu'_s(dp)\, p\, x^2 < \infty,$$

which is simply equivalent to

$$\int_{(0,1]} \nu'_s(dp)\, p < \infty.$$

Otherwise the simplest sort of measure ν_s can be obtained by parameterizing its second component μ, for example if μ is a Beta distribution $\mu_{a,b}(dq) = c_{a,b}\, q^{a-1}(1-q)^{b-1}\, dq$, where $a, b > 0$ and $c_{a,b} = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}$, we can consider ν_s under the form

$$\nu_s(dp, d\mu) = \nu'_s(dp, da, db)\, \delta_{\mu_{a,b}}(d\mu).$$

In this case, the condition (3.6a) reads

$$\int_{(0,1]\times(0,\infty)\times(0,\infty)} \nu'_s(dp, da, db)\, p^2 < \infty,$$

and (3.6b) becomes

$$\int_{(0,1]\times(0,\infty)\times(0,\infty)} \nu'_s(dp,da,db)\, p \int_{[0,1]} c_{a,b}\, q^{a+1}(1-q)^{b-1}\, dq < \infty,$$

which can be rewritten

$$\int_{(0,1]\times(0,\infty)\times(0,\infty)} \nu'_s(dp,da,db)\, \frac{pa(a+1)}{(a+b)(a+b+1)} < \infty.$$

Note that the idea to use a Beta distribution here is inspired by the Λ-coalescent setting [35], where Beta distributions appear as natural candidates for the parametrization of the measure Λ, as the coalescence rate of each k-tuple of blocks among a total number of b blocks is expressed in the form

$$\int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx).$$

2.2 Notation

For any $n \in \bar{\mathbb{N}} := \mathbb{N} \cup \{+\infty\}$, let $\mathcal{P}_n$ be the set of partitions of $[n]$. A partition π is called *simple* if at most one of its non-empty blocks is not a singleton. We denote the set of simple partitions of $[n]$ by $\mathcal{P}'_n$, that is,

$$\mathcal{P}'_n = \{\pi \in \mathcal{P}_n,\ \text{Card}\{i,\ |\pi_i| > 1\} \leq 1\}$$

where $\pi_1, \pi_2, \dots$ denote the blocks of π ordered by their least element and $|\pi_i|$ stands for the number of elements in the block π_i. Recall that a partition π can be viewed as an equivalence relation, in the sense that $i \overset{\pi}{\sim} j$ if and only if i and j belong to the same block of the partition π. If π^g and π^s belong to $\mathcal{P}_n$, we will say that the bivariate partition $\pi = (\pi^s, \pi^g)$ is *nested* (or equivalently that π^g is finer than π^s) when

$$i \overset{\pi^g}{\sim} j \implies i \overset{\pi^s}{\sim} j.$$

Note that this is defines a natural partial order on $\mathcal{P}_n$, and we can write $\pi^g \preceq \pi^s$ if (π^g, π^s) is nested. The set of nested partitions of $[n]$ is denoted in the sequel by $\mathcal{N}_n$. We will sometimes use the notation $\mathbf{1}_n := \{[n]\}$ for the coarsest partition of $[n]$, and $\mathbf{0}_n := \{\{1\}, \{2\}, \dots\}$ for the finest partition of $[n]$.

Example 2.1. An example of nested partition of $\{1, 2, \dots, 10\}$ is given by

$$\pi^s = \big\{\{1,5,7\},\ \{2,4,8,10\},\ \{3,6,9\}\big\}$$
$$\pi^g = \big\{\{1\},\ \{2,4\},\ \{3\},\ \{5,7\},\ \{6,9\},\ \{8\},\ \{10\}\big\}.$$

The notation (π^s, π^g) owes to our modeling inspiration (see Introduction) where gene lineages are enclosed into species lineages.

Notation related to and properties of $\mathcal{P}_n$ can naturally be extended to the framework of bivariate partitions. For the sake of completeness we specify here the ones we will use repeatedly. The number of non-empty blocks of a bivariate partition $\pi = (\pi_1, \pi_2) \in \mathcal{P}_{n_1} \times \mathcal{P}_{n_2}$ is merely $|\pi| := (|\pi_1|, |\pi_2|)$. If $m_1 < n_1$ and $m_2 < n_2$, we write $\pi_{|m_1 \times m_2}$ for the restriction of π to $\mathcal{P}_{m_1} \times \mathcal{P}_{m_2}$, that is, $\pi_{|m_1\times m_2} = (\pi_{1\,|m_1}, \pi_{2\,|m_2})$. If $m \leq \min(n_1, n_2)$, we will write $\pi_{|m} := \pi_{|m\times m}$ for its restriction to $\mathcal{P}^2_m := \mathcal{P}_m \times \mathcal{P}_m$. A sequence $\pi^{(1)}, \pi^{(2)}, \dots$ of elements of $\mathcal{P}^2_1, \mathcal{P}^2_2, \dots$ is called *consistent* if for all integers $k' \leq k$, $\pi^{(k')}$ coincides with the restriction of $\pi^{(k)}$ to $[k']^2$. Moreover, a sequence of partitions $(\pi^{(n)} : n \in \mathbb{N})$ is consistent if and only if there exists $\pi \in \mathcal{P}^2_\infty$ such that $\pi_{|n} = \pi^{(n)}$ for every $n \in \mathbb{N}$.

Given a nested partition we can use the coagulation operator Coag (more details in Chapter 3 in Bertoin [4]) to write the species partition in terms of the labels of the gene partition. Recall that if $\pi \in \mathcal{P}_n$ and $\tilde{\pi} \in \mathcal{P}_m$ with $m \geq |\pi|$, then define $\pi' = \text{Coag}(\pi, \tilde{\pi})$ as the partition of $\mathcal{P}_n$ such that

$$\pi'_j = \bigcup_{i \in \tilde{\pi}_j} \pi_i.$$

For every $n \in \bar{\mathbb{N}}$, let $\pi = (\pi^s, \pi^g)$ be an element of $\mathcal{N}_n$ and write $m = |\pi^g|$. The unique partition $\bar{\pi} \in \mathcal{P}_m$ such that $\pi^s = \text{Coag}(\pi^g, \bar{\pi})$ is called the *link* partition of π. We sometimes say that π is linked by $\bar{\pi}$. To illustrate the previous definition, observe that the nested partition defined in Example 2.1 has link partition $\bar{\pi} = \{\{1,4\},\{2,6,7\},\{3,5\}\}$.

We can next get a partition of $\mathcal{P}_{n_1} \times \mathcal{P}_{n_2}$ through the coagulation of two pairs of partitions. More precisely, if $(\pi^1, \tilde{\pi}^1) \in \mathcal{P}_{n_1} \times \mathcal{P}_{n'_1}$ and $(\pi^2, \tilde{\pi}^2) \in \mathcal{P}_{n_2} \times \mathcal{P}_{n'_2}$ with $n'_1 \geq |\pi^1|$ and $n'_2 \geq |\pi^2|$, then $(\text{Coag}(\pi^1, \tilde{\pi}^1), \text{Coag}(\pi^2, \tilde{\pi}^2))$ is well defined and it is an element of $\mathcal{P}_{n_1} \times \mathcal{P}_{n_2}$. If we denote $\pi = (\pi^1, \pi^2)$ and $\tilde{\pi} = (\tilde{\pi}^1, \tilde{\pi}^2)$ we will say that the pair $(\pi, \tilde{\pi})$ is *admissible* and denote the latter operation by $\text{Coag}_2(\pi, \tilde{\pi})$. In the following we will sometimes call the partition $\tilde{\pi}$ as the *recipe* partition.

In the sequel, we are interested in the coagulation of a nested partition, say $\pi = (\pi^s, \pi^g)$, with a pair of simple partitions $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$. Nevertheless, we should observe that the resulting partition, $\text{Coag}_2(\pi, \tilde{\pi})$ is not necessarily nested. For instance, if we coagulate the partition π of Example 2.1, with $\tilde{\pi}^s = \{\{1,2\},\{3\}\}$, and $\tilde{\pi}^g = \{\{1,3\},\{2\},\{4\},\{5\},\{6\},\{7\}\}$ then $\text{Coag}(\pi^g, \tilde{\pi}^g)$ is not nested in $\text{Coag}(\pi^s, \tilde{\pi}^s)$. In order to maintain the nested property while coagulating a nested partition we need to watch out the way the gene blocks do merge together and if they respect the species structure. To this end, for any $n \in \bar{\mathbb{N}}$ and $\pi \in \mathcal{N}_n$, we can define the set $\widetilde{\mathcal{P}}(\pi) \subset (\mathcal{P}'_n)^2$ of simple recipe partitions permitting a consistent merger of species and genes, i.e.

$$\widetilde{\mathcal{P}}(\pi) = \left\{ \tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g) \in (\mathcal{P}'_n)^2,\ i \overset{\tilde{\pi}^g}{\sim} j \implies i \overset{\bar{\pi}}{\sim} j, \text{ or } k \overset{\tilde{\pi}^s}{\sim} l, \text{ where } \pi^g_i \subset \pi^s_k \text{ and } \pi^g_l \subset \pi^s_l \right\},$$

where $\bar{\pi}$ denotes as usual the link partition of π. Simply put, $\widetilde{\mathcal{P}}(\pi)$ is the subset of $(\mathcal{P}'_n)^2$ such that

$$\tilde{\pi} \in \widetilde{\mathcal{P}}(\pi) \iff \text{Coag}_2(\pi, \tilde{\pi}) \in \mathcal{N}_n.$$

Finally the natural partial order on partitions can be extended to bivariate partitions by defining $(\pi^{1,s}, \pi^{1,g}) \preceq (\pi^{2,s}, \pi^{2,g}) \iff \pi^{1,s} \preceq \pi^{2,s}$ and $\pi^{1,g} \preceq \pi^{2,g}$. This partial order allows us to see coalescent processes as nondecreasing processes in the space of nested partitions.

3 Simple nested exchangeable coalescents

In the aim to describe the joint dynamics of the species and gene partitions, we will now define a nondecreasing process with values in the nested partitions, called *nested coalescent process*. In this work we are only interested in *simple* nested coalescents in the sense that at any jump event, called coalescence event, all blocks undergoing a modification merge into one single block. Simple exchangeable coalescent processes were first introduced independently by Pitman [35] and Sagitov [37], and are usually called in the literature Λ-coalescents (see Introduction). Here we use the term *simple* as in [4], to denote the analog of a Λ-coalescent process in the case of (nested) bivariate partitions.

Note that for any partition $\pi \in \mathcal{P}_\infty$ and any *injection* $\sigma : \mathbb{N} \to \mathbb{N}$, there is a partition $\sigma(\pi)$ defined by

$$i \overset{\sigma(\pi)}{\sim} j \iff \sigma(i) \overset{\pi}{\sim} \sigma(j).$$

For bivariate partitions we define in the same way $\sigma(\pi^s, \pi^g) := (\sigma(\pi^s), \sigma(\pi^g))$. For random partitions, exchangeability is usually defined as invariance under the action of permutations $\sigma : \mathbb{N} \to \mathbb{N}$. Here, to avoid degenerate processes we will define our processes as being invariant under the action of all injections $\sigma : \mathbb{N} \to \mathbb{N}$. Indeed, by making this assumption we avoid dependence on, for instance, the total number of blocks of the partition. An example of what we consider here a degenerate process with values in $\mathcal{P}_\infty$ would be a modified Kingman coalescent where any pair of blocks merge at rate $a = a(n)$, a function of n the total number of blocks. While this process would be invariant under permutations of $\mathbb{N}$, it is in general not invariant under injections, as their action can change the total number of blocks in a partition of $\mathbb{N}$. Furthermore, given $(\Pi(t),\, t \geq 0)$ such a process and n an integer, the restriction $(\Pi(t)_{|n},\, t \geq 0)$ would not be a Markov process, as the jump rates of $\Pi(t)_{|n}$ depend on the whole partition $\Pi(t)$. Invariance under injections ensures us that processes can be consistently defined, i.e. that $(\Pi(t)_{|n},\, t \geq 0)$ will always be a Markov process. It will also be useful in forthcoming proofs to consider invariance under injections rather than only permutations.

Since we consider processes with values in the space $\mathcal{P}_\infty$, let us endow it with the natural topology generated by the sets of the form $\{\pi' \in \mathcal{P}_\infty,\, \pi'_{|n} = \pi\}$ for $n \in \mathbb{N}$ and $\pi \in \mathcal{P}_n$. It is readily checked that this topology is metrizable and makes $\mathcal{P}_\infty$ compact. Also, note that the product topology on $\mathcal{P}_\infty^2$, and that induced on $\mathcal{N}_\infty$ also makes them compact.

Definition 3.1. *Let $\mathcal{R} := ((\mathcal{R}^s(t), \mathcal{R}^g(t)),\, t \geq 0)$ be a càdlàg Markov process with values in $\mathcal{P}_\infty^2$. This process is called a* simple nested exchangeable coalescent*, SNEC for short, if*

i) For any $t \geq 0$, $\mathcal{R}(t)$ is nested;

ii) The process $(\mathcal{R}(t),\ t \geq 0)$ evolves with simple coalescence events, that is for any time $t \geq 0$ such that $\mathcal{R}(t-) \neq \mathcal{R}(t)$, there is a random bivariate partition $\widetilde{\mathcal{R}}(t) = (\widetilde{\mathcal{R}}^s(t), \widetilde{\mathcal{R}}^g(t))$ taking values in $\widetilde{\mathcal{P}}(\mathcal{R}(t-))$ such that

$$\mathcal{R}(t) = \mathit{Coag}_2(\mathcal{R}(t-), \widetilde{\mathcal{R}}(t));$$

iii) The semigroup of the process $(\mathcal{R}(t),\ t \geq 0)$ is exchangeable, in the sense that for any $t, t' \geq 0$ and any injection *$\sigma : \mathbb{N} \to \mathbb{N}$,*

$$(\sigma(\mathcal{R}(t+t')) \mid \mathcal{R}(t) = \pi) \overset{(d)}{=} (\mathcal{R}(t+t') \mid \mathcal{R}(t) = \sigma(\pi))\,. \tag{3.1}$$

To start the analysis of SNEC processes we would like to make some observations related to Definition 3.1. First note that $\mathcal{R}$ is a $\mathcal{N}_\infty$-valued process such that for every $t, t' \geq 0$, the conditional distribution of $\mathcal{R}(t+t')$ given $\mathcal{R}(t) = \pi$ is the law of $\mathrm{Coag}_2(\pi, \tilde{\pi})$, where $\tilde{\pi} \in \widetilde{\mathcal{P}}(\pi)$, hence the law of $\tilde{\pi}$ depends on t' but also on π. Also, it will be clear from our main result (see Theorem 3.4) that $(\mathcal{R}^s(t),\ t \geq 0)$ is an exchangeable coalescent, however $(\mathcal{R}^g(t),\ t \geq 0)$ is not a Markov process in general, because the distribution of $\mathcal{R}^g(t+t')$ may depend on $\mathcal{R}^s(t)$.

We now turn to investigate the transitions of the restrictions of a SNEC to finite partitions, which relies on the following lemma.

Lemma 3.2 (Projective Markov property)**.** *Let $\mathcal{R} = (\mathcal{R}(t),\ t \geq 0)$ be a process with values in $\mathcal{N}_\infty$ and for every integer n, write $\mathcal{R}_{|n} = (\mathcal{R}_{|n}(t),\ t \geq 0)$ for its restriction to $\mathcal{N}_n$. Then $\mathcal{R}$ is a SNEC in $\mathcal{N}_\infty$ if and only if for all $n \in \mathbb{N}$, $\mathcal{R}_{|n}$ is a continuous-time Markov chain on the space $\mathcal{N}_n$ satisfying the analog of statements i) – iii) of Definition 3.1, namely:*

i) For all $t \geq 0$, $\mathcal{R}_{|n}$ is nested;

ii) For $\varrho, \pi \in \mathcal{N}_n$, the rate from ϱ to π is zero if π can not be obtained from a simple coalescence event;

iii) The Markov chain $(\mathcal{R}_{|n}(t),\ t \geq 0)$ is exchangeable, in the sense that for any $t, t' \geq 0$, $\varrho, \pi \in \mathcal{N}_n$ and σ permutation *of n, the rate from ϱ to π is equal to that from $\sigma(\varrho)$ to $\sigma(\pi)$.*

Proof. Let $\mathcal{R}$ be a SNEC in $\mathcal{N}_\infty$ and let $n \in \mathbb{N}$. Let us prove that $\mathcal{R}_{|n}$ satisfies the claimed properties. Let $\varrho \in \mathcal{N}_n$. Pick $\varrho^\star \in \mathcal{N}_\infty$ such that $\varrho^\star_{|n} = \varrho$, and which contains an infinite number of species blocks, each of which containing an infinite number of gene blocks, each of them being an infinite subset of $\mathbb{N}$. Now for any $\varrho' \in \mathcal{N}_\infty$ such that $\varrho'_{|n} = \varrho$, there is an injection $\sigma : \mathbb{N} \to \mathbb{N}$ such that $\sigma(\varrho^\star) = \varrho'$ and such that $\sigma_{|[n]} = \mathrm{id}_{[n]}$, so for any $t, t' \geq 0$,

$$\begin{aligned}(\mathcal{R}_{|n}(t+t') \mid \mathcal{R}(t) = \varrho') &\overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}(t) = \sigma(\varrho^\star)) \\ &\overset{(d)}{=} (\sigma(\mathcal{R})_{|n}(t+t') \mid \mathcal{R}(t) = \varrho^\star) \\ &\overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}(t) = \varrho^\star).\end{aligned}$$

Since this is valid for any ϱ' such that $\varrho'_{|n} = \varrho$, this conditional distribution depends only on $\{\mathcal{R}_{|n}(t) = \varrho\}$, which proves that $\mathcal{R}_{|n}$ is a Markov process. Now the assumption that $\mathcal{R}$ has càdlàg paths ensures us that the process $\mathcal{R}_{|n}$ stays some positive time in each visited state *a.s.* Therefore $\mathcal{R}_{|n}$ is a continuous-time Markov chain. Now statements *i) – iii)* are easily deduced from Definition 3.1.

Conversely, let $\mathcal{R} = (\mathcal{R}(t),\ t \geq 0)$ be a process with values in $\mathcal{N}_\infty$ such that for all $n \in \mathbb{N}$, $\mathcal{R}_{|n}$ is a Markov chain satisfying *i) – iii)* of the lemma. Then *i)* and *ii)* of Definition 3.1 follow immediately, and it remains to check that for any injection $\sigma : \mathbb{N} \to \mathbb{N}$, the equality in distribution (3.1) holds.

Let $\sigma : \mathbb{N} \to \mathbb{N}$ be an injection and fix $n \in \mathbb{N}$. Define $N = \max\{\sigma(1), \sigma(2), \ldots, \sigma(n)\}$, and consider $\widetilde{\sigma} : [N] \to [N]$ a permutation such that for all $1 \leq i \leq n$, $\widetilde{\sigma}(i) = \sigma(i)$. For instance, one can define inductively for $n+1 \leq i \leq N$,

$$\widetilde{\sigma}(i) := \min([N] \setminus \{\sigma(1), \sigma(2), \ldots, \sigma(i-1)\}).$$

Now notice that for any $t \geq 0$ and any $\pi \in \mathcal{N}_\infty$,

$$\sigma(\pi)_{|n} = \widetilde{\sigma}(\pi_{|N})_{|n},$$

which enables us to write, for any $t, t' \geq 0$,

$$\begin{aligned}(\sigma(\mathcal{R})_{|n}(t+t') \mid \mathcal{R}(t) = \pi) &\overset{(d)}{=} (\widetilde{\sigma}(\mathcal{R}_{|N}(t+t'))_{|n} \mid \mathcal{R}(t) = \pi) \\ &\overset{(d)}{=} (\mathcal{R}_{|N}(t+t')_{|n} \mid \mathcal{R}_{|N}(t) = \widetilde{\sigma}(\pi_{|N})) \\ &\overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}_{|n}(t) = \widetilde{\sigma}(\pi_{|N})_{|n}) \\ &\overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}_{|n}(t) = \sigma(\pi)_{|n}) \\ &\overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}(t) = \sigma(\pi)).\end{aligned}$$

The passage to the second line in the last display is a consequence of *iii)* of the lemma, and we used the fact that restrictions are Markov chains, i.e. $(\mathcal{R}_{|n}(t+t') \mid \mathcal{R}_{|n}(t) = \pi_{|n}) \overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}(t) = \pi)$. Since n is arbitrary in

$$(\sigma(\mathcal{R})_{|n}(t+t') \mid \mathcal{R}(t) = \pi) \overset{(d)}{=} (\mathcal{R}_{|n}(t+t') \mid \mathcal{R}(t) = \sigma(\pi)),$$

we have shown (3.1), concluding the proof. □

This key lemma enables us to give the following first properties of SNEC processes.

Proposition 3.3. *Let $\mathcal{R}$ be a SNEC.*

- *If the process $\mathcal{R}$ starts from an exchangeable nested partition $\mathcal{R}(0)$, then for any $t \geq 0$, $\mathcal{R}^g(t)$ and $\mathcal{R}^s(t)$ are exchangeable partitions.*
- *The process $\mathcal{R}$ is a Feller process, so in particular it satisfies the strong Markov property.*
- *Conditional on $\mathcal{R}(t)$, if $\bar{\mathcal{R}}(t)$ denotes the link partition of $\mathcal{R}(t)$ then for any $t, t' \geq 0$, the distribution of $\mathcal{R}^g(t+t')$ is the law of Coag$(\mathcal{R}^g(t), \tilde{\pi}^g)$, where $\tilde{\pi}^g$ is a random partition such that $\sigma(\tilde{\pi}^g) \overset{d}{=} \tilde{\pi}^g$ for any permutation σ preserving $\bar{\mathcal{R}}(t)$ i.e., such that*

$$i \overset{\bar{\mathcal{R}}(t)}{\sim} j \Rightarrow \sigma(i) \overset{\bar{\mathcal{R}}(t)}{\sim} \sigma(j). \tag{3.2}$$

Another property is that the process $(\mathcal{R}^s(t),\ t \geq 0)$ is a simple exchangeable coalescent process, but we do not prove it at this point as it will be clear from Theorem 3.4.

Proof. The first point of the proposition is immediate considering *iii)* of Definition 3.1.

As for the second point, recall that $\mathcal{N}_\infty$ is endowed with the topology generated by the sets of the form $\{\pi \in \mathcal{N}_\infty,\ \pi_{|n} = \widehat{\pi}\}$, for $n \in \mathbb{N}$, $\widehat{\pi} \in \mathcal{N}_n$. It is easy to see that this topology is metrized by $d(\pi, \pi') := (\sup\{n \in \mathbb{N},\ \pi_{|n} = \pi'_{|n}\})^{-1}$ (with $(\sup \mathbb{N})^{-1} = 0$) and that $(\mathcal{N}_\infty, d)$ is compact.

We need to show that for any continuous (then bounded) function $f : \mathcal{N}_\infty \to \mathbb{R}$, the function $P_t f : \pi \mapsto \mathbb{E}_\pi f(\mathcal{R}(t))$ (where $\mathbb{E}_\pi(\cdot) = \mathbb{E}(\cdot \,|\, \mathcal{R}(0) = \pi)$) is continuous, and that $P_t f(\pi) \to f(\pi)$ as $t \to 0$. By definition the process is càdlàg so we have almost surely $f(\mathcal{R}(t)) \to f(\mathcal{R}(0))$ so clearly by taking expectations $P_t f(\pi) \to f(\pi)$ as $t \to 0$. Now to show that $P_t f$ is continuous, consider $n \in \mathbb{N}$ and let $\{\widehat{\pi}^1, \ldots, \widehat{\pi}^k\}$ be an enumeration of $\mathcal{N}_n$. We pick $\pi^1, \ldots, \pi^k \in \mathcal{N}_\infty$ such that $\pi^i_{|n} = \widehat{\pi}^i$, and define $\widehat{f}_n : \mathcal{N}_\infty \to \mathbb{R}$ by

$$\widehat{f}_n(\pi) = f(\pi^i) \quad \text{if } \pi_{|n} = \widehat{\pi}^i.$$

Now since f is continuous on $(\mathcal{N}_\infty, d)$ which is compact, f is uniformly continuous, which means that

$$\omega_n := \sup_{\pi \in \mathcal{N}_\infty} |f(\pi) - \widehat{f}_n(\pi)| \to 0 \quad \text{as } n \to \infty.$$

For $t > 0$ and $\pi, \pi' \in \mathcal{N}_\infty$, we have

$$|P_t f(\pi) - P_t f(\pi')| \leq \left| \mathbb{E}_\pi \widehat{f}_n(\mathcal{R}(t)) - \mathbb{E}_{\pi'} \widehat{f}_n(\mathcal{R}(t)) \right| + 2\omega_n. \tag{3.3}$$

Now suppose $\pi_{|n} = \pi'_{|n}$. Since $\widehat{f}_n$ depends only on $\pi_{|n}$ and by Lemma 3.2 the process $\mathcal{R}_{|n}$ has the same distribution under $\mathbb{P}_\pi$ or $\mathbb{P}_{\pi'}$, we have the equality $\mathbb{E}_\pi \widehat{f}_n(\mathcal{R}(t)) = \mathbb{E}_{\pi'} \widehat{f}_n(\mathcal{R}(t))$, and plugging that into (3.3), we get

$$\sup\{|P_t f(\pi) - P_t f(\pi')|,\ \pi, \pi' \in \mathcal{N}_\infty,\ \pi_{|n} = \pi'_{|n}\} \leq 2\omega_n \to 0 \quad \text{as } n \to \infty,$$

showing that $P_t f$ is continuous.

For the third point of the proposition, $\mathcal{R}^g(t+t')$ is clearly of the form Coag$(\mathcal{R}^g(t), \tilde{\pi}^g)$, where $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ is a random recipe partition whose distribution depends on $\mathcal{R}(t)$ and t'. Let us show that the conditional distribution of $\tilde{\pi}$ given $\mathcal{R}(t)$ is invariant under the action of permutations preserving $\bar{\mathcal{R}}(t)$.

Without loss of generality, we can work under the conditioning $\{\mathcal{R}(t) = (\varrho, \mathbf{0}_\infty)\}$, where ϱ is any partition and $\mathbf{0}_\infty$ is the partition into singletons, so that for all π, we have $\mathrm{Coag}(\mathcal{R}^g(t), \pi) = \pi$. In particular, note that in this case we have $\mathcal{R}^g(t+t') = \tilde{\pi}^g$, and $\bar{\mathcal{R}}(t) = \varrho$. Let σ be a permutation such that $\sigma(\varrho) = \varrho$. The problem then reduces to showing that

$$(\sigma(\mathcal{R}^g(t+t')) \mid \mathcal{R}(t) = (\varrho, \mathbf{0}_\infty)) \overset{(d)}{=} (\mathcal{R}^g(t+t') \mid \mathcal{R}(t) = (\varrho, \mathbf{0}_\infty)),$$

which is now an immediate consequence of *iii)* in Definition 3.1. □

Let us now investigate the transition rates of the Markov chains $\mathcal{R}_{|n}$ appearing in Lemma 3.2, for every $n \in \mathbb{N}$. In this direction fix $n \in \mathbb{N}$, let $\varrho \in \mathcal{N}_\infty$ and $\pi \in \mathcal{N}_n$ and denote the jump rate of $\mathcal{R}_{|n}$ from $\varrho_{|n}$ to π by

$$q_n(\varrho, \pi) := \lim_{t \to 0+} \frac{1}{t} \mathbb{P}_\varrho(\mathcal{R}_{|n}(t) = \pi) \tag{3.4}$$

where $\mathbb{P}_\varrho(\cdot) = \mathbb{P}(\cdot \,|\, \mathcal{R}(0) = \varrho)$. The index n is not necessary in the notation as it can be read in the partition π. However we keep it as it will ease reading. Remind that $q_n(\varrho, \pi)$ only depends on ϱ through $\varrho_{|n}$. As is remarked in Lemma 3.2, $q_n(\varrho, \pi)$ equals zero if π is not obtained from $\varrho_{|n}$ by coagulating blocks according to a partition in $\widetilde{\mathcal{P}}(\varrho_{|n})$, that is by merging some species blocks of $\varrho^s_{|n}$ into one and some gene blocks of the new species into one. Also observe that the rates do not depend on the sizes of the gene blocks in the starting configuration so there is no loss of generality if we consider that $\varrho^g = \mathbf{0}_\infty$, the trivial partition made of singletons. Of course changing the starting partition ϱ has some effect on the arrival partition π. This is why we will need to write transition rates in another way, giving more emphasis on the dependence of the coagulation mechanism upon the starting partition.

Fix $n \in \bar{\mathbb{N}}$ and suppose that $\mathcal{R}_{|n}$ starts from n singleton gene blocks allocated into b species. Since labels of genes do not affect the transition rates, we will keep the data of the number of genes in each species in a vector $\mathbf{g} = (g_1, \ldots, g_b)$. This vector suffices to describe the starting position. Indeed $|\mathbf{g}| = b$ gives the number of species and $\sum_{i=1}^b g_i = n$ gives the number of genes.

Now the coagulation mechanism will be described by two terms. We will say that a gene block *participates in the coalescence event* if it merges with other gene blocks. We will say that a species block *participates in the coalescence event* if it merges with other species blocks or if it contains gene blocks that participate in the coalescence event.

The behaviour of the species blocks will be encoded in a vector $\mathbf{s} = (s_1, \ldots, s_b)$ with coordinates taking values in $\{0, 1\}$. Namely, $s_i = 1$ if the i-th species participates in the coalescence event and $s_i = 0$ otherwise. The total number of species involved in the event is $k = \sum_{i=1}^b s_i$.

The behaviour of the gene blocks will be encoded by an array $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_b)$ where $\mathbf{c}_i$ is a vector describing which gene blocks in the i-th species participate in the coalescence event. If $s_i = 1$ (i-th species block participating in the coalescence event), then $\mathbf{c}_i = (c_{i1}, \ldots, c_{ig_i})$ is such that $c_{ij} = 1$ if j-th gene block inside i-th species block participates in the coalescence event and $c_{ij} = 0$ otherwise. If $s_i = 0$, the i-th species block is not participating in the event and so neither will the gene blocks within it. In this case we set $\mathbf{c}_i = (0, 0, \ldots, 0) = \mathbf{0}$ and nothing happens at the gene level. Note that the number of gene blocks participating in the coalescence event is $\sum_{i,j} c_{ij}$.

Note that all such arrays $(\mathbf{g}, \mathbf{s}, \mathbf{c})$ do not necessarily code for observable coalescence events, so we will define a restricted set of arrays of interest for our study. First, note that one needs to have $\sum_i s_i \geq 2$ in order to observe a species merger. If $\sum_i s_i = 1$, then there is a gene coalescence if and only if $\sum_{i,j} c_{ij} \geq 2$. Also, we will restrict ourselves

to the arrays $(\mathbf{g},\mathbf{s},\mathbf{c})$ such that $\sum_{i,j} c_{ij} \neq 1$, because *a sole gene coalescing* is not distinguishable from *no gene coalescing*.

Formally, we consider finite arrays $(\mathbf{g},\mathbf{s},\mathbf{c})$ satisfying the assumptions

$$\begin{aligned} &\text{If } |\mathbf{g}| = b, \text{ then } \mathbf{s} \in \{0,1\}^b \text{ and } \mathbf{c} = (\mathbf{c}_1,\dots,\mathbf{c}_b),\ \mathbf{c}_i \in \{0,1\}^{g_i}, \\ &\text{with } s_i = 0 \implies \forall j, c_{ij} = 0 \end{aligned} \tag{\textbf{H1}}$$

$$\sum_{i,j} c_{ij} < 2 \implies \sum_i s_i \geq 2, \tag{\textbf{H2}}$$

$$\text{and } \sum_{i,j} c_{ij} \neq 1. \tag{\textbf{H3}}$$

We denote by $\mathcal{C}$ the set of arrays $(\mathbf{g},\mathbf{s},\mathbf{c})$ satisfying (**H1**), (**H2**) and (**H3**).

We then denote the transition rate of $\mathcal{R}_{|n}$ from a partition described by $\mathbf{g}$ (such that $\sum g_i = n$) to a new partition obtained by merging species and genes according to $\mathbf{s}$ and $\mathbf{c}$ by

$$q_{b,k}(\mathbf{g},\mathbf{s},\mathbf{c}).$$

Here again indices b and k are not necessary but permit to read easily the coalescence event at the species level ($k = \sum s_i$ species merging among $b = |\mathbf{g}|$). We insist on the fact that we consider only arrays $(\mathbf{g},\mathbf{s},\mathbf{c}) \in \mathcal{C}$ when we study the rates $q_{b,k}(\mathbf{g},\mathbf{s},\mathbf{c})$, and that these quantities determine uniquely the law of a SNEC $\mathcal{R}$, since they describe completely the rates associated to each finite-space continuous-time Markov chain $\mathcal{R}_{|n}$.

We introduce a notation that we will use in the next result for ease of writing. For μ a probability on $[0,1]$, consider any probability space where $Z_1, Z_2, \dots$ are i.i.d. with distribution μ and denote the expectation $\mathbb{E}_\mu$. Now take a vector $(g_i,\, i \in S)$ of integers, where S is a finite subset of $\mathbb{N}$. We define

$$\begin{aligned} \mathcal{U}(\mu,(g_i, i\in S)) &= \mathbb{E}_\mu\Big[\sum_{i\in S} g_i Z_i (1-Z_i)^{g_i-1} \prod_{j\in S\,:\,j\neq i} (1-Z_j)^{g_j}\Big] \\ &= \sum_{i\in S} g_i \int_{[0,1]} \mu(dq) q (1-q)^{g_i-1} \prod_{j\in S\,:\,j\neq i} \int_{[0,1]} \mu(dq)(1-q)^{g_j}. \end{aligned} \tag{3.5}$$

This can be thought of as the probability that a random array $(c_{ij},\, i \in S, 1 \leq j \leq g_i)$ does not satisfy (**H3**), where conditional on $(Z_i,\, i \in S)$ the variables (c_{ij}) are independent, and for all i,j, $c_{ij} = 1$ with probability Z_i. We can now state our main result.

Theorem 3.4. *There exist two non-negative real numbers $a_s, a_g \geq 0$ and two measures:*

- *ν_s on $E = (0,1] \times \mathcal{M}_1([0,1])$;*
- *ν_g on $(0,1]$;*

such that

$$\int_E \nu_s(dp,d\mu) p^2 < \infty, \tag{3.6a}$$

$$\int_E \nu_s(dp,d\mu) p \int_{[0,1]} \mu(dq) q^2 < \infty, \tag{3.6b}$$

$$\text{and } \int_{(0,1]} \nu_g(dq) q^2 < \infty, \tag{3.7}$$

and such that for any array $(\mathbf{g},\mathbf{s},\mathbf{c}) \in \mathcal{C}$ such that $|\mathbf{g}| = b$, $\sum_i s_i = k$ and $\sum_j c_{ij} = l_i$,

$$
\begin{aligned}
q_{b,k}(\mathbf{g},\mathbf{s},\mathbf{c}) = \int_E \nu_s(dp,d\mu)p^k(1-p)^{b-k}\Bigg(&\prod_{i\,:\,s_i=1}\int_{[0,1]}\mu(dq)q^{l_i}(1-q)^{g_i-l_i}\\
&+\mathbb{1}\{\mathbf{c}=\mathbf{0}\}\mathcal{U}(\mu,(g_i,\,1\le i\le b \text{ with } s_i=1))\Bigg)\\
&+a_s\mathbb{1}\{k=2,\mathbf{c}=\mathbf{0}\}\\
&+\mathbb{1}\{k=1\}\left(a_g\mathbb{1}\{l_I=2\}+\int_{(0,1]}\nu_g(dq)q^{l_I}(1-q)^{g_I-l_I}\right),
\end{aligned}
\tag{3.8}
$$

where the functional $\mathcal{U}$ is defined in (3.5) *and $I=I(\mathbf{g},\mathbf{s},\mathbf{c})$, in the case $k=1$, is the unique index in $\{1,2,\dots,b\}$ such that $s_I=1$.*

Furthermore, this correspondence between laws of SNEC processes and quadruplets (a_s,a_g,ν_s,ν_g) satisfying (3.6) *and* (3.7) *is bijective.*

Remark 3.5. We will show the *surjective* part of the theorem's last statement in Section 5, using an explicit Poissonian construction. For now we prove the existence and uniqueness of the characteristics (a_s,a_g,ν_s,ν_g).

4 Proof of Theorem 3.4

Consider a SNEC process $\mathcal{R}=((\mathcal{R}^s(t),\mathcal{R}^g(t)),\ t\ge 0)$ with values in $\mathcal{N}_\infty$ and recall its jump rates $q_n(\varrho,\pi)$ defined in (3.4). Also recall the alternative notation $q_{b,k}(\mathbf{g},\mathbf{s},\mathbf{c})$. Here, $\mathbf{g}$ is a vector of size b such that $\sum g_i=n$, $\mathbf{s}$ is a vector having the same size as $\mathbf{g}$ with coordinates in $\{0,1\}$ such that $\sum s_i=k$, and $\mathbf{c}$ is a family of $|\mathbf{g}|$ elements denoted by $\mathbf{c}_1,\mathbf{c}_2\dots$ where $\mathbf{c}_i$ is a vector of $\{0,1\}^{g_i}$ if $s_i=1$ and $\mathbf{c}_i=\mathbf{0}$ if $s_i=0$.

Lemma 4.1. *For any initial value $\varrho=(\varrho_s,\varrho_g)\in\mathcal{N}_\infty$, there exists a unique measure μ_ϱ on $\mathcal{N}_\infty$ such that*

$$
\mu_\varrho(\{\varrho\})=0 \quad \textit{and} \quad \forall n\ge 1,\ \mu_\varrho(\Pi_{|n}\neq\varrho_{|n})<\infty \tag{4.1}
$$

and such that the transition rate of the Markov chain $\mathcal{R}_{|n}$ from $\varrho_{|n}$ to $\pi\in\mathcal{N}_n$ is given by

$$
q_n(\varrho,\pi)=\mu_\varrho(\Pi_{|n}=\pi). \tag{4.2}
$$

Furthermore, for any permutation $\sigma:\mathbb{N}\to\mathbb{N}$,

$$
\mu_\varrho(\sigma(\Pi)\in\cdot)=\mu_{\sigma(\varrho)}(\Pi\in\cdot). \tag{4.3}
$$

Note that we write $\mu_\varrho(\Pi\in A)$ instead of $\mu_\varrho(A)$ because we implicitly work on the canonical space $\mathcal{N}_\infty$ and we denote by Π the generic element of $\mathcal{N}_\infty$.

Proof. Let $n<m$. We first note that since $\mathcal{R}_{|m}$ and $\mathcal{R}_{|n}=(\mathcal{R}_{|m})_{|n}$ are Markov chains, the transition rates can be expressed, for any $\pi\in\mathcal{N}_n\setminus\{\varrho_{|n}\}$,

$$
q_n(\varrho,\pi)=\sum_{\pi'\in\mathcal{N}_m\,:\,\pi'_{|n}=\pi} q_m(\varrho,\pi'). \tag{4.4}
$$

Let us now check that this consistency property along with Carathéodory's extension theorem ensures us that there exists a measure μ_ϱ on $\mathcal{N}_\infty\setminus\{\varrho\}$ satisfying (4.2).

Here the family $\mathcal{A}:=\big\{\{\Pi_{|n}=\pi\},\ n\in\mathbb{N},\pi\in\mathcal{N}_n\setminus\{\varrho_{|n}\}\big\}\cup\{\emptyset\}$ clearly forms a semi-ring of subsets of $\mathcal{N}_\infty$, and it remains to check that the functional $\widetilde{\mu}:\mathcal{A}\to[0,+\infty]$, defined by

$$
\widetilde{\mu}(\emptyset):=0 \quad \text{and} \quad \widetilde{\mu}(\{\Pi_{|n}=\pi\}):=q_n(\varrho,\pi),
$$

is a pre-measure. Equation (4.4) shows that $\widetilde{\mu}$ is finitely additive, and the only difficulty lies in understanding that $\widetilde{\mu}$ is countably additive. Now observe that the topology of $\mathcal{N}_\infty\setminus\{\varrho\}$ is generated by $\mathcal{A}$, and that each of the non-empty sets in $\mathcal{A}$ is both open and

closed (thus compact), because

$$\mathcal{N}_\infty \setminus \{\Pi_{|n} = \pi\} = \bigcup_{\varrho \in \mathcal{N}_n \setminus \{\pi\}} \{\Pi_{|n} = \varrho\}.$$

This implies that if $(A_n)_{n\geq 1}$ is a family of pairwise disjoint elements of $\mathcal{A}$ such that $\bigcup_n A_n \in \mathcal{A}$, then at most a finite number of the A_n are non-empty (because since $\bigcup_n A_n$ is compact, there is a finite subcover), so countable additivity reduces to finite additivity. Therefore Carathéodory's extension theorem applies, hence the existence of a measure μ_ϱ on $\mathcal{N}_\infty \setminus \{\varrho\}$ satisfying (4.2).

Considering μ_ϱ as a measure on $\mathcal{N}_\infty$ such that $\mu_\varrho(\{\varrho\}) = 0$, we check easily (4.1) by noticing that

$$\mu_\varrho(\Pi_{|n} \neq \varrho_{|n}) = \sum_{\pi \in \mathcal{N}_n \setminus \{\varrho_{|n}\}} q_n(\varrho, \pi) < \infty.$$

Furthermore, for any n, $\pi \in \mathcal{N}_n \setminus \{\varrho_{|n}\}$ and $\sigma : \mathbb{N} \to \mathbb{N}$ permutation, we have by the exchangeability property (3.1) of a SNEC, that

$$\mu_\varrho(\sigma(\Pi)_{|n} = \pi) = \lim_{t\to 0} \frac{1}{t} \mathbb{P}_\varrho(\sigma(\mathcal{R}(t))_{|n} = \pi) = \lim_{t\to 0} \frac{1}{t} \mathbb{P}_{\sigma(\varrho)}(\mathcal{R}_{|n}(t) = \pi) = \mu_{\sigma(\varrho)}(\Pi_{|n} = \pi),$$

which proves that (4.3) holds on $\mathcal{A}$. Since the topology of $\mathcal{N}_\infty$ is generated by $\mathcal{A}$, the proof is complete. □

The latter lemma implies that there exists a family of exchangeable measures on $\mathcal{N}_\infty$ characterizing (i.e. acting as an analog of a Markov kernel for continuous-space pure-jump Markov chains) the SNEC process $\mathcal{R}$. Furthermore, since we are dealing with a simple coalescent, it is clear from the characterization (4.2) that μ_ϱ is simple in the sense that it is supported by all the possible bivariate partitions obtained from a simple coalescence from ϱ. To put it simply,

$$\mu_\varrho\left(\mathcal{N}_\infty \setminus \{\text{Coag}_2(\varrho, \tilde{\pi}),\ \tilde{\pi} \in \widetilde{\mathcal{P}}(\varrho)\}\right) = 0.$$

The measure μ_ϱ can be translated as a measure on arrays of random variables in $\{0, 1\}$. Informally, we can associate to each species in ϱ a 1 entry if it participates in the coalescence and a 0 entry otherwise. Inside the species participating to the coalescence event, we can also associate a 1 entry to the genes participating in the coalescence event and a 0 entry otherwise. To tally with the definition of the SNEC we will need a certain partial exchangeability structure for this array. This picture can be formalized as follows. Let $((X_i, (Y_{ij},\ j \in \mathbb{N})),\ i \in \mathbb{N})$ be an array of Bernoulli random variables and denote by Z_i the i-th line vector $(X_i, (Y_{ij},\ j \in \mathbb{N}))$. We say that this array is *hierarchically exchangeable* if

(A1) the family $(Z_i,\ i \in \mathbb{N})$ is exchangeable;

(A2) for any $i \in \mathbb{N}$, the family $(X_i, (Y_{ij},\ j \in \mathbb{N}))$ is invariant under any permutation over the j's.

We also naturally extend this definition to measures on the space $(\{0, 1\} \times \{0, 1\}^{\mathbb{N}})^{\mathbb{N}}$. We say that such a measure μ is *hierarchically exchangeable* if it is invariant both under the permutations of the rows, and the permutations within a row.

For an initial state $\varrho = (\varrho^s, \varrho^g) \in \mathcal{N}_\infty$ and an arrival state $\pi = \text{Coag}_2(\varrho, \tilde{\pi}) \in \mathcal{N}_\infty$, with $\tilde{\pi}$ a simple bivariate partition $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g) \in (\mathcal{P}'_\infty)^2$, define the array $\mathbf{Z}(\varrho, \pi) = (\mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2, \ldots)$ by

$$\begin{aligned} &X_i = 1 \text{ if the } i\text{-th block in } \varrho^s \text{ has coalesced in } \pi, \\ &Y_{ij} = 1 \text{ if the } I(i,j)\text{-th block in } \varrho^g \text{ has coalesced in } \pi, \end{aligned} \tag{4.5}$$

where $I(i, j) := k$ if the k-th block of ϱ^g is the j-th gene block of the i-th species block.

Now choose a state ϱ with an infinite number of species blocks, each containing an infinite number of gene blocks. Let ν be the push-forward of μ_ϱ by the application

$$\pi \longmapsto \mathbf{Z}(\varrho, \pi).$$

Then the exchangeability property of μ_ϱ (4.3) implies that ν is a hierarchically exchangeable measure on $(\{0,1\} \times \{0,1\}^{\mathbb{N}})^{\mathbb{N}}$, and (4.1) implies that

$$\nu(\mathbf{Z} = \mathbf{0}) = 0, \quad \text{and} \quad \nu\left(\sum_{i=1}^{n} X_i \geq 2 \text{ or } \exists i \leq n, \sum_{j=1}^{n} Y_{ij} \geq 2\right) < \infty, \tag{4.6}$$

where $\mathbf{0}$ denotes the null array on $(\{0,1\} \times \{0,1\}^{\mathbb{N}})^{\mathbb{N}}$. Also, note that the application $(\mu_\varrho,\ \varrho \in \mathcal{N}_\infty) \mapsto \nu$ is one-to-one. Indeed, we can conversely define for any $\mathbf{Z}$ and any nested partition $\varrho \in \mathcal{N}_\infty$, the nested partition $C(\varrho, \mathbf{Z}) \in \mathcal{N}_\infty$ obtained from ϱ by merging exactly the blocks that *participate in the coalescence* where

- The i-th block of ϱ^s participates iff $X_i = 1$;
- The j-th block in ϱ^g of the i-th block of ϱ^s participates iff $X_i = 1$ and $Y_{ij} = 1$.

With this definition, μ_ϱ is obtained as the push-forward of ν by the map $\mathbf{Z} \mapsto C(\varrho, \mathbf{Z})$.

Now recall the alternative notation $q_{b,k}(\mathbf{g}, \mathbf{s}, \mathbf{c})$ for the transition rate of $\mathcal{R}_{|n}$ (where $n = \sum_i g_i$) from a nested partition with b species blocks and $g_1, \ldots, g_b$ gene blocks inside them, to a nested partition obtained by merging k species blocks according to the vector $\mathbf{s}$ and gene blocks inside those species according to the array $\mathbf{c}$. For any array $(\mathbf{g}, \mathbf{s}, \mathbf{c}) \in \mathcal{C}$, note that (4.2) translates in terms of our push-forward ν in the following way:

$$\begin{aligned} q_{b,k}(\mathbf{g}, \mathbf{s}, \mathbf{c}) &= \nu\big(\forall 1 \leq i \leq b,\ X_i = s_i, \text{ and } \forall 1 \leq j \leq g_i, Y_{ij} = c_{ij}\big) \\ &\quad + \mathbb{1}_{\{\mathbf{c}=\mathbf{0}\}} \nu\left(\forall 1 \leq i \leq b,\ X_i = s_i, \text{ and } \sum_{i=1}^{b} \sum_{j=1}^{g_i} Y_{ij} = 1\right). \end{aligned} \tag{4.7}$$

Indeed, the first line is quite straightforward and comes from our representation of coalescence events by those arrays $(\mathbf{g}, \mathbf{s}, \mathbf{c}) \in \mathcal{C}$ (see Section 3) which basically means that blocks participating in a coalescence event are those associated with a 1. However in the case when $\mathbf{c} = \mathbf{0}$, there is an additional probability to observe the coalescence of species blocks associated to $\mathbf{s}$ with no coalescence of gene blocks (the case when all the Y_{ij}'s are 0 is included in the first term), which is when exactly one of the Y_{ij}'s is equal to 1. This gives rise to the second line of (4.7).

We now have to establish a de Finetti representation of hierarchically exchangeable arrays to express the measure of an event of the form $\{\forall 1 \leq i \leq b,\ X_i = s_i, \text{ and } \forall 1 \leq j \leq g_i, Y_{ij} = c_{ij}\}$. Note that we consider random measures in the following, but only on Borel spaces $(S, \mathcal{S})$ (i.e. spaces isomorphic to a Borel subset of $\mathbb{R}$ endowed with the Borel σ-algebra), which will enable us to use de Finetti's theorem [23]. For this we write $\mathcal{M}_1(S)$ for the space of probability measures on S, which is endowed with the σ-algebra generated by the maps $\mu \mapsto \mu(B)$ for all $B \in \mathcal{S}$. The spaces $(S, \mathcal{S})$ that we consider will be for instance $[0,1]$ with its Borel sets or $\{0,1\}^{\mathbb{N}}$ equipped with the product σ-algebra, which are clearly Borel spaces.

Proposition 4.2. *Let $\mathbf{Z} = (Z_i,\ i \in \mathbb{N}) = ((X_i, (Y_{ij},\ j \in \mathbb{N})),\ i \in \mathbb{N})$ be a hierarchically exchangeable array (with variables in $\{0,1\}$). Then there exists a unique probability measure Λ on $E' = [0,1] \times \mathcal{M}_1([0,1]) \times \mathcal{M}_1([0,1])$ (and we will write any element μ of E'*

as (p, μ_0, μ_1)*) such that for all* $n \geq 1$

$$\begin{aligned}
&\mathbb{P}(X_i = x_i, Y_{ij} = y_{ij},\ i,j \in [n]) \\
&= \int_{E'} \Lambda(d\mu) \prod_{i=1}^{n} \Bigg[(p\mathbb{1}_{\{x_i=1\}} + (1-p)\mathbb{1}_{\{x_i=0\}}) \\
&\qquad\qquad \int_{[0,1]} \mu_{x_i}(dq_i) \prod_{j=1}^{n} (q_i \mathbb{1}_{\{y_{ij}=1\}} + (1-q_i)\mathbb{1}_{\{y_{ij}=0\}}) \Bigg].
\end{aligned} \tag{4.8}$$

Proof. Let us first observe that if a sequence $(X, (Y_j,\ j \in \mathbb{N}))$ satisfies Hypothesis **(A2)**, then, conditional on $X = x \in \{0,1\}$, the sequence $(Y_j,\ j \in \mathbb{N})$ is exchangeable. We can thus apply de Finetti's theorem: conditional on $X = x$ there is a unique probability measure μ_x giving the distribution of the asymptotic frequency q of the variables $(Y_j,\ j \in \mathbb{N})$, and conditional on q they are i.i.d. Bernoulli with parameter q. This implies that, for any $\{0,1\}$-valued finite sequence $(x, y_1, y_2, \ldots, y_k)$,

$$\begin{aligned}
&\mathbb{P}(X = x, Y_1 = y_1, \ldots, Y_k = y_k) \\
&= \mathbb{P}(X = x) \int_{[0,1]} \mu_x(dq) \prod_{j=1}^{k} (q\mathbb{1}_{\{y_j=1\}} + (1-q)\mathbb{1}_{\{y_j=0\}}).
\end{aligned} \tag{4.9}$$

Also observe that since X is binary, there exists $p \in [0,1]$ such that $\mathbb{P}(X = x) = p\mathbb{1}_{\{x=1\}} + (1-p)\mathbb{1}_{\{x=0\}}$.

As a consequence of Hypothesis **(A1)**, we can apply once again de Finetti's theorem: there exists a unique law $\tilde{\Lambda}$ on $\mathcal{M}_1(\{0,1\}^{\mathbb{N}})$ such that the law of $(Z_i,\ i \in \mathbb{N})$ equals $\int_{\mathcal{M}_1(\{0,1\}^{\mathbb{N}})} \tilde{\Lambda}(d\tilde{\mu}) \bigotimes_{i \geq 1} \tilde{\mu}$. Furthermore it has been seen that $\tilde{\mu}$ can be expressed as in (4.9).

Now let F stand for the measurable mapping such that $F(\tilde{\mu}) = (p, \mu_0, \mu_1) \in E'$ and let Λ be the push-forward of $\tilde{\Lambda}$ by the mapping F. We obtain that if A and $(B_i,\ i \in A)$ are finite subsets of $\mathbb{N}$, then

$$\begin{aligned}
&\mathbb{P}(X_i = x_i, Y_{ij_i} = y_{ij_i},\ i \in A, j_i \in B_i) \\
&= \int_{E'} \Lambda(d\mu) \prod_{i \in A} \Bigg[(p\mathbb{1}_{\{x_i=1\}} + (1-p)\mathbb{1}_{\{x_i=0\}}) \\
&\qquad\qquad \int_{[0,1]} \mu_{x_i}(dq_i) \prod_{j_i \in B_i} (q_i \mathbb{1}_{\{y_{ij_i}=1\}} + (1-q_i)\mathbb{1}_{\{y_{ij_i}=0\}}) \Bigg].
\end{aligned}$$

This ends the proof. □

This result is almost enough to express (4.7) but one has to be careful because the measure ν might not be finite. However, it is σ-finite because by (4.6),

$$\nu = \lim_{n \to \infty} \uparrow \nu \left(\left\{ \sum_{i=1}^{n} X_i \geq 2 \text{ or } \exists i \leq n, \sum_{j=1}^{n} Y_{ij} \geq 2 \right\} \cap \cdot \right),$$

and those events have finite measure. The idea behind the following lemma is to make use of those events and hierarchical exchangeability to express ν as a limit of finite measures which, thanks to an application of Proposition 4.2, have a representation under the form (4.8).

Let us introduce some notation that will enable us to make this argument formal. For a fixed vector $(\mathbf{g}, \mathbf{s}, \mathbf{c}) \in \mathcal{C}$, such that $|\mathbf{g}| = b$, let us examine the event

$$A = A(\mathbf{g}, \mathbf{s}, \mathbf{c}) := \{\forall 1 \leq i \leq b,\ X_i = s_i, \text{ and } \forall 1 \leq j \leq g_i, Y_{ij} = c_{ij}\}$$

and its measure $\nu(A)$. Let us define, for all $n \geq 1$ the shifted random array

$$\mathbf{Z}_n := (X_{i+n}, Y_{(i+n)j},\ i,j \in \mathbb{N}). \tag{4.10}$$

We decompose naturally $A = (A \cap \{\mathbf{Z}_b \neq \mathbf{0}\}) \cup (A \cap \{\mathbf{Z}_b = \mathbf{0}\})$, where $b = |\mathbf{g}|$.

Recall that the array $\mathbf{Z}$ encodes which species blocks and which gene blocks are participating in a coalescence. Therefore the event $A \cap \{\mathbf{Z}_b \neq 0\}$ indicates that there are merging species blocks outside of the first b blocks. In fact we will see that this implies that such merging blocks are infinitely many (a random proportion p of them), and within each of these blocks, a random proportion q of gene blocks are also participating in the coalescence event. The following technical lemma makes this statement rigorous.

Lemma 4.3. *For an array $(\mathbf{g}, \mathbf{s}, \mathbf{c})$ satisfying assumptions* (**H1**) *and* (**H2**)*, there exists a unique measure ν_s on $E = (0,1] \times \mathcal{M}([0,1])$ such that*

$$\nu(A \cap \{\mathbf{Z}_b \neq \mathbf{0}\}) = \int_E \nu_s(dp, d\mu) p^k (1-p)^{b-k} \prod_{i\,:\,s_i=1} \int_{[0,1]} \mu(dq) q^{l_i} (1-q)^{g_i - l_i}, \tag{4.11}$$

where $b := |\mathbf{g}|$, $k := \sum_i s_i$ and $l_i := \sum_j c_{ij}$. Moreover, ν_s satisfies (3.6)*.*

Proof. We define some events that will be used to express $\nu(A \cap \{\mathbf{Z}_b \neq \mathbf{0}\})$.

$$A_n = A_n(\mathbf{g}, \mathbf{s}, \mathbf{c}) := \{\forall 1 \leq i \leq b,\ X_{i+n} = s_i,\ \text{and } \forall 1 \leq j \leq g_i, Y_{(i+n)j} = c_{ij}\}$$

$$B_n := \left\{ \sum_{i=1}^{n} X_i \geq 2 \right\}$$

$$B'_n = B'_n(\mathbf{g}, \mathbf{s}, \mathbf{c}) := \left\{ \sum_{i=b+1}^{b+n} X_i \geq 2 \right\}.$$

Note that $(\mathbf{g}, \mathbf{s}, \mathbf{c})$ satisfies (**H1**) and (**H2**), so we have $A \subset \{\sum_{i=1}^m X_i \geq 2 \text{ or } \sum_{i,j=1}^m Y_{ij} \geq 2\}$ for $m = \max(b, g_1, \dots g_b)$. Now because ν satisfies (4.6), necessarily $\nu(A) < \infty$, which implies that

$$\nu(A \cap \{\mathbf{Z}_b \in \cdot\})$$

is a finite hierarchically exchangeable measure on $(\{0,1\} \times \{0,1\}^{\mathbb{N}})^{\mathbb{N}}$. The de Finetti representation (Proposition 4.2) implies that on the event A, $\mathbf{Z}_b$ is either $\mathbf{0}$, or has an infinite number of entries with value 1. In particular,

$$A \cap \{\mathbf{Z}_b \neq \mathbf{0}\} = A \cap \{\mathbf{Z}_b \text{ has at least two entries at } 1\},$$

therefore, there is the equality

$$A \cap \{\mathbf{Z}_b \neq \mathbf{0}\} = \bigcup_{n \geq 1} A \cap B'_n,$$

where the union is increasing. Therefore,

$$\begin{aligned} \nu(A \cap \{\mathbf{Z}_b \neq \mathbf{0}\}) &= \lim_{n\to\infty} \nu(A \cap B'_n). \\ &= \lim_{n\to\infty} \nu(B_n \cap A_n), \end{aligned}$$

where we used the hierarchical exchangeability of ν to get the second equality. Now we know from (4.6) and because ν is exchangeable that the measure

$$\nu(B_n \cap \{\mathbf{Z}_n \in \cdot\})$$

is a finite hierarchically exchangeable measure on $(\{0,1\} \times \{0,1\}^{\mathbb{N}})^{\mathbb{N}}$. Because it is finite we can apply Proposition 4.2 to deduce that there exists a finite measure Λ_n on $E' = (0,1] \times \mathcal{M}([0,1])^2$ such that

$$\nu(B_n \cap A_n) = \int_{E'} \Lambda_n(dp, d\mu_0, d\mu_1) \prod_{i=1}^{b} \Bigg[(p\mathbb{1}_{\{s_i=1\}} + (1-p)\mathbb{1}_{\{s_i=0\}}) \int_{[0,1]} \mu_{s_i}(dq_i) \prod_{j=1}^{g_i} (q_i \mathbb{1}_{\{c_{ij}=1\}} + (1-q_i)\mathbb{1}_{\{c_{ij}=0\}}) \Bigg].$$

We can simplify this expression since ν is supported by the set $\{\forall i \in \mathbb{N},\, X_i = 0 \Rightarrow \forall j \in \mathbb{N}, Y_{ij} = 0\}$. This implies that Λ_n-a.e. the measure μ_0 is δ_0 the Dirac measure at 0. Therefore we write $\widetilde{\Lambda}_n$ for the push forward measure on $E := (0,1] \times \mathcal{M}([0,1])$ of Λ_n by the application $(p, \mu_0, \mu_1) \mapsto (p, \mu_1)$. We now have

$$\nu(B_n \cap A_n) = \int_E \widetilde{\Lambda}_n(dp, d\mu) p^k (1-p)^{b-k} \prod_{i\,:\,s_i=1} \int_{[0,1]} \mu(dq) q^{l_i} (1-q)^{g_i - l_i}. \tag{4.12}$$

To be able to pass to the limit, let us check that the sequence of measures $(\widetilde{\Lambda}_n)$ is increasing. Indeed, recall that Λ_n is obtained from two applications of de Finetti's theorem to the exchangeable array $\mathbf{Z}_n$, so the asymptotic parameters p and μ appearing in (4.12) are a deterministic, measurable functional of $\mathbf{Z}_n$. Let us write this functional $F(\mathbf{Z}_n) = (p, \mu)$, so now $\widetilde{\Lambda}_n$ is simply the measure

$$\nu(B_n \cap \{F(\mathbf{Z}_n) \in \cdot\}).$$

But p and μ are asymptotic quantities of the array $\mathbf{Z}_n$, which do not depend on the first row of $\mathbf{Z}_n$, so $F(\mathbf{Z}_{n+1}) = F(\mathbf{Z}_n)$ and we have

$$\begin{aligned} \widetilde{\Lambda}_n &= \nu(B_n \cap \{F(\mathbf{Z}_n) \in \cdot\}) \\ &= \nu(B_n \cap \{F(\mathbf{Z}_{n+1}) \in \cdot\}) \\ &\leq \nu(B_{n+1} \cap \{F(\mathbf{Z}_{n+1}) \in \cdot\}) \\ &= \widetilde{\Lambda}_{n+1}, \end{aligned}$$

where the passage from the second to the third line is simply because $B_n \subset B_{n+1}$. Therefore there is a limiting measure ν_s on E such that

$$\nu(A \cap \{\mathbf{Z}_b \neq \mathbf{0}\}) = \lim_{n\to\infty} \nu(B_n \cap A_n) = \int_E \nu_s(dp, d\mu) p^k (1-p)^{b-k} \prod_{i\,:\,s_i=1} \int_{[0,1]} \mu(dq) q^{l_i} (1-q)^{g_i - l_i},$$

so we recover (4.11). To prove the uniqueness of this measure, consider any measure ν'_s on E such that (4.11) holds. Then we have simply

$$\tilde{\Lambda}_n(dp, d\mu) = \nu(B_n \cap \{F(\mathbf{Z}_n) \in (dp, d\mu)\}) = (1 - (1-p)^n - np(1-p)^{n-1}) \nu'_s(dp, d\mu),$$

where the first equality is by definition and the second because we assumed that (4.11) holds for ν'_s. Taking limits on both sides yields

$$\nu_s(dp, d\mu) = \nu'_s(dp, d\mu).$$

It remains to prove (3.6). Note that the condition (4.6) implies that

$$\nu(X_1 = X_2 = 1) < \infty \quad \text{and} \quad \nu(X_1 = 1, Y_{1,1} = Y_{1,2} = 1) < \infty.$$

Translating these conditions with the formula (4.11), we recover exactly (3.6). □

Let us now examine $\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\})$. Recall that the event $A \cap \{\mathbf{Z}_b = 0\}$ indicates that there are no other merging species blocks than those within the first b blocks. The next lemma shows that this implies that we are either in a Kingman-type coalescence (a pair of species blocks are merging, occurring at rate a_s, or a pair of gene blocks within one species are merging, occurring at rate a_g), or in a multiple gene coalescence within a single species block (in which case a random proportion q of gene blocks are merging).

The key idea is to use exchangeability and the σ-finiteness property (4.6) of the measure ν to show by contradiction that $\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\})$ is zero in certain cases.

Lemma 4.4. *For an array $(\mathbf{g}, \mathbf{s}, \mathbf{c})$ satisfying assumptions* (**H1**) *and* (**H2**)*, there exist unique real numbers $a_s, a_g \geq 0$ and a unique measure ν_g on $(0,1]$ satisfying* (3.7) *such that*

$$\begin{aligned}\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\}) &= a_s \mathbb{1}\{k = 2, \mathbf{c} = \mathbf{0}\} \\ &\quad + \mathbb{1}\{k = 1\} \left(a_g \mathbb{1}\{l_I = 2\} + \int_{(0,1]} \nu_g(dq) q^{l_I} (1-q)^{g_I - l_I} \right),\end{aligned} \tag{4.13}$$

where $b := |\mathbf{g}|$, $k := \sum_i s_i$, $l_i := \sum_j c_{ij}$ and in the case $k = 1$, I is the unique index in $\{1, 2, \ldots, b\}$ such that $s_I = 1$.

Proof. The measure $\nu(\mathbf{X} \in \cdot)$ is an exchangeable measure on $\{0,1\}^{\mathbb{N}}$ such that, because of (4.6), $\nu(X_1 = X_2 = X_3 = 1) < \infty$. Note that exchangeability implies that for any $n, i \geq 3$,

$$\nu(\{X_1 = X_2 = X_3 = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) = \nu(\{X_1 = X_2 = X_i = 1\} \cap \{\mathbf{Z}_i = \mathbf{0}\}), \tag{4.14}$$

But the events $(\{X_1 = X_2 = X_i = 1\} \cap \{\mathbf{Z}_i = \mathbf{0}\},\ i \geq 3)$ are disjoint and all included in $\{X_1 = X_2 = 1\}$, so that

$$\sum_{i \geq 3} \nu(\{X_1 = X_2 = X_i = 1\} \cap \{\mathbf{Z}_i = \mathbf{0}\}) \leq \nu(\{X_1 = X_2 = 1\}) < \infty.$$

From (4.14) we deduce $\nu(\{X_1 = X_2 = X_3 = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) = 0$. This implies immediately that for a finite array $(\mathbf{g}, \mathbf{s}, \mathbf{c})$ such that $k = \sum_i s_i > 2$, we have $\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\}) = 0$.

- In the case $k = 2$ (suppose $s_1 = s_2 = 1$), one must examine several cases.
  - Suppose first $c_{1,1} = c_{1,2} = 1$. This means that the first two gene blocks of the first species block coalesce while the first two species blocks coalesce. Then we note that for any $n, i \geq 2$,

    $$\begin{aligned}&\nu(\{X_1 = X_2 = 1,\, Y_{1,1} = Y_{1,2} = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) \\ &= \nu(\{X_1 = X_i = 1,\, Y_{1,1} = Y_{1,2} = 1\} \cap \{\mathbf{Z}_i = \mathbf{0}\}).\end{aligned}$$

    However,

    $$\sum_{i \geq 2} \nu(\{X_1 = X_i = 1,\, Y_{1,1} = Y_{1,2} = 1\} \cap \{\mathbf{Z}_i = \mathbf{0}\}) \leq \nu(\{Y_{1,1} = Y_{1,2} = 1\}) < \infty,$$

    so that necessarily $\nu(\{X_1 = X_2 = 1,\, Y_{1,1} = Y_{1,2} = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) = 0$. So in the case $c_{1,1} = c_{1,2} = 1$, we have $\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\}) = 0$.
  - Now suppose $c_{1,1} = c_{2,1} = 1$, and all the other c_{ij} are zero. From our previous point, note that

$$\nu(\{X_1 = X_2 = 1,\, Y_{1,1} = Y_{2,1} = 1,\ \text{and } \exists j \geq 2, Y_{1,j} = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) = 0,$$

which implies that the events $(\{X_1 = X_2 = 1,\, Y_{1,j} = Y_{2,1} = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\},\, j \geq 1)$ are ν-a.e. disjoint. Therefore for any $n \geq 2$,

$$\sum_{j \geq 1} \nu(\{X_1 = X_2 = 1,\, Y_{1,j} = Y_{2,1} = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) \leq \nu(\{X_1 = X_2 = 1\}) < \infty,$$

So necessarily $\nu(\{X_1 = X_2 = 1,\, Y_{1,j} = Y_{2,1} = 1\} \cap \{\mathbf{Z}_n = \mathbf{0}\}) = 0$. This implies that in the case $c_{1,1} = c_{2,1} = 1$, we have $\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\}) = 0$.

– The previous two points show that in the case $k = 2$, the only way to have $\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\}) > 0$ is if $\mathbf{c} = \mathbf{0}$. In that case, define

$$\begin{aligned} a_s &:= \nu(\{X_1 = X_2 = 1\} \cap \{\mathbf{Z}_2 = \mathbf{0}\}) \\ &= \nu(\{X_1 = X_2 = 1\} \cap \{\mathbf{Y} = \mathbf{0} \text{ and } \forall k \notin \{1, 2\},\, X_k = 0\}). \end{aligned}$$

Then by exchangeability, for all $i, j \in \mathbb{N}$ with $i \neq j$, we have

$$a_s = \nu(\{X_i = X_j = 1\} \cap \{\mathbf{Y} = \mathbf{0} \text{ and } \forall k \notin \{i, j\},\, X_k = 0\}),$$

and in conclusion, for any array $(\mathbf{g}, \mathbf{s}, \mathbf{c})$ such that $k = 2$, we have

$$\nu(A \cap \{\mathbf{Z}_b = \mathbf{0}\}) = \mathbb{1}_{\{\mathbf{c}=\mathbf{0}\}} a_s.$$

• In the case $k = 1$, suppose that $s_1 = 1$. On the event

$$\{X_1 = 1,\, X_2 = X_3 = \cdots = X_b = 0\} \cap \{\mathbf{Z}_b = \mathbf{0}\},$$

we have simply $\mathbf{Z}_1 = \mathbf{0}$, and then the measure

$$\nu' := \nu\big(\{(Y_{1,j})_{j \in \mathbb{N}} \in \cdot\} \cap \{X_1 = 1,\, \mathbf{Z}_1 = \mathbf{0}\}\big)$$

is an exchangeable measure on $\{0, 1\}^{\mathbb{N}}$ such that for all $n \in \mathbb{N}$, $\nu'\left(\sum_{j=1}^n Y_j \geq 2\right) < \infty$. Therefore (see for instance Bertoin [4]) there exist a unique constant $a_g \geq 0$ and ν_g a unique measure on $(0, 1]$ satisfying (3.7) such that ν' can be written

$$\nu'(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = a_g \mathbb{1}_{l=2} + \int_{(0,1]} \nu_g(dq) q^l (1 - q)^{n-l},$$

for any vector $(y_1, y_2, \ldots, y_n) \in \{0, 1\}^n \setminus \{\mathbf{0}\}$ such that $l := \sum_i y_i \geq 2$.

Putting all the previous considerations together yields (4.13). □

Now it remains to put together Lemma 4.3 and Lemma 4.4. Recall that we restricted the rate function q to arrays in $\mathcal{C}$, i.e. satisfying (**H1**) to (**H3**). The reason for assuming (**H3**) is that then we can always write $q_{b,k}(\mathbf{g}, \mathbf{s}, \mathbf{c})$ as in (4.7), that is

$$\begin{aligned} q_{b,k}(\mathbf{g}, \mathbf{s}, \mathbf{c}) &= \nu\big(\forall 1 \leq i \leq b,\, X_i = s_i,\ \text{and } \forall 1 \leq j \leq g_i, Y_{ij} = c_{ij}\big) \\ &\quad + \mathbb{1}_{\{\mathbf{c}=\mathbf{0}\}} \nu\left(\forall 1 \leq i \leq b,\, X_i = s_i,\ \text{and } \sum_{i=1}^{b} \sum_{j=1}^{g_i} Y_{ij} = 1\right). \end{aligned}$$

Using the previous two lemmas to decompose the two lines on the events $\{\mathbf{Z}_b \neq \mathbf{0}\}$ and $\{\mathbf{Z}_b = \mathbf{0}\}$, we obtain the formula (3.8), concluding the proof of Theorem 3.4.

5 Poissonian construction

The goal of the present section is to show how any simple nested exchangeable coalescent can be constructed from a Poisson point process. Consider two real coefficients $a_s, a_g \geq 0$ and two measures: ν_s on $E = (0,1] \times \mathcal{M}_1([0,1])$ satisfying (3.6), and ν_g on $(0,1]$, satisfying (3.7). Recall the measures $\mathtt{K}_s, \mathtt{K}_g, P^g_x$ and $P^s_{x,\mu}$ introduced in Section 2, and the measure $\nu(d\mathbf{Z})$ defined on the space $\widehat{E}$ of doubly indexed arrays of 0's and 1's $\mathbf{Z} = (\mathbf{X}, (\mathbf{Y}_i,\, i \geq 1)) = (X_i, Y_{ij},\, i,j \geq 1)$

$$\nu := a_s \mathtt{K}_s + a_g \mathtt{K}_g + \int_{(0,1]} \nu_g(dx)\, P^g_x + \int_{(0,1]\times\mathcal{M}_1([0,1])} \nu_s(dx, d\mu)\, P^s_{x,\mu}.$$

Note that ν characterizes the distribution of the SNEC through the relation (4.7). The key idea of the construction is that ν necessarily satisfies (4.6), which is easily shown using exchangeability and conditions (3.6) and (3.7). First, note that $\nu(\mathbf{Z} = \mathbf{0}) = 0$ is trivial from our definitions, and that a straightforward union bound yields

$$\begin{aligned}
&\nu\Big(\textstyle\sum_{i=1}^n X_i \geq 2 \text{ or } \exists i \leq n, \sum_{j=1}^n Y_{ij} \geq 2\Big) \\
&\qquad \leq \sum_{1\leq i<i'\leq n} \nu(X_i = X_{i'} = 1) + \sum_{i=1}^n \sum_{1\leq j<j'\leq n} \nu(X_i = Y_{ij} = Y_{ij'} = 1) \\
&\qquad = \frac{n(n-1)}{2}\nu(X_1 = X_2 = 1) + \frac{n^2(n-1)}{2}\nu(X_1 = Y_{1,1} = Y_{1,2} = 1),
\end{aligned}$$

therefore we need only check that these two quantities are finite. Now by definition, we have

$$\begin{aligned}
\mathtt{K}_s(X_1 = X_2 = 1) = 1, &\quad \mathtt{K}_s(X_1 = Y_{1,1} = Y_{1,2} = 1) = 0, \\
\mathtt{K}_g(X_1 = X_2 = 1) = 0, &\quad \mathtt{K}_g(X_1 = Y_{1,1} = Y_{1,2} = 1) = 1, \\
P^g_x(X_1 = X_2 = 1) = 0, &\quad P^g_x(X_1 = Y_{1,1} = Y_{1,2} = 1) = x^2, \\
P^s_{x,\mu}(X_1 = X_2 = 1) = x^2, &\quad P^s_{x,\mu}(X_1 = Y_{1,1} = Y_{1,2} = 1) = x\int_{[0,1]} \mu(dq) q^2.
\end{aligned}$$

Integrating the last two lines with respect to ν_g and ν_s and summing, we see that (3.6) and (3.7) imply that both $\nu(X_1 = X_2 = 1)$ and $\nu(X_1 = Y_{1,1} = Y_{1,2} = 1)$ are finite, proving (4.6).

Now to start the construction of our process, consider an initial partition $\pi_0 \in \mathcal{N}_\infty$. Let M be a Poisson point process on $(0,\infty)\times\widehat{E}$ with intensity $dt \otimes \nu(d\mathbf{Z})$. We will construct on the same probability space the processes $\mathcal{R}^n = (\mathcal{R}^n(t),\, t \geq 0)$, for $n \in \mathbb{N}$ thanks to M.

Recall that for any $\mathbf{Z} = (\mathbf{X}, (\mathbf{Y}_i,\, i \geq 1)) = (X_i, Y_{ij},\, i,j \geq 1)$ and any nested partition $\pi \in \mathcal{N}_n$, we denote by $C(\pi, \mathbf{Z})$ the nested partition of $\mathcal{N}_n$ obtained from π by merging exactly the blocks that *participate in the coalescence* where

- The i-th block of π^s participates iff $X_i = 1$;
- The j-th block in π^g of the i-th block of π^s participates iff $X_i = 1$ and $Y_{ij} = 1$.

Fix $n \in \mathbb{N}$, and let M_n denote the subset of M consisting of points $(t, \mathbf{Z})$ such that $\sum_{i=1}^n X_i \geq 2$ or $\exists i \leq n,\, X_i \sum_{j=1}^n Y_{ij} \geq 2$. Because of (4.6), there are only a finite number of such points with t in a compact set of $[0,+\infty)$. Therefore one can label the atoms of the set $M_n := \{(t_k, \mathbf{Z}^{(k)}), k \in \mathbb{N}\}$ in increasing order, i.e. such that $0 \leq t_1 \leq t_2 \ldots$

We set $\mathcal{R}^n(t) = (\pi_0)_{|n}$ for $t \in [0, t_1)$. Then define recursively

$$\mathcal{R}^n(t) = C(\mathcal{R}^n(t_i-), \mathbf{Z}^{(i)}), \quad \text{for every } t \in [t_i, t_{i+1}).$$

These processes are consistent in n as we show in the following result.

Proposition 5.1. *For every $t \geq 0$, the sequence of random bivariate partitions $(\mathcal{R}^n(t), n \in \mathbb{N})$ is consistent. If we denote by $\mathcal{R}(t)$ the unique partition of $\mathcal{N}_\infty$ such that $\mathcal{R}_{|n}(t) = \mathcal{R}^n(t)$ for every $n \in \mathbb{N}$, then the process $\mathcal{R} = (\mathcal{R}(t), t \geq 0)$ is a SNEC started from π_0, with rates given as in Theorem 3.4.*

The proof uses similar arguments as in the proof of consistency of exchangeable coalescents given in Proposition 4.5 of [4].

Proof. The key idea (basically (4.4) in [4]) is that by definition, the coagulation operator satisfies

$$\mathrm{Coag}_2(\pi, \tilde{\pi})_{|n} = \mathrm{Coag}_2(\pi_{|n}, \tilde{\pi}) = \mathrm{Coag}_2(\pi_{|n}, \tilde{\pi}_{|n}) \tag{5.1}$$

for any $\pi, \tilde{\pi}$ and n for which this is well defined.

Recall that we defined M_n as the subset of M consisting of points $(t, \mathbf{Z})$ such that $\sum_{i=1}^n X_i \geq 2$ or $\exists i \leq n,\ X_i \sum_{j=1}^n Y_{ij} \geq 2$. Fix $n \geq 2$ and write $(t_1, \mathbf{Z}^{(1)})$ for the first atom of M_n on $(0, \infty) \times \widehat{E}$. Plainly, $\mathcal{R}^{n-1}(t) = \mathcal{R}^n_{|n-1}(t) = (\pi_0)_{|n-1}$ for every $t \in [0, t_1)$.

Consider first the case when $\sum_{i=1}^{n-1} X_i^{(1)} \geq 2$ or $\exists i \leq n-1,\ X_i^{(1)} \sum_{j=1}^{n-1} Y_{ij}^{(1)} \geq 2$. Then $(t_1, \mathbf{Z}^{(1)})$ is also the first atom of M_{n-1} and by definition and using (5.1), $\mathcal{R}^{n-1}(t_1) = \mathcal{R}^n_{|n-1}(t)$.

Now suppose $\sum_{i=1}^{n-1} X_i^{(1)} \leq 1$ and $\forall i \leq n-1,\ X_i^{(1)} \sum_{j=1}^{n-1} Y_{ij}^{(1)} \leq 1$. This implies that at time t_1, there is no species (resp. genes) coalescence between the $n-1$ first species (resp. genes) of $\mathcal{R}^n(t_1-)$. Therefore the coalescence event in $\mathcal{R}^n$ at time t_1 leaves the first $n-1$ blocks of $\mathcal{R}^n(t_1-)^s$ or $\mathcal{R}^n(t_1-)^g$ unchanged, though there may be a coalescence involving the n-th block (in that case, necessarily a singleton $\{n\}$) and one of the $n-1$ first blocks. So finally $\mathcal{R}^n(t_1)_{|n-1} = \mathcal{R}^n(t_1-)_{|n-1} = \mathcal{R}^{n-1}(t_1)$.

In both cases we have $\mathcal{R}^n(t_1)_{|n-1} = \mathcal{R}^{n-1}(t_1)$, and by an obvious induction this is true for any further jump of the process $\mathcal{R}^n$, so that for all $t \geq 0$,

$$\mathcal{R}^n(t)_{|n-1} = \mathcal{R}^{n-1}(t).$$

This shows the existence of $\mathcal{R}$ such that for all n, $\mathcal{R}_{|n} = \mathcal{R}^n$.

From this Poissonian construction $\mathcal{R}^n$ is a Markov process, and by definition the arrays $\mathbf{Z}^{(i)}_{|[n]^2}$ are hierarchically exchangeable, which implies that $\mathcal{R}^n$ is an exchangeable process. Clearly by construction $\mathcal{R}^n(t)$ is nested for all t, and the only jumps of the process $\mathcal{R}^n$ are coalescence events. According to Lemma 3.2, the process $\mathcal{R}$ is a SNEC process. Because the arrays $\mathbf{Z}$, where $(t, \mathbf{Z}) \in M$, are the same arrays that appear in the proof of Theorem 3.4, it is clear that the jump rates of $\mathcal{R}^n$ are those given in Theorem 3.4. □

6 Marginal coalescents – coming down from infinity

Consider a SNEC process $\mathcal{R} = (\mathcal{R}^s, \mathcal{R}^g)$, with rates given as in Theorem 3.4 by two coefficients $a_s, a_g \geq 0$ and two measures, ν_s on $E = (0,1] \times \mathcal{M}_1([0,1])$ and ν_g on $(0,1]$ satisfying (3.6) and (3.7).

It is obvious from Proposition 5.1 that $(\mathcal{R}^s(t),\ t \geq 0)$ is a simple coalescent process, with Kingman coefficient a_s and coagulation measure $\widehat{\nu}_s$ satisfying (2.1) which is the push-forward of $\nu_s(dp, d\mu)$ by the application $(p, \mu) \mapsto p$. Let us call this univariate coalescent the *(marginal) species coalescent* of the SNEC process $\mathcal{R}$.

Now, notice that under an initial condition with a unique species block (i.e., $\mathcal{R}^s$ is constant to the coarsest partition $\mathbf{1}_\infty$), the process $(\mathcal{R}^g(t),\ t \geq 0)$ also behaves as a simple coalescent process, with Kingman coefficient a_g and coagulation measure $\widehat{\nu}_g$ defined by

$$\forall B \text{ Borel set of } (0,1], \quad \widehat{\nu}_g(B) := \nu_g(B) + \int_{(0,1]\times\mathcal{M}_1([0,1])} \nu_s(dp, d\mu)p\,\mu(B).$$

We call the simple coalescent thus defined the *(marginal) gene coalescent* of the SNEC process $\mathcal{R}$.

Equivalently, in terms of Λ-coalescents, the marginal species coalescent is a Λ_s-coalescent with Λ_s defined by

$$\forall B \text{ Borel set of } [0,1], \quad \Lambda_s(B) = a_s\delta_0(B) + \int_{B\times\mathcal{M}_1([0,1])} \nu_s(dp, d\mu)p^2, \tag{6.1}$$

and the marginal gene coalescent is a Λ_g-coalescent with Λ_g defined for all B Borel set of $[0,1]$ by

$$\Lambda_g(B) = a_g\delta_0(B) + \int_B \nu_g(dq)q^2 + \int_{(0,1]\times\mathcal{M}_1([0,1])} \nu_s(dp, d\mu)p \int_B \mu(dq)q^2. \tag{6.2}$$

These two marginal processes allow us to express properties of the initial bivariate SNEC process. Consider an initial state $\varrho_0 \in \mathcal{N}_\infty$ with infinitely many species blocks, each containing infinitely many gene blocks. In a way analogous to the one-dimensional case, recalling that $|\mathcal{R}^g(t)| \geq |\mathcal{R}^s(t)|$ for all $t \geq 0$, we will say that a SNEC *comes down from infinity* (CDI) if for all $t > 0$

$$|\mathcal{R}^g(t)| < \infty \qquad \mathbb{P}_{\varrho_0}\text{-a.s.}$$

In the univariate case, characterizing which coalescent processes come down from infinity has been solved [39] for Λ-coalescents, with the following necessary and sufficient condition for coming down from infinity:

$$\sum_{n\geq 2}\left(\sum_{k=2}^{n}(k-1)\binom{n}{k}\int_{[0,1]}\Lambda(dp)p^{k-2}(1-p)^{n-k}\right)^{-1} < \infty.$$

Note that the previous condition is true as soon as Λ has an atom at 0 ($\Lambda(\{0\})$ is the Kingman coefficient of the process). An equivalent criterion (see [6], and [1] for a probabilistic proof) is the integrability of $1/\psi$ near $+\infty$, where

$$\psi(q) := \int_{[0,1]} \left(e^{-qx} - 1 + qx\right) x^{-2}\,\Lambda(dx). \tag{6.3}$$

We will now see that in the case of simple nested coalescents, we can give a general characterization of the different CDI properties of a SNEC process, depending only on the properties of the marginal species and marginal gene coalescents.

First notice that if the marginal gene coalescent does not CDI, then any species block with infinitely many gene blocks at some time t clearly keeps infinitely many gene blocks for any $t' \geq t$. Also in any case the process $\mathcal{R}^s$ has the distribution of the marginal species coalescent, so determining whether the number of species comes down from infinity is trivial.

Proposition 6.1. *We assume here that $\widehat{\nu}_s(\{1\}) = \widehat{\nu}_g(\{1\}) = 0$ and that the marginal gene coalescent comes down from infinity (CDI). Then we have the following three cases.*

i) If the marginal species coalescent CDI as well, then $\mathcal{R}$ CDI.

ii) If the marginal species coalescent does not CDI but $\int_{[0,1]} \widehat{\nu}_s(dx)\,x = \infty$, then for any initial condition with infinitely many species blocks and for each time $t > 0$, the number of gene blocks in each species block of $\mathcal{R}(t)$ is infinite a.s.

iii) If the marginal species coalescent does not CDI and $\int_{[0,1]} \widehat{\nu}_s(dx)\, x < \infty$*, then for any initial condition and for each time* $t > 0$*, the number of gene blocks in each species block of* $\mathcal{R}(t)$ *is finite a.s.*

As a consequence of this proposition, it is clear that $\mathcal{R}$ comes down from infinity if and only if both the marginal species coalescent and the marginal gene coalescent come down from infinity.

A simple example of a SNEC process coming down from infinity is the nested Kingman coalescent ('Kingman in Kingman'), given by its marginal rates $a_s, a_g > 0$, defined so that each pair of species coalesces at rate a_s independently of the others, and each pair of genes within the same species coalesces at rate a_g independently of the rest. Since the marginal coalescents are precisely two Kingman coalescents, they both come down from infinity.

Note that the Bolthausen-Sznitman coalescent [9] (denoted U-coalescent in [35] because the measure Λ is uniform on $[0,1]$) satisfies the conditions of the peculiar case *ii)*. So for a SNEC $\mathcal{R}$ defined by a Kingman gene coalescent evolving within a species U-coalescent, at each positive time the number of gene blocks within a species block is infinite (if the initial state ϱ_0 has an infinite number of species blocks).

Case *iii)* can easily be obtained by considering a "slow" species coalescent, such as a δ_x-coalescent for $x \in (0,1)$, or any $\beta(a,b)$-coalescent with $a > 1, b > 0$ (that is a Λ-coalescent with $\Lambda(dx) = c_{a,b} x^{a-1}(1-x)^{b-1} dx$).

Proof. ***i)*** Suppose both marginal coalescents come down from infinity, and consider an initial state $\varrho \in \mathcal{N}_\infty$ with infinitely many species blocks, each containing infinitely many gene blocks.

Choose $t > 0$. Since $\mathcal{R}^s$ comes down from infinity, we have $\mathbb{P}_\varrho(|\mathcal{R}^s(t/2)| < \infty) = 1$, and necessarily, $\mathcal{R}^s$ stays constant on an interval $[t/2, T[$, where T is its next jump time. Now on the interval $[t/2, \min(T,t)[$, within each of the $|\mathcal{R}^s(t/2)|$ species block, the gene blocks undergo independent coalescent processes which CDI, therefore there are finitely many gene blocks in each species at time $\min(T,t)$, which implies

$$\mathbb{P}_\varrho(|\mathcal{R}^g(t)| < \infty) = 1.$$

Let us say a few words before proving *ii)* and *iii)*. Pick $t > 0$ and focus on the species containing 1 (the first species). To this aim, write $M(t)$ for the number of genes within the first species, at time t. By exchangeability, to show *ii)* it is sufficient to show $\mathbb{P}_\varrho(M(t) = \infty) = 1$, for any initial condition ϱ with infinitely many species blocks, and to show *iii)* it is sufficient to show $\mathbb{P}_\varrho(M(t) < \infty) = 1$, for any initial condition ϱ.

ii) Suppose $\int_{[0,1]} \widehat{\nu}_s(dx)\, x = \infty$. First, note that since the species coalescent does not CDI and $\widehat{\nu}_s(\{1\}) = 0$, there are at all times $t \geq 0$ infinitely many species blocks (see for instance [35, Proposition 23]). Now let us fix $\delta \in (0,t]$ and $\varepsilon \in (0,1]$, and investigate the random number of coalescence events in the time interval $[t-\delta, t]$ involving the first species and at least a proportion ε of all other species. More precisely, we consider the number of atoms $(s, \mathbf{Z})$ in the Poissonian construction such that $s \in [t-\delta, t]$, $X_1 = 1$ and $\lim_{n\to\infty} \sum_{i=1}^n X_i/n \geq \varepsilon$. From the Poissonian construction, it is easy to see that this number is a Poisson random variable with mean

$$\delta \int_{[\varepsilon,1]} \widehat{\nu}_s(dx)\, x.$$

Pick any $A \in \mathbb{N}$ and $\eta > 0$. We will show $\mathbb{P}_\varrho(M(t) \leq A) < 2\eta$, which is sufficient to conclude that $M(t) = \infty$ a.s. Note that we assumed that the marginal gene coalescent

CDI, so for $\Pi = (\Pi(t),\, t \geq 0)$ a version of this univariate coalescent started from $\mathbf{0}_\infty$, we have $\mathbb{P}(|\Pi(\delta)| < \infty) = 1$ for all $\delta > 0$. In addition, Π is right-continuous, so $|\Pi(\delta)| \uparrow \infty$ as $\delta \to 0$. Therefore, one can choose $\delta > 0$ small enough, and then $\varepsilon > 0$ such that

$$\mathbb{P}(|\Pi(\delta)| \leq A) < \eta \quad \text{and} \quad e(\varepsilon) := \int_{[\varepsilon,1]} \widehat{\nu}_s(dx)\, x \geq \frac{-\log(\eta)}{\delta}. \tag{6.4}$$

Now consider the stopping time defined by

$$T := \inf\{s \geq t - \delta, \text{ the first species participates at time } s \text{ in a coalescence event involving at least a proportion } \varepsilon \text{ of other species}\}.$$

By the Poisson construction, $T - (t - \delta)$ is an exponential random variable with parameter $e(\varepsilon)$, so from (6.4) we deduce

$$\mathbb{P}_\varrho(T \geq t) \leq \eta.$$

Now since T is a coalescence time for the first species, we have $M(T) = \infty$ almost surely. Indeed, the assumption $\widehat{\nu}^g(\{1\}) = 0$ implies that not every gene participates in the coalescence. But since an infinite number of species participate in the coalescence, the law of large numbers implies that in the newly formed species, there is an infinite number of genes which do not coalesce at time T. Since $M(T) = \infty$, we can define a random injection $\sigma : \mathbb{N} \to \mathbb{N}$ mapping k to the first element of the k-th gene of the first species at time T. We then define $\widetilde{\Pi}(u) := \sigma(\mathcal{R}^g(T + u))$, which has by the strong Markov property the distribution of a marginal gene coalescent started from $\mathbf{0}_\infty$, independent of T. Furthermore, by construction we have $M(T + u) \geq |\widetilde{\Pi}(u)|$ a.s., so that finally

$$\begin{aligned}
\mathbb{P}_\varrho(M(t) \leq A) &\leq \mathbb{P}_\varrho(T > t) + \mathbb{P}_\varrho(t - \delta \leq T \leq t)\, \mathbb{P}_\varrho(|\widetilde{\Pi}(t - T)| \leq A \mid t - \delta \leq T \leq t) \\
&\leq \mathbb{P}_\varrho(T > t) + \mathbb{P}(|\Pi(\delta)| \leq A) \\
&\leq 2\eta.
\end{aligned}$$

iii) Now supposing $\int_{[0,1]} \widehat{\nu}_s(dx)\, x < \infty$, with the same argument as previously, the first species participates in coalescence events at some random times $0 < T_1 < T_2 < \ldots$, distributed as a Poisson process with parameter $\int_{[0,1]} \widehat{\nu}_s(dx)\, x$, and all these events involve infinitely many species blocks (recall the marginal species coalescent does not CDI and so in particular has $a_s = 0$). Let $T_0 := 0$ by convention and for each i, we can define a random injection $\sigma_i : \mathbb{N} \to \mathbb{N}$ mapping k to the first element of the k-th gene of the first species at time T_i. Now because the first species does not change during the intervals $[T_i, T_{i+1})$, the process $\widetilde{\Pi}_i$ defined by

$$\widetilde{\Pi}_i(u) := \sigma_i(\mathcal{R}^g(T_i + u))$$

is a marginal gene coalescent (and so CDI by assumption), which is independent of T_i, and there is the following equality between processes, for $u < T_{i+1} - T_i$,

$$M(T_i + u) = \widetilde{\Pi}_i(u).$$

Finally, we have for any $t > 0$, and any initial $\varrho \in \mathcal{N}_\infty$,

$$\begin{aligned}
\mathbb{P}_\varrho(M(t) < \infty) &= \sum_{i \geq 0} \mathbb{P}_\varrho(T_i < t < T_{i+1})\, \mathbb{P}_\varrho(M(t) < \infty \mid T_i < t < T_{i+1}) \\
&= \sum_{i \geq 0} \mathbb{P}_\varrho(T_i < t < T_{i+1})\, \mathbb{P}_\varrho(\widetilde{\Pi}_i(t - T_i) < \infty \mid T_i < t < T_{i+1}) \\
&= \sum_{i \geq 0} \mathbb{P}_\varrho(T_i < t < T_{i+1}) = 1,
\end{aligned}$$

which concludes the proof. □

References

[1] Julien Berestycki, Nathanaël Berestycki, and Vlada Limic, *A small-time coupling between Λ-coalescents and branching processes*, The Annals of Applied Probability **24** (2014), no. 2, 449–475 (en). MR-3178488

[2] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg, *The genealogy of branching Brownian motion with absorption*, The Annals of Probability **41** (2013), no. 2, 527–618 (en). MR-3077519

[3] Nathanaël Berestycki, *Recent progress in coalescent theory*, Ensaios Matemáticos **16** (2009), no. 1, 1–193. MR-2574323

[4] Jean Bertoin, *Random fragmentation and coagulation processes*, Cambridge University Press, Cambridge, 2006. MR-2253162

[5] Jean Bertoin and Jean-François Le Gall, *Stochastic flows associated to coalescent processes*, Probability Theory and Related Fields **126** (2003), no. 2, 261–288. MR-1990057

[6] Jean Bertoin and Jean-François Le Gall, *Stochastic flows associated to coalescent processes. III. Limit theorems*, Illinois Journal of Mathematics **50** (2006), no. 1-4, 147–181. MR-2247827

[7] Airam Blancas, Tim Rogers, Jason Schweinsberg, and Arno Siri-Jégousse, *The nested Kingman coalescent: Speed of coming down from infinity*, arXiv:1803.08973 [math] (2018).

[8] Airam Blancas and Anton Wakolbinger, *A representation for the semigroup of a two-level Fleming-Viot process in terms of the Kingman nested coalescent*, In preparation.

[9] Erwin Bolthausen and Alain-Sol Sznitman, *On Ruelle's probability cascades and an abstract cavity method*, Communications in Mathematical Physics **197** (1998), no. 2, 247–276. MR-1652734

[10] Éric Brunet and Bernard Derrida, *Genealogies in simple models of evolution*, Journal of Statistical Mechanics: Theory and Experiment **2013** (2013), no. 01, P01006. MR-3036206

[11] Donald A. Dawson, *Multilevel mutation-selection systems and set-valued duals*, Journal of Mathematical Biology **76** (2018), no. 1-2, 295–378 (en). MR-3742789

[12] James H Degnan and Noah A Rosenberg, *Gene tree discordance, phylogenetic inference and the multispecies coalescent*, Trends in ecology & evolution **24** (2009), no. 6, 332–340.

[13] Michael M Desai, Aleksandra M Walczak, and Daniel S Fisher, *Genetic diversity and the structure of genealogies in rapidly adapting populations*, Genetics **193** (2013), no. 2, 565–585.

[14] Jeff J Doyle, *Trees within trees: genes and species, molecules and morphology*, Systematic Biology **46** (1997), no. 3, 537–553.

[15] Jean-Jil Duchamps, *Trees within trees II: Nested Fragmentations*, arXiv:1807.05951 (2018).

[16] Rick Durrett and Jason Schweinsberg, *A coalescent model for the effect of advantageous mutations on the genealogy of a population*, Stochastic processes and their applications **115** (2005), no. 10, 1628–1657. MR-2165337

[17] Bjarki Eldon and John Wakeley, *Coalescent processes when the distribution of offspring number among individuals is highly skewed*, Genetics **172** (2006), no. 4, 2621–2633.

[18] Alison Etheridge, *Some mathematical models from population genetics: École d'été de probabilités de Saint-Flour XXXIX-2009*, Lecture notes in mathematics, Springer, 2011. MR-2759587

[19] Joseph Felsenstein, *Inferring phylogenies*, vol. 2, Sinauer associates Sunderland, MA, 2004.

[20] Félix Foutel-Rodier, Amaury Lambert, and Emmanuel Schertzer, *Exchangeable coalescents, ultrametric spaces, nested interval-partitions: a unifying approach*, arXiv:1807.05165 (2018).

[21] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes, *Unifying the epidemiological and evolutionary dynamics of pathogens*, Science **303** (2004), no. 5656, 327–332.

[22] Joseph Heled and Alexei J Drummond, *Bayesian inference of species trees from multilocus data*, Molecular biology and evolution **27** (2009), no. 3, 570–580.

[23] Olav Kallenberg, *Probabilistic symmetries and invariance principles*, Probability and Its Applications, Springer-Verlag, New York, 2005 (en). MR-2161313

[24] J.F.C. Kingman, *The coalescent*, Stochastic processes and their applications **13** (1982), no. 3, 235–248. MR-0671034

[25] Amaury Lambert, *Population dynamics and random genealogies*, Stochastic Models **24** (2008), no. sup1, 45–163. MR-2466449

[26] Amaury Lambert, *Probabilistic models for the (sub)tree(s) of life*, Braz. J. Probab. Stat. **31** (2017), no. 3, 415–475. MR-3693976

[27] Amaury Lambert, *Random ultrametric trees and applications*, ESAIM: Procs **60** (2017), 70–89. MR-3772473

[28] Amaury Lambert and Emmanuel Schertzer, *Coagulation-transport equations and the nested coalescents*, arXiv:1807.09153 (2018).

[29] Wayne P Maddison, *Gene trees in species trees*, Systematic biology **46** (1997), no. 3, 523–536.

[30] Sebastian Matuszewski, Marcel E Hildebrandt, Guillaume Achaz, and Jeffrey D Jensen, *Coalescent processes with skewed offspring distributions and non-equilibrium demography*, Genetics (2017), genetics–300499.

[31] Richard A. Neher and Oskar Hallatschek, *Genealogies of rapidly adapting populations*, Proceedings of the National Academy of Sciences **110** (2013), no. 2, 437–442.

[32] Masatoshi Nei and Sudhir Kumar, *Molecular evolution and phylogenetics*, Oxford university press, 2000.

[33] Roderic DM Page and Michael A Charleston, *From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem*, Molecular phylogenetics and evolution **7** (1997), no. 2, 231–240.

[34] Roderic DM Page and Michael A Charleston, *Trees within trees: phylogeny and historical associations*, Trends in Ecology & Evolution **13** (1998), no. 9, 356–359.

[35] Jim Pitman, *Coalescents with multiple collisions*, The Annals of Probability **27** (1999), no. 4, 1870–1902 (en). MR-1742892

[36] Noah A Rosenberg, *The probability of topological concordance of gene trees and species trees*, Theoretical population biology **61** (2002), no. 2, 225–247.

[37] Serik Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, Journal of Applied Probability **36** (1999), no. 4, 1116–1125. MR-1742154

[38] Jason Schweinsberg, *Coalescents with simultaneous multiple collisions*, Electronic Journal of Probability **5** (2000) (EN). MR-1781024

[39] Jason Schweinsberg, *A necessary and sufficient condition for the Λ-coalescent to come down from infinity.*, Electronic Communications in Probability **5** (2000), 1–11 (EN). MR-1736720

[40] Jason Schweinsberg, *Coalescent processes obtained from supercritical Galton-Watson processes*, Stochastic Processes and their Applications **106** (2003), no. 1, 107–139. MR-1983046

[41] Jason Schweinsberg, *Rigorous results for a population model with selection II: genealogy of the population*, Electronic Journal of Probability **22** (2017). MR-3646064

[42] Charles Semple and Mike A Steel, *Phylogenetics*, vol. 24, Oxford University Press, 2003. MR-2060009

[43] Gergely J Szöllősi, Eric Tannier, Vincent Daubin, and Bastien Boussau, *The inference of gene trees with species trees*, Systematic biology **64** (2014), no. 1, e42–e62.

[44] Aurelien Tellier and Christophe Lemaire, *Coalescence 2.0: A multiple branching of recent theoretical developments and their applications*, Molecular ecology **23** (2014), no. 11, 2637–2652.

[45] Erik M Volz, Katia Koelle, and Trevor Bedford, *Viral phylodynamics*, PLoS Computational Biology **9** (2013), no. 3, e1002947. MR-3048921

Acknowledgments. The four authors thank the *Center for Interdisciplinary Research in Biology* (Collège de France) for travel funding. ABB and ASJ would like to thank CIMAT, A.C. and especially Víctor Rivero for support and for his comments on this project, which started during Airam's PhD thesis. ABB is supported by CONACYT postdoctoral grant 234823, and ASJ by CONACYT grant CB-2014/243068.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System `http://www.vtex.lt/en/ejms.html`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`