CIMAT

Centro de Investigación en Matemáticas A.C.

Two contributions to the theory of stochastic population dynamics

T H E S I S

In partial fulfillment of the requirements for the degree of

Doctor of Sciences

with orientation in

Probability and Statistic

by

Airam Aseret Blancas Benítez

Advisors:
Dr. Víctor Rivero, Dr. Arno Siri-Jégousse

GUANAJUATO, GUANAJUATO, MEXICO FEBRUARY 2016

A la mujer que le debo todo lo que soy,
mi maravillosa madre.

This work is part of the project CONACyT CB-2014/243068 “Modelos aleatorios en Evolución, Genética y Ecología” directed by Dr. Arno Siri-Jégousse.

Agradecimientos

Primero quiero expresar un profundo agradecimiento a mis directores de tesis los doctores Víctor Rivero y Arno Siri-Jégousse, porque su dedicación fue fundamental para la elaboración de este trabajo.

A Víctor le agradezco por haberle dado dirección a mi carrera profesional cuando apenas iniciaba. Así mismo le doy las gracias por haberme permitido conocer a Maika y ver crecer a Felipe y a Luna.

A Arno le doy gracias por todo su apoyo, por brindarme una mano solidaria cuando la necesitaba y por su dinámismo.

Un agradecimiento especial es para el Professor Amaury Lambert, quien hizo posible que realizara una estancia de investigación en la Universidad Paris 6. Por permitirme ser parte del equipo SMILE y por la confianza que me tuvo para proponerme el desarrollo de un proyecto tan apasionante como es el coalescente anidado.

Al comité evaluador integrado por los Profesores Amaury Lambert y Andreas Kyprianou. Así como por los doctores José Alfredo López Mimbela, Juan Carlos Pardo, Arno Siri-Jégousse, Víctor Rivero y Gerónimo Uribe, les agradezco el tiempo que invirtieron en la revisión de esta tesis. En particular quiero agradecer a Juan Carlos por seguir mi carrera e impulsarme cuando fue necesario.

A la población mexicana que hace posible la existencia de becas de posgrado. Para la obtención del grado de doctor me fue otorgada la beca No. 301357 vía el Consejo Nacional de Ciencia y Tecnología.

Así mismo quiero agradecer al Laboratorio Internacional Solomon Lefschetz CONACyT-CNRS por el apoyo financiero recibido para la realización de este trabajo.

Gracias al CIMAT porque es la institución donde académicamente me he formado. A su afectuosa comunidad le agradezco por contagiarme día a día con su sonrisa. En especial quisiera decir gracias a Rosy Dávalos, Larry, Odal, Ciri, Tere, Memo, Josesito, Lolita y Lalo porque de ellos he aprendido a valorar mi trabajo.

A Stephanie quiero expresarle mi más profunda gratitud porque sus conversaciones me hicieron ver la vida de un modo distinto, haciéndome una persona más consciente.

Un cariñoso agradecimiento es para mis amigos Henry, Miraine, Elena y Sandra.

A Henry, mi hermano (académico) mayor, le agradezco sus palabras porque ellas me ayudan a encontrarme a mi misma.

A Miraine, mi cómplice y casi hermana, le agradezco sus enseñanzas, su energía, su incondicional amistad, todos los inolvidables momentos que vivimos durante mi estancia en Paris. También quiero agradecer a su esposo Pedro por su amistad y a la pequeña Irene que me impulsó para entregar esta tesis.

A Elena porque sus líneas me reconfortaron cuando lo necesité, en la distancia ha sido un apoyo muy fuerte para mi.

A Sandra le agradezco su compresión y su generosidad.

A Cécile, Fanny y Miraine, quienes fueron mi compañeras de oficina en el equipo SMILE, les agradezco sus expresiones de afecto.

A Cécile le doy las gracias porque con su ejemplo de gentileza, dedicación y entrega, siempre me motivó.

A Fanny le agradezco porque con su entusiasmo me hizo dejar de lado mi timidez y sacar de mi el lado artísitico.

A mi familia porque es mi mayor bendición.

A mi mamá por todo su amor, por estar a mi lado en cada instante a pesar de la distancia, por darme la fuerza para seguir en los momentos díficiles, por hacerme conocer a *DIOS*.

Papi gracias por estar siempre ahí mostrándome tu serenidad y fortaleza.

A mis hermanos Airamara, Analhí y Adamir porque sus acciones siempre me recuerdan las enseñanzas de nuestros padres.

A Sergio, mi querido tío, porque desde pequeña me ha apoyado y me ha dado ánimos.

A mis abuelos Conchita y Chuyito porque aunque ya no están con nosotros sigo recibiendo de sus bendiciones.

Finalmente quiero agradecer a Ernesto por darme la fortaleza para enfrentar los momentos difíciles y alcanzar mis metas. Él es para mi un ejemplo de perseverancia. Más aún, le doy gracias por sus innumerables muestras de amor.

Airam Aseret Blancas

Acknowledgments

Firstly, I want to express a deep gratitude to my thesis supervisors Dr. Víctor Rivero and Dr. Arno Siri-Jégousse. Their dedication has been essential for the realization of this work, particularly

> To Víctor, who gave direction to my career since the begging. I also appreciate that he let me meet Maika and I had the opportunity to see growing his children Felipe and Luna.
>
> To Arno by his support, solidarity and dynamism.

Special thanks to Professor Amaury Lambert, who provided me an opportunity to join their team SMILE as intern, and who gave access to the laboratory and research facilities. For suggesting an exciting project as the nested coalescent, addressed in this thesis.

I am also grateful to the committee for this thesis the professors Amaury Lambert and Andreas Kyprianou, as well as the doctors José Alfredo López Mimbela, Juan Carlos Pardo, Arno Siri-Jégousse, Víctor Rivero and Gerónimo Uribe, who carefully revised this work. Most specially, thanks to Juan Carlos for following and propel my career when it was necessary.

A special recognition to the Mexican people who made possible the existence of post-graduate fellowships. Through the Mexican Council of Science, CONACyT, I received the grant number 301357 to do my PhD.

I also would like to thank the Internacional Laboratory Solomon Lefschetz CONACyT-CNRS for the financial support.

I thank CIMAT for supporting my academic training. To its community because its joy and enthusiasm. Specially I would like to say thanks to Rosy Dávalos, Larry, Odal, Ciri, Tere, Memo, Josesito, Lolita and Lalo because from them I learnt to appreciate my work.

I would like to express my deep gratitude to Stephanie because in her English classes I also learned life lessons.

My sincere appreciation to my friends:

Henry for helping me to find myself.

Miraine for her teachings, energy, unconditional friendship and all the unforgettable moments that we lived during my stay in Paris. I extend my gratitude to Pedro, her husband, for his friendship and the little Irene who prompted me to deliver this thesis.

Elena for her words, these always comforted me.

Sandra for her understanding and generosity.

I am also grateful to those who were my colleagues.

Cécile for her kindness, dedication and commitment.

Fanny for her joy and enthusiasm, for making me put aside my shyness.

I want to express my admiration to my family.

To my mom for her love, for being by my side despite the distance, for giving me strength and for teaching me that *GOD* exists.

To my father for being there as a model of serenity and strength.

To my sisters Airamara and Analhí and my brother Adamir, because they always remember me the lessons given by our parents.

To my uncle Sergio because since I was a child he supports and encourages me.

To my grandparents "Conchita" and "Chuyito" because I still feeling their blessings, though they are now in a better place.

Last but not the least, I want to thank Ernesto for giving me the strength to reach my goals and teaching me to face difficulties. He is my symbol of perseverance. Specially, I thank him for his countless expressions of love.

Airam Aseret Blancas

Contents

Introducción

En esta tesis se estudia la teoría de procesos estocásticos utilizada para modelar fenómenos de interés en el campo de la biología poblacional y molecular. Específicamente, estudiamos la teoría de procesos de ramificación y procesos coalescentes. Hemos dividido éste trabajo en tres capítulos, en el primero presentamos un modelo que describe la estructura de la genealogía de una población con mutaciones neutrales, propuesto por Bertoin (2010). Suponiendo hipótesis mas débiles en la ley de reproducción de los individuos, se establecen los resultados asintóticos obtenidos en Bertoin (2010) sobre la estructura de las subfamilias alélicas. En los restantes capítulos, se estudian desde dos enfoques los árboles de genes y árboles de especies que surgen en biología molecular. Más precisamente, en el Capítulo 2 nos interesamos por la probabilidad de que las topologías asociadas a dichos árboles coincidan, mientras que en el Capítulo 3 definimos una nueva clase de procesos coalescentes con la finalidad de modelar su dinámica.

Capítulo 1. Un proceso de ramificación con mutationes neutrales raras

Sea una población modelada por un proceso de Galton Watson, recordemos que esto significa que los individuos ahí presentes son asexuales y se reproducen en generaciones discretas, dando nacimiento a un número aleatorio de hijos independientemente de los otros y con la misma distribución. Suponiendo que aparecen mutaciones que modifican el tipo genético de los individuos pero no la ley con que se reproducen, y que cada evento de mutación genera un nuevo alelo, se tiene un proceso de Galton Watson con mutaciones neutrales $\{Z_n^{(+)} : n \in \mathbb{Z}_+\}$. El número de hijos del mismo tipo genético que el padre (clones) y los de tipo distinto (mutantes), están determinados respectivamente por las variables $\xi^{(c)}$ y $\xi^{(m)}$. De esta manera el tamaño de una familia típica es $\xi^{(+)} := \xi^{(c)} + \xi^{(m)}$. La genealogía de este proceso se representa mediante árboles aleatorios enraízados, con marcas entre las aristas que unen a padres e hijos mutantes. De acuerdo al trabajo de Bertoin (2010), llamamos *individuos del n-ésimo tipo*, a todos aquellos asociados a vértices con n marcas en su línea ancestral, denotando el total de la población de individuos del tipo n por T_n, ver Figura 1 (izq). Vértices que corresponden a individuos cuyo padre es un individuo del tipo $n-1$, son conocidos como *mutantes del tipo n*. Escribimos M_n para el número total

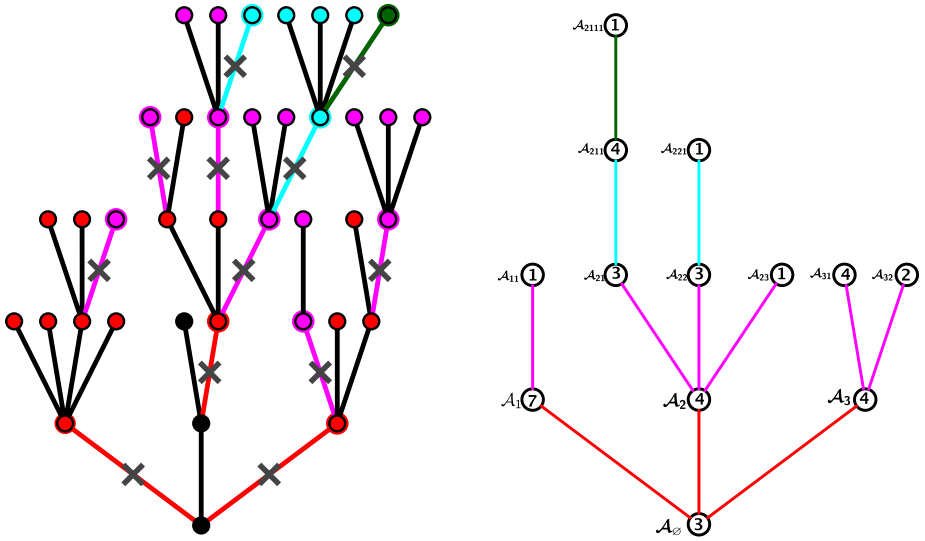

Figura 1: Árbol de la genealogía de una población con mutaciones (izq) y árbol de alelos (der). El color del vértices representa el tipo de alelo, así por ejemplo vértices en rojo corresponden a individuos del 1-tipo y los coloreados con magenta a los del 3-tipo. Las aristas en color están asociadas a las líneas de paro generadas por los mutantes. Las etiquetas en el árbol de alelos corresponden a los tamaños de las sub-familias alélicas.

de estos individuos, con la convención de que los mutantes del tipo cero son los ancestros. El árbol de Galton-Watson con mutaciones neutrales tiene una propiedad de ramificación llamada general, la cual asegura que los subárboles con raíz, mutantes del n-ésimo tipo, son copias independientes del árbol original. En consecuencia $\{M_k : k \in \mathbb{Z}_+\}$, es un proceso de Galton-Watson y $\{(T_k, M_{k+1}) : k \in \mathbb{Z}_+\}$ es una cadena de Markov. La observación de estos hechos conduce a la construcción de un nuevo árbol, el cual se conoce como *árbol de alelos*, ver Figura 1 (der). Heurísticamente podemos decir que se obtiene colapsando en un solo vértice los subárboles cuya raíz es un mutante del k-ésimo tipo, y asignando al vértice una etiqueta con el número de vértices de dicho subárbol. Por convención se ordenan las etiquetas de cada uno de los niveles aleatoria y uniformemente. Observe que la suma de las etiquetas de los vértices del k-ésimo nivel es T_k. Además, el número de vertices en el k-ésimo nivel es M_k, de manera que la estructura del árbol de alelos como tal, es decir, considerando únicamente los vértices del árbol de alelos y no sus etiquetas, este describe la genealogía de un proceso de Galton-Watson usual debido a que $\{M_k : k \in \mathbb{Z}_+\}$ lo es. Motivados por esta observación, en nuestro trabajo se establecen las versiones de algunos resultados clásicos de la teoría de procesos de ramificación para un proceso de Galton-Watson con mutaciones neutrales.

Considerando que el tiempo de extinción de mutantes

$$T = \inf\{n \geq 1 : M_n = 0\},$$

es finito casi seguramente, construimos la versión condicionada a la no extinción de mu-

tantes de la cadena $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$. En este sentido primero encontramos una medida de probabilidad mediante el método de la h-transformada y luego verificamos que bajo esta medida el proceso tiene las transiciones del proceso de interés, lo cual se traduce respectivamente en el siguiente par de resultados.

Teorema 1. Sea $a \in \mathbb{Z}_+$ y $\mathcal{F}_n$ la filtración natural del proceso $\{(T_{n-1}, M_n) : n \in \mathbb{N}\}$. Entonces, existe una medida de probabilidad $\mathbb{P}_a^{\uparrow}$ que es localmente absolutamente continua con respecto a $\mathbb{P}_a$, con martingala de Radon-Nikodim

$$Y_n = \frac{M_n q^{M_n - a}}{(f'(q))^n} \mathbf{1}_{\{n<T\}},$$

donde $f(y) = \mathbb{E}_1(y^{M_1})$ y $q = \mathbb{P}_1(0 < T < \infty)$, es decir,

$$d\mathbb{P}_a^{\uparrow}|_{\mathcal{F}_n} = \frac{Y_n}{a}\, d\mathbb{P}_a|_{\mathcal{F}_n}, \quad n \in \mathbb{N}.$$

Más aún, $\mathbb{P}_a^{\uparrow}$ es la ley de una cadena de Markov $\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\}$ con probabilidad de transición a n pasos,

$$Q^n_{(i,j),(k,l)} = \frac{l q^{l-j}}{j(f'(q))^n} P^n_{(i,j),(k,l)}, \quad j, l \geq 1,$$

donde $\{P^n_{(i,j),(k,l)} : i, j, k, l \in \mathbb{Z}_+\}$ denota la probabilidad de transición en n pasos de $\{(T_n, M_{n+1}), n \in \mathbb{Z}_+\}$.

Notemos que bajo $\mathbb{P}_a^{\uparrow}$ el tiempo de extinción de mutantes es infinito casi seguramente debido a que

$$\mathbb{P}_a^{\uparrow}(n < T) = \mathbb{P}_a\left(\frac{Y_n}{a}\mathbf{1}_{\{n<T\}}\right) = 1.$$

Teorema 2. Supongamos que $\mathbb{E}(\xi^{(c)}) < 1$ y $\mathbb{E}(\xi^{(+)}) \leq 1$.

i) Sea $a, n \in \mathbb{N}$ con n fija. La ley condicional del proceso $\{(T_k, M_{k+1}) : 0 \leq k \leq n-1\}$ bajo $\mathbb{P}_a(\cdot | n + k < T < \infty)$ converge, cuando $k \to \infty$, hacia la medida de probabilidad $\mathbb{P}_a^{\uparrow}$, en el sentido de que para cualquier n

$$\lim_{k\to\infty} \mathbb{P}_a(A | n + k < T < \infty) = \mathbb{P}_a^{\uparrow}(A), \quad \forall A \in \mathcal{F}_n,$$

donde $T = \text{ínf}\{n \geq 1 : M_n = 0\} < \infty, \quad \mathbb{P}_a\text{-}c.s.$

ii) El límite de Yaglom

$$\lim_{n\to\infty} \mathbb{P}(T_{n-1} = i, M_n = j | n < T < \infty), \quad i, j \in \mathbb{Z}_+,$$

existe y tiene función generadora $\widehat{\varphi}(x, y)$ tal que para todo $n \in \mathbb{N}$,

$$m^n \widehat{\varphi}(x, y) = \widehat{f}(\varphi_n(x, y)) - \widehat{f}(\varphi_n(x, 0)), \quad x, y \in [0, 1].$$

Por otro lado, es bien sabido de la teoría clásica de procesos de ramificación que una sucesión de procesos de Galton Watson, apropiadamente reescalada en espacio de estados, converge a los llamados procesos de Jiřina (1958). Buscando obtener un resultado similar para un proceso de Galton-Watson con mutaciones neutrales, en Bertoin (2010) se define un *CSBP indexado por un árbol* con ley de reproducción ν, como un proceso a valores en $(0,\infty)$, indexado en el conjunto de las sucesiones de los naturales

$$\mathbb{U} = \bigcup_{k\in\mathbb{Z}_+} \mathbb{N}^k,$$

donde $\mathbb{N} = \{1,2,...\}$ y $\mathbb{N}^0 = \{\varnothing\}$, tal que condicionalmente a los vértices de niveles precedentes a un nivel dado, los vértices en ese nivel se distribuyen como la familia de los átomos de una medida aleatoria de Poisson. Recordando la construcción a partir de subordinadores de procesos de Jirina, vemos que los vértices en el nivel k representan los tamaños de las subfamilias de la generación k de un proceso de Jirina con ley de reproducción ν, los cuales descienden de un padre en la generación $k-1$.

Sea $\mathbb{P}_a^p$, la ley de probabilidad de un proceso de Galton Watson con mutaciones neutrales que inicia con a ancestros y con probabilidad de que un individuo engendre un mutante igual a p. Interesados en analizar el comportamiento asintótico de una sucesión de árboles de alelos desde un contexto mas amplio al planteado en Bertoin (2010), donde la ley de reproducción tiene varianza finita, consideramos un proceso de Galton Watson con mutaciones neutrales con ley de reproducción $\pi_k^+ = \mathbb{P}(\xi^{(+)} = k)$ crítica y que presenta el comportamiento

$$\bar{\pi}^+(j) := \mathbb{P}(\xi^{(+)} > j) \in VR_\infty^{-\alpha}, \quad \alpha \in [1,2], \tag{1}$$

donde $VR_\infty^{-\alpha}$ denota la clase de funciones que varían regularmente con índice $-\alpha$ en ∞. Esto implica que existe una función r que varía regularmente tal que

$$r(n)\mathbb{P}(\xi^+ > ny) \xrightarrow[n\to\infty]{} c_\alpha y^{-\alpha}, \quad \forall y > 0, \tag{2}$$

donde $c_\alpha = 1/\Gamma(3-\alpha)$. La notación $f \sim g$ significa que $\lim_{x\to\infty} f(x)/g(x) = 1$. A diferencia de Bertoin (2010), donde el número de ancestros y la tasa de mutación se comporta asintóticamente como sigue,

$$a(n) \sim nx \quad \text{y} \quad p(n) \sim cn^{-1}, \quad n \to \infty, \tag{3}$$

donde c y x constantes positivas, en nuestro trabajo utilizamos las siguientes hipótesis

$$a(n) \sim xr(n)p(n) \quad \text{y} \quad p(n) \sim cn^{-1}, \quad n \to \infty. \tag{4}$$

En adelante usaremos la notación $\Longrightarrow$, para referirnos a la convergencia en distribución cuando $n \to \infty$ y $\mathcal{L}\left(\cdot, \mathbb{P}_{a(n)}^{p(n)}\right)$ para la distribución de una variable aleatoria bajo la medida $\mathbb{P}_{a(n)}^{p(n)}$. Uno de nuestros principales resultados determina las constantes de normalización para las cuales se tiene la convergencia de una sucesión de árboles de alelos, asociada a procesos de Galton Watson con mutaciones neutrales definido en el marco anterior.

Teorema 3. Si las hipótesis (1) y (4) se cumplen, entonces se tiene la siguiente convergencia, en el sentido de las distribuciones finito dimensionales

$$\mathcal{L}\left(\left(((r(n))^{-1}\mathcal{A}_u^{(n)}, (r(n)p(n))^{-1}d_u^{(n)}) : u \in \mathbb{U}\right), \mathbb{P}_{a(n)}^{p(n)}\right) \Longrightarrow ((\mathcal{Z}_u^{1/\alpha}, \mathcal{Z}_u^{1/\alpha}) : u \in \mathbb{U}),$$

donde $\{\mathcal{Z}_u : u \in \mathbb{U}\}$ es un proceso de ramificación continuo indexado por un árbol CSBP, con medida de reproducción

$$\nu^\alpha(dy) = c' x^{-1-1/\alpha} dy, \quad y > 0,\ \alpha \in (1,2), \tag{5}$$

donde $c' = \alpha^{-1}/\Gamma(1-\alpha^{-1})$.

A diferencia de la convergencia clásica de Galton Watson, con este tipo de resultados no solo tenemos convergencia de los tamaños de las generaciones de la población dada, sino que además obtenemos la convergencia de sus genealogías.

También hemos probado la convergencia de las distribuciones finito dimensionales de la cadena $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$ condicionada a la no extinción de mutantes, hacia un proceso de Jirina con inmigración en tiempo discreto.

Teorema 4. Si la ley de reproducción es crítica, entonces existen sucesiones $b_1(n)$ y $b_2(n)$ tales que la siguiente convergencia se cumple en el sentido de las distribuciones finito dimensionales:

$$\mathcal{L}\left((b_1(n)T_{k-1}, b_2(n)M_k) : k \in \mathbb{Z}_+), \mathbb{P}_{a(n)}^{p(n)\uparrow}\right) \Longrightarrow ((Y_k, \beta Y_k) : k \in \mathbb{Z}_+),$$

donde $\{Y_k : k \in \mathbb{Z}_+\}$ es un CSBP con inmigación, caracterizado por las siguientes condiciones:

i) si la ley de reproducción tiene varianza finita y (3) se cumple, entonces su medida reproducción esta dada por

$$\nu(dy) = \frac{c}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{c^2 y}{2\sigma^2}\right) dy, \quad y > 0, \tag{6}$$

y la medida de inmigración es $z\nu(dz)$ y $\beta = c$; más aún $b_1(n) = n^{-2}$ y $b_2(n) = n^{-1}$;

ii) si las condiciones (1) y (4) se satisfacen, entonces la medida de reproducción es $\nu^\alpha(dz)$, definida como en el Teorema 3, la medida de inmigración es $z\nu^\alpha(dz)$ y $\beta = 1$; las constantes de normalización están dadas por $b_1(n) = (r(n)p(n))^{-1}$ y $b_2(n) = (r(n))^{-1}$.

Capítulo 2. Árboles de genes y árboles de especies

El objetivo de este capítulo es brindar al lector algunas de las ideas y nociones importantes de la genética de poblaciones, precisando las relaciones y diferencias entre los árboles de genes y especies. En este sentido, el presente capítulo puede verse como una motivación para el modelo que se introduce en el Capítulo 3, así que aquellos familiarizados con el tema pueden dirigirse directamente a éste último.

De acuerdo a la teoría moderna de la evolución, todos los organismos tienen un ancestro común, esto significa que todas las especies existentes y extintas se relacionan. La filogenética es la rama de la biología que busca determinar dichas relaciones evolutivas, la representación gráfica de éstas da lugar a los árboles filogenéticos.

En sus origenes, los árboles filogenéticos se obtenían a partir de las características comunes entre los organismos. Cuánto se parecen en lo que respecta a caracteres como morfología, anatomía y embriología, indicaba la distancia genética entre estos organismos y por lo tanto su evolución. No fue hasta 1858, que a raíz de la publicación del origen de las especies de Darwin, se sentaron las bases de la teoría de la evolución por medio de la selección natural. El desarrollo de la biología evolutiva y filogenética continuó a lo largo de los años, hasta que en la década de los 60's, Emile Zuckerkandl junto con Linus Pauling descubrieron que las moléculas de ADN y las proteínas que codifican son "documentos de la historia evolutiva". Las proteínas son responsables de lo que es un ser vivo y lo que puede hacer en un sentido físico, mientras que los ácidos nucléicos codifican la información necesaria para producir proteínas y son responsables de transmitir "esta receta" a generaciones subsecuentes. Así, en la actualidad la evolución es un proceso molecular basado en las proteínas y los ácidos nucléicos, que se deriva bajo los mismos principios que Darwin, una molécula varía y algunas de estas variaciones se transmiten a través de las generaciones.

En el intento por reconstruir la historia filogenética, los modelos matemáticos son herramientas indispensables para caracterizar el proceso evolutivo debido a que muchos aspectos no son fácilmente susceptibles a la experimentación directa. Idealmente, un modelo de evolución debe proporcionar una buena descripción de los datos y al mismo tiempo ser parametrizado en una forma que facilite una visión biológica. A la fecha se dispone de una gran cantidad de datos moleculares, se conocen incluso sucesiones completas de todas las secuencias de ADN de un individuo o de una especie. Los modelos estadísticos son una herramienta particularmente importante en el estudio de la evolución molecular ya que utilizan los datos para evaluar tasas, procesos y constricciones en el cambio molecular a lo largo del tiempo. Esto último les permite aignar valores a los parámetros involucrados en el modelo y conocer como ocurrió la evolución molecular.

Bajo esta premisa se han desarrollado métodos de simulación de secuencias y rigurosos marcos de filogénetica estadística, tanto frecuentistas como Bayesianos, a partir de las unidades de ADN (genes) que se conocen. El punto mas débil de estos modelos es que se basan en el árbol de genes para hacer inferencia sobre el árbol de las especies. Sin embargo,

aunque ambos árboles guardan características en común son objetos completamente diferentes. En este capítulo describimos los árboles de genes y árboles de especies, precisando la razón por la cual son diferentes. Así mismo presentamos el coalescente multiespecies, modelo base de los modelos estadísticos. A su vez, proponemos un nuevo intento para calcular la probabilidad de que ambos árboles sean "concordantes".

Capítulo 3. Coalescentes anidados simples

La teoría de coalescencia fue inicialmente formulada por Kingman (1982a). El propósito fue describir la genealogía de una población haploide donde los individuos tienen una reproducción asexual binaria. Por lo tanto es un modelo inadecuado cuando la población tiene fluctuaciones importantes, o cuando la selección natural está presente y no puede ser ignorada. Las situaciones antes descritas corresponden a poblaciones donde una proporción considerable de las líneas ancestrales de la población coalescen. Los modelos utilizados para modelarlas son los procesos Λ-*coalescentes*, introducidos por Pitman (1999), e independientemente por Sagitov (1999). Si suponemos que en los tiempos de coalescencia de un Λ-coalescente, diferentes porciones de la población colisionan obtenemos los procesos *coalescentes con colisiones simultáneas*, definidos por por Schweinsberg (2000) e independientemente por Möhle y Sagitov (2001). Por otra parte, Bertoin y Le Gall (2003) definieron ésta última familia de procesos utilizando el *operador de coagulación* y la llamaron *coalescentes intercambiables*. Para ser un poco más precisos, coagular las particiones π y π' significa unir los bloques de π de acuerdo a una partición π' que llamamos "receta", escribiremos ésta operación como $\mathrm{Coag}(\pi, \pi')$. Para $\pi = (\{1,6,7\},\{2,4,5\},\{3,8\},\{9,10\})$ y $\pi' = (\{1,3\},\{2,4\})$ tenemos que

$$\mathrm{Coag}(\pi, \pi') = (\{1,3,6,7,8\},\{2,4,5,9,10\}).$$

Un coalescente intercambiable $\Pi(t) := (\Pi(t) : t \geq 0)$ es por definición un proceso de Markov a valores en las particiones cuyo semigrupo satisface para todo $t, t' \geq 0$, que condicionalmente a $\Pi(t) = \pi$ la distribución de $\Pi(t + t')$ es la ley de $\mathrm{Coag}(\pi, \pi')$, donde π' es alguna partición intercambiable (cuya ley solo depende de t'). En particular, el Λ-coalascente se conoce como *coalescente intercambiable simple*.

Naturalmente la teoría de las particiones aleatorias juega un papel importante dentro los coalescentes. Es sabido que para cada $n \in \bar{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$, una partición de $B \subset \mathbb{N}$ es una colección numerable $\pi = (\pi_i : i \in \mathbb{N})$ de subcojuntos de B disjuntos dos a dos. En particular, denotamos por $\mathcal{P}_n$ el espacio de particiones de $[n] := \{1, 2, ..., n\}$, donde por convención $n = \infty$ corresponde a $\mathbb{N}$. Las particiones con uno y solamente un bloque que no es un singulete, se llaman simples, escribimos $\mathcal{P}'_n$ para el conjunto de particiones simples de $[n]$.

La creciente demanda de la genética poblacional por el desarrollo y el análisis de modelos que incorporen hipótesis mas realistas ha hecho que la teoría de coalescencia haya sido

expandida. Una población con migración puede ser descrita por los procesos *coalescentes distinguidos* definidos en Foucart (2011). Poblaciones separadas por barreras geográficas son analizadas por Limic y Sturm (2006), a través de los Λ-*coalescente espaciales*. Interesados en plantear un modelo probabilista que describa la dinámica de los árboles de genes y árboles de especies, presentados en el Capítulo 2, definimos una nueva clase de procesos coalescentes que hemos llamado *coalescentes intercambiables anidados simples*, o *snec* por sus iniciales en inglés.

El punto de partida para definir un proceso snec es observar que la dinámica del árbol de especies es independiente del árbol de genes y se puede modelar por un proceso coalescente intercambiable $\mathcal{R}^s := (\mathcal{R}^s(t) : t \geq 0)$. Entonces para cada n, $\mathcal{R}^s_{|_n}$ es un $[n]$-coalescente intercambiable tal que $\mathcal{R}^s_{|_n}(0) = \pi^s$, para alguna partición de $[n]$. De manera que para cada n, el bloque π^s_i es la etiqueta asociada con la i-ésima especie. En la Figura 2 (a), $\pi^s = (\{1,5,7\},\{2,4,8,10\},\{3,6,9\})$ lo que implica que la primera especie es etiquetada con el bloque $\{1,5,7\}$. Por otro lado podemos pensar en identificar los genes que corresponden a la i-ésima especie con bloques de una partición de π^s_i. En la Figura 2 (b), se utilizó la partición $(\{1\},\{5,7\})$ para etiquetar los dos genes de la especie con $\{1,5,7\}$.

Sea $\pi = (\pi^s, \pi^g)$ una partición de $B \subseteq \mathbb{N}^2$ tal que π^s modela la evolución de las especies y π^g la de los genes. Observe que $i \overset{\pi^g}{\sim} j$ implica que $i \overset{\pi^s}{\sim} j$. En este sentido vamos a decir que π^g esta anidada en π^s, escribiremos $\pi^g \subseteq \pi^s$. El conjunto de *particiones anidadas* de $[n]^2$, lo denotamos por $\mathcal{N}_n$. En consecuencia, para cada $n \in \mathbb{N}$ la genealogía del árbol de especies y del árbol de genes puede ser descrita por un proceso $\mathcal{R}(t) := ((\mathcal{R}^s(t), \mathcal{R}^g(t)) : t \geq 0)$ a valores en las particiones anidadas, tal que la distribución condicional de $\mathcal{R}(t+t')$ dado $\mathcal{R}(t) = (\pi^s, \pi^g)$, es la ley de $\mathrm{Coag}_2(\pi, \tilde{\pi})$, donde $\pi = (\pi^s, \pi^g)$, $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ y $\mathrm{Coag}_2(\pi, \tilde{\pi}) = (\mathrm{Coag}(\pi^s, \tilde{\pi}^s), \mathrm{Coag}(\pi^g, \tilde{\pi}^g))$.

Dada una partición anidada π, definimos para $n_1 \geq |\pi^s|$ y $n_2 \geq |\pi^g|$, el conjunto $\tilde{\mathcal{P}}_{n_1,n_2}(\pi)$ de particiones *conservativas* $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ de $\mathbb{N}^2$, tales que la partición $\mathrm{Coag}_2(\pi, \tilde{\pi})$ es anidada. En particular, decimos que una partición $\tilde{\pi}$ es *estrictamente conservativa* para π, si $n_1 = |\pi^s|$ y $n_2 = |\pi^g|$, denotamos el conjunto de estas particiones por $\tilde{\mathcal{P}}(\pi)$. Sin lugar a dudas una de las propiedades clave de los procesos coalescentes es la intercambiabilidad, para extender esta noción al proceso $\mathcal{R}$, definimos una clase de permutaciones que heurísticamente no son más que permutaciones que respetan la propiedad de anidamiento, es decir, al permutar la partición de los genes obtenemos una partición anidada a la partición de las especies (permutada). Para precisar establecemos que cada partición anidada $\pi = (\pi^s, \pi^g)$ de $\mathcal{N}_n$, tiene asociada una única partición $\bar{\pi} \in \mathcal{P}_{|\pi^g|}$ que llamamos *liga*, tal que $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi})$. En términos del modelo, los elementos en el i-ésimo bloque de la partición liga corresponden a las etiquetas de los bloques de la partición de genes anidados a la i-ésima especie. Por convención los bloques de las particiones son ordenados en orden creciente de su mínimo elemento. Para la partición anidada

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

que aparece en la Figura 2, $\bar{\pi} = (\{1,4\},\{2,6,7\},\{3,5\})$. Ahora bien, si $\tilde{\pi}^g$ es una partición

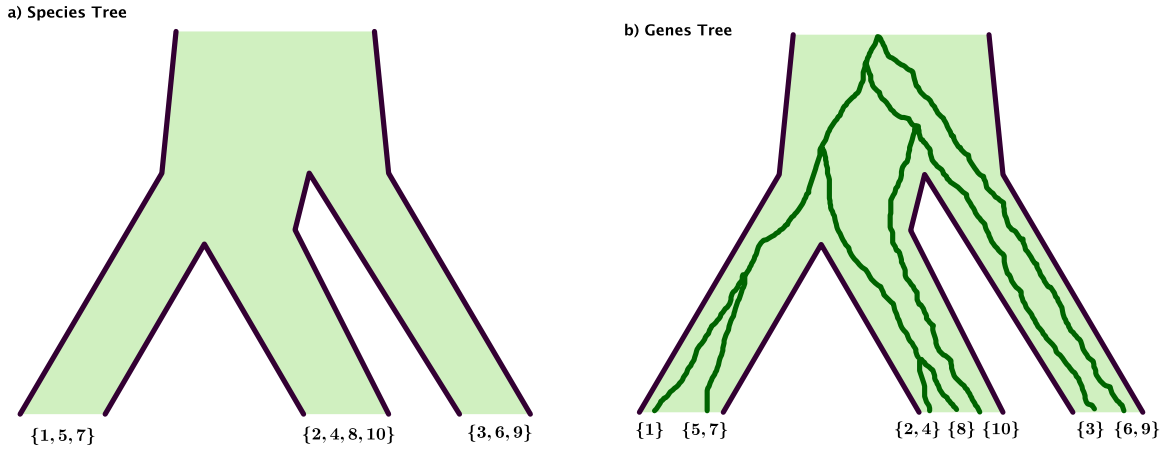

Figura 2: a) Árbol de especies etiquetado con una partición de [10]. b) Árbol de genes asociado al árbol de especies en (a).

aleatoria tal que $\sigma(\tilde{\pi}^g) \overset{\mathcal{L}}{=} \tilde{\pi}^g$, decimos que una permutación σ preserva $\bar{\pi}$ si cumple la siguiente implicación

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

Finalmente definimos un proceso snec como sigue:

Definición 1. Fijo $n \in \mathbb{N}$, para cada $t \geq 0$ sea $\mathcal{R}(t) := ((\mathcal{R}^s(t), \mathcal{R}^g(t)) : t \geq 0)$ un proceso de Markov con valores en $\mathcal{P}_n^2$. Este proceso es llamado *coalescente intercambiable anidado simple*, snec, si

i) Para $t \geq 0$ fijo, $\mathcal{R}^s(t)$ y $\mathcal{R}^g(t)$ son particiones aleatorias intercambiables.

ii) Para cada $t \geq 0$, $\mathcal{R}^g(t) \subseteq \mathcal{R}^g(t)$.

iii) El proceso $(\mathcal{R}^s(t) : t \geq 0)$ es un coalescente intercambiable, esto es, para cada $t, t' \geq 0$ condicionalmente a $\mathcal{R}^s(t) = \pi^s$, la distribución de $\mathcal{R}^s(t+t')$ es la ley de $\text{Coag}(\pi^s, \tilde{\pi}^s)$, donde $\tilde{\pi}^s$ es una partición aleatoria intercambiable simple independiente de $\mathcal{R}^s(t)$, cuya ley sólo depende de t'.

iv) Condicionalmente a $\mathcal{R}(t) = (\pi^s, \pi^g)$, si $\bar{\mathcal{R}}(t)$ denota la partición liga de $\mathcal{R}(t)$ entonces tal que para todo $t \geq 0$, $\mathcal{R}^g(t + t')$ tiene la misma ley que $\text{Coag}(\pi^g, \tilde{\pi}^g)$, donde $\tilde{\pi}^g$ es una partición aleatoria con la misma ley que $\sigma(\tilde{\pi}^g)$, con σ una permutación que preserva $\bar{\mathcal{R}}(t)$.

Ahora estamos interesados en analizar las transiciones de un proceso snec $(\mathcal{R}(t) : t \geq 0)$, en este sentido es suficiente considerar las tasas de transición de sus restricciones,

$$q_{\pi,\pi'} := \lim_{t \to 0+} \frac{1}{t} \mathbb{P}(\mathcal{R}_{|n}(t) = \pi' \,|\, \mathcal{R}_{|n}(0) = \pi).$$

Sea $\tilde{\pi}$ la receta para obtener π' y $\bar{\pi}$ la partición liga de π, i.e. $\pi' = \mathrm{Coag}_2(\pi, \tilde{\pi})$ y $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi})$. Una primera observación es que las tasas de salto son cero cuando $\tilde{\pi}$ es no conservativa, recordemos que en este caso π' no es un elemento de $\mathcal{N}_n$. Ahora bien, para todo n las transiciones de $\mathcal{R}_{|\mathbf{n}}$ sólo dependen del número de bloques de las particiones de especies y genes, notemos también que si $\bar{\pi}$ es un elemento de $\mathcal{P}_k$, para algún k, no es difícil ver que su cardinalidad coincide con la de la partición π^s, y $k = |\pi^g|$. En consecuencia para todo $m > n$, si π'' es un elemento de $\mathcal{N}_m$ con partición liga $\bar{\pi}$, entonces la tasa de salto para $\mathcal{R}_{|\mathbf{m}}$ de π'' a $\mathrm{Coag}_2(\pi'', \tilde{\pi})$ es también $q_{\pi,\pi'}$. Por lo tanto, los procesos snec $\mathcal{R}$ están completamente caracterizados por $(\tilde{q}_{\bar{\pi},\tilde{\pi}} : \bar{\pi} \in \mathcal{P}_\infty, \tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}), \rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n), n \in \mathbb{N})$ donde $\tilde{q}_{\bar{\pi},\tilde{\pi}} := q_{\rho_{|\mathbf{n}},\mathrm{Coag}_2(\rho_{|\mathbf{n}},\tilde{\pi})}$, es decir,

$$\tilde{q}_{\bar{\pi},\tilde{\pi}} = \lim_{t\to 0+} \frac{1}{t}\mathbb{P}(\mathcal{R}_{|\mathbf{n}}(t) = \mathrm{Coag}_2(\rho_{|\mathbf{n}}, \tilde{\pi}) \,|\, \mathcal{R}_{|\mathbf{n}}(0) = \rho_{|\mathbf{n}}). \tag{7}$$

Uno de resultados obtenidos asegura que las tasas de transición $\tilde{q}_{\bar{\pi},\tilde{\pi}}$, pueden ser caracterizadas en términos de una medida en $\tilde{\mathcal{P}}(\bar{\pi})$.

Proposición 1. Sea $(\tilde{q}_{\bar{\pi},\tilde{\pi}} : \bar{\pi} \in \mathcal{P}_\infty, \tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}), \rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n), n \in \mathbb{N})$ la familia de saltos de algún snec $\mathcal{R}$. Existe una única familia $(\mu_{\bar{\pi}}, \bar{\pi} \in \mathcal{P}_\infty)$ donde $\mu_{\bar{\pi}}$ es una medida sobre $\tilde{\mathcal{P}}(\bar{\pi})$ tal que para cualquier $\bar{\pi}$, $\mu_{\bar{\pi}}(\mathbf{0}_{\infty^2}) = 0$ y

$$\mu_{\bar{\pi}}(\mathcal{P}_{\infty,\tilde{\pi}}(\rho)) = \tilde{q}_{\bar{\pi},\tilde{\pi}}.$$

Concluimos este capítulo presentando la construcción de un ejemplo de proceso snec, con tasas de transición determinadas por una medida aleatoria de Poisson. Para ser más precisos consideramos una sucesión $\zeta = (\zeta_i)_{i\in\mathbb{N}}$ de ensayos Bernoulli con probabilidad de éxito x, y una sucesión independiente $(\xi^i_j)_{i,j\in\mathbb{N}}$ de variables Bernoulli con parámetro y. Dada una partición anidada π ligada por $\bar{\pi}$, se construye una partición receta $\tilde{\pi}$ conservativa para π, asociando el par (ζ_i, ξ^i_j) con el j-ésimo elemento del bloque $\bar{\pi}_i$. Así las especies y genes que se juntan corresponden a los pares $(\zeta_i, \xi^i_j) = (1, 1)$, ver Figura 3. Sea P_{xy} la ley de $\tilde{\pi}$ y ν_{sg} una medida sigma finita en $[0, 1]^2$ tal que

$$\nu_{sg}(\{(0, 0)\}) = 0 \quad \text{y} \quad \int_{[0,1]}\int_{[0,1]} (x^2 + xy^2)\nu_{sg}(dx, dy) < \infty. \tag{8}$$

Luego

$$\varrho_{\nu_{sg}}(d\tilde{\pi}) = \int_{[0,1]}\int_{[0,1]} \nu_{sg}(dx, dy)P_{xy}(d\tilde{\pi}),$$

es una medida en $(\mathcal{P}'_\infty)^2$ que satisface la siguiente condición

$$\varrho_{\nu_{sg}}(\tilde{\pi} \in \tilde{\mathcal{P}}(\rho) : \tilde{\pi}_{|\mathbf{n}} \neq \mathbf{0}_{|\rho_{|\mathbf{n}}|}) < \infty, \quad \text{para cada } n \in \mathbb{N}.$$

Ahora podemos construir un proceso snec $\mathscr{R} = (\mathscr{R}(t) : t \geq 0)$ con transiciones determinadas por los saltos de un proceso puntual de Poisson M en $(0, \infty) \times (\mathcal{P}'_\infty)^2$ con intensidad

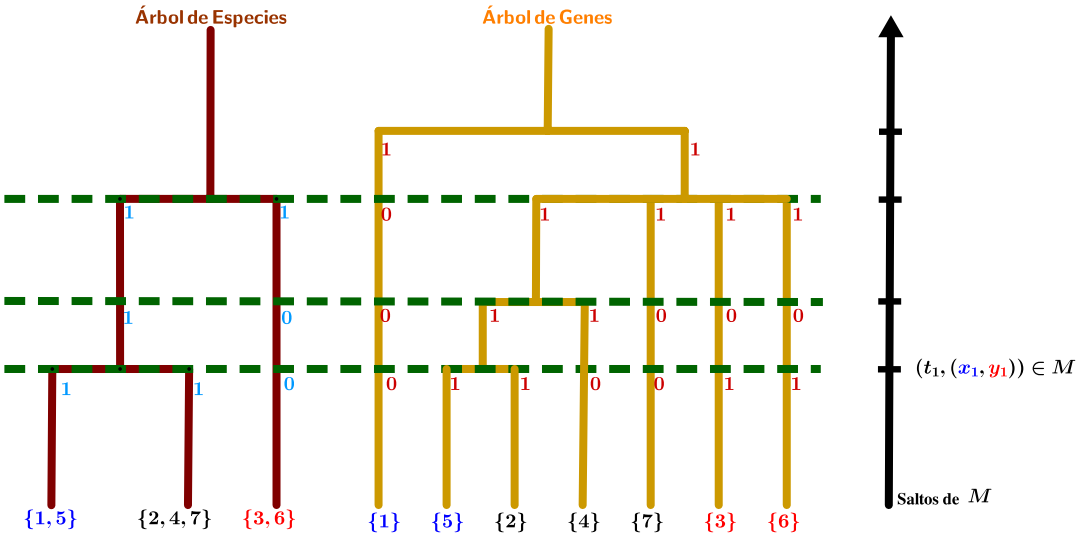

Figura 3: Construcción poissoniana del proceso snec $\mathcal{R}_{|\mathbf{n}}$. Observe que los genes sólo fusionan cuando la especie a la que pertenece ha sido "activada", es decir, tiene asociada una variable Bernoulli(x) con éxito.

$dt \otimes \varrho_{\nu_{sg}}(d\tilde{\pi})$. Más precisamente, para cada $n \in \mathbb{N}$ consideramos M_n, la imagen de M por el mapeo $(t, \tilde{\pi}) \to (t, \tilde{\pi}_{|\mathbf{n}})$. Así como la familia de sus átomos en $(0,\infty) \times ((\mathcal{P}'_\infty)^2 \setminus \mathbf{0}_{|\rho_{|\mathbf{n}|}})$, ordenados en forma creciente por la primera coordenada. Definimos $\mathscr{R}_{|\mathbf{n}}(t) = \rho_{|\mathbf{n}}$ para $t \in [0, t_1)$, y recursivamente

$$\mathscr{R}^{\mathbf{n}}(t_i) = \mathrm{Coag}_2(\mathscr{R}^{\mathbf{n}}(t_i-), \tilde{\pi}^{(i)}(t_i)), \quad \text{para todo } t \in [t_i, t_{i+1}).$$

Como un siguiente paso probamos que estas sucesiones son consistentes y en consecuencia tenemos que existe un proceso snec $\mathscr{R} = (\mathscr{R}(t), t \geq 0)$ tal que cada $n \in \mathbb{N}$, $\mathscr{R}_{|\mathbf{n}}(t) = \mathscr{R}^{\mathbf{n}}(t)$, de hecho establecimos la siguiente proposición.

Proposición 2. Para cada $t \geq 0$, la sucesión de particiones aleatorias $(\mathscr{R}^{\mathbf{n}}(t), n \in \mathbb{N})$ es consistente. Si denotamos por $\mathscr{R}(t)$ la única partición de $\mathcal{N}_\infty$ tal que $\mathscr{R}_{|\mathbf{n}}(t) = \mathscr{R}^{\mathbf{n}}(t)$ para cada $n \in \mathbb{N}$, entonces el proceso $\mathscr{R} = (\mathscr{R}(t), t \geq 0)$ es un snec con tasa de salto $\mu = \varrho_{\nu_{sg}}$, que inicia en $\rho = (\bar{\pi}, \mathbf{0}_\infty)$. Observe que en este caso la tasa las tasas de salto no dependen del estado.

Por otro lado, gracias a la condición (8) tenemos que las medidas definidas como sigue

$$u_s(dx) = \int_{y\in[0,1]} \nu_{sg}(dx, dy), \qquad u_g(dy) = \int_{x\in[0,1]} x\nu_{sg}(dx, dy).$$

satisfacen

$$u_g(0) = u_s(0) = 0 \quad \text{y} \quad \int_{[0,1]} x^2 u_s(dx), \int_{[0,1]} y^2 u_s(dy) < \infty.$$

Luego por un resultado de Bertoin (2006), existen coalescentes intercambiables simples Π^s y Π^g con medida de coagulación u_s y u_g, respectivamente. Más aún $\Pi^s \overset{\mathcal{L}}{=} \mathscr{R}^s$. Así podemos pensar que las propiedades de los coalescentes intercambiables se tienen en algún sentido para un proceso snec $\mathcal{R} = (\mathcal{R}^s, \mathcal{R}^g)$. En particular, si decimos que $\mathcal{R}$ baja del infinito cuando

$$\mathbb{P}(|\mathcal{R}^g(t)| < \infty, \text{ para todo } t > 0) = 1.$$

Las condiciones necesarias y suficientes para que el proceso snec antes construido baje del infinito, están dadas por el siguiente resultado.

Proposición 3. El coalescente anidado intercambiable simple $\mathscr{R}$ con tasa de salto $\mu = \varrho_{\nu_{sg}}$ baja del infinito si y solamente si los procesos coalescentes intercambiables simple Π^s y Π^g, antes definidos, bajan del infinito.

La construcción del proceso $\mathscr{R}$ da lugar a intentar construir todos los procesos $\mathcal{R}$ de manera Poissoniana, y así poder eliminar la dependencia del estado presente que aparece en la medida μ, introducida en la Proposición 1. Esto es parte de un trabajo en curso, al final del capítulo mencionamos algunas de sus siguientes direcciones.

Introduction

This thesis is devoted to stochastic processes theory used to model phenomena with interest in biology both population and molecular. To be specific, we study the branching processes theory and coalescent theory. We divided this work in three chapters. In the first one we present a model to describe the genealogy in a population with rare neutral mutation presented by Bertoin (2010). Assuming weaker hypothesis about the reproduction law, we establish the asymptotic result obtained by him about the allelic subfamilies structure. Immediately, we study the genes and species tree. More precisely, in Chapter 2 we are interested in the probability gene trees concordance with species tree topology. Finally, in Chapter 3 we define a new class of coalescent process to model the dynamic of these trees.

Chapter 1. On branching process with rare neutral mutations

A Galton Watson process models a population in which at every generation each individual begets according to a fixed offspring distribution and independently of the other individuals. Imagine that neutral mutations may happen, so that a child can be either a clone of its parent or a mutant, and the reproduction laws of clones and mutants are identical. We shall further suppose that each time a mutation occurs, it produces a mutant with a genetic type which has never been observed before. Hence we get a Galton Watson process with neutral mutations $\{Z_n^{(+)} : n \in \mathbb{Z}_+\}$. Let $\xi^{(c)}$ and $\xi^{(m)}$ be non-negative integer-valued random variables, which describe respectively the number of clone children and the number of mutant children of a typical individual. Denote the size of a typical family by $\xi^{(+)} = \xi^{(c)} + \xi^{(m)}$. The genealogy of a Galton Watson process is described by a planar rooted tree, with edges connecting parents to children an assigning marks to the edges between parents an the mutant children. According to Bertoin (2010), an individual has the n-th type if its genotype has been affected by n mutations, that if its ancestral line comprises exactly n marks. Denote by T_n the total population of individuals of the n-th type, see Figure 4. Those individuals with parent of $(n-1)$-type are known mutants of the n-type. We write M_n for the total number of these individuals, agreeing that mutants of the 0-th type are the ancestors. The Galton-Watson tree with

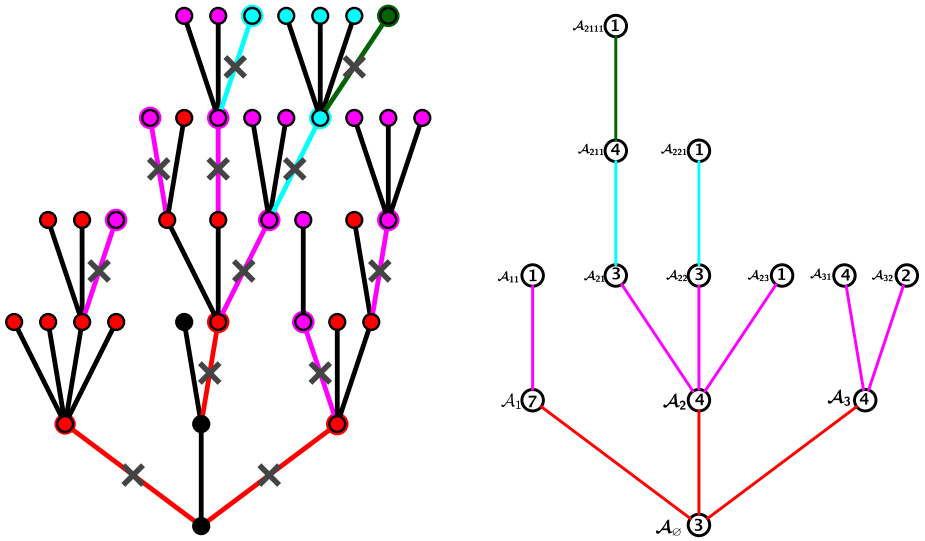

Figure 4: Genealogical tree with mutations (left) and tree of alleles (right). The colors represent the different alleles. Vertices in red correspond to 1-type individual and vertices in magenta are individuals of the 3-type. The edges in color are stopping lines associated with the mutants. The labels on the tree of alleles are the sizes of the corresponding allelic sub-failies.

neutral mutations has a branching property called general, subtrees with root, a mutant of the n-type are independent copies of the original tree. As a consequence $\{M_k : k \in \mathbb{Z}_+\}$ is a Galton Watson process and, $\{(T_k, M_{k+1}) : n \in \mathbb{Z}_+\}$ is a Markov chain with transition probabilities depends on the second coordinate. From this latter observation a new tree, called tree of alleles, is constructed. Heuristically, it is constructed colapsed in one vertex the subtrees with root a mutant of the k-type, and labeling this vertex with the total number of vertex of the subtree. By convention the labels are ordered uniformly at random. Observe that the sum of the sizes of allelic sub-families of the type k is T_k, moreover the number of vertices at level k is M_k. In this sense, the tree structure of the alleles tree, take into account only the vertices of the tree and not the labels, describe the genealogy of a Galton-Watson because of $\{M_n : n \in \mathbb{Z}_+\}$ is it. Motivated by this observation in our work, we establish the versions of some classical results of branching processes for a Galton-Watson process with neutral mutations.

We consider the extinction time of mutants

$$T = \inf\{n \geq 1 : M_n = 0\},$$

assuming that is finite almost surely, we constuct the version conditioned to non extinction of mutants of the chain $\{(T_n, M_{n+1} : n \in \mathbb{Z}_+)\}$, as a first step we find a probability measure using the h-transform and then we verify that under this measure the transitions of this process has the transitions of the process that we are interested, this is written in the following result.

Theorem 1. Let $a \in \mathbb{Z}_+$ and $\mathcal{F}_n$ the natural filtration of the process $\{(T_{n-1}, M_n) : n \in \mathbb{N}\}$. There exits a probability measure $\mathbb{P}_a^{\uparrow}$ that can be expressed as a h-transform of $\mathbb{P}_a$ using the $(\mathcal{F}_n)$-martingale

$$Y_n = \frac{M_n q^{M_n - a}}{(f'(q))^n} \mathbf{1}_{\{n<T\}},$$

where $f(y) = \mathbb{E}_1(y^{M_1})$ and $q = \mathbb{P}_1(0 < T < \infty)$. That is

$$d\mathbb{P}_a^{\uparrow}|_{\mathcal{F}_n} = \frac{Y_n}{a}\, d\mathbb{P}_a|_{\mathcal{F}_n}, \quad n \in \mathbb{N}.$$

Furthermore, $\mathbb{P}_a^{\uparrow}$ is the law of a Markov chain $\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\}$ with n-step transition probabilities,

$$Q^n_{(i,j),(k,l)} = \frac{l q^{l-j}}{j(f'(q))^n} P^n_{(i,j),(k,l)}, \quad j, l \geq 1,$$

where $\{P^n_{(i,j),(k,l)} : i, j, k, l \in \mathbb{Z}_+\}$ denotes the n-step transition probabilities of $\{(T_n, M_{n+1}), n \in \mathbb{Z}_+\}$.

We observe that under $\mathbb{P}_a^{\uparrow}$ the extinction time of mutants is infinite almost surely because of

$$\mathbb{P}_a^{\uparrow}(n < T) = \mathbb{P}_a\left(\frac{Y_n}{a}\mathbf{1}_{\{n<T\}}\right) = 1.$$

Theorem 2. Suppose that $\mathbb{E}(\xi^{(c)}) < 1$ and $\mathbb{E}(\xi^{(+)}) \leq 1$.

i) Let $a, n \in \mathbb{N}$ with n fixed. The conditional law of the process $\{(T_k, M_{k+1}) : 0 \leq k \leq n-1\}$ under $\mathbb{P}_a(\cdot | n + k < T < \infty)$ converges, as $k \to \infty$, towards the probability measure $\mathbb{P}_a^{\uparrow}$, in the sense that for any n

$$\lim_{k\to\infty} \mathbb{P}_a(A | n + k < T < \infty) = \mathbb{P}_a^{\uparrow}(A), \quad \forall A \in \mathcal{F}_n.$$

ii) The Yaglom limit

$$\lim_{n\to\infty} \mathbb{P}(T_{n-1} = i, M_n = j | n < T < \infty),$$

exists and has a generating function $\widehat{\varphi}(x, y)$ such that for all $n \in \mathbb{N}$,

$$m^n \widehat{\varphi}(x, y) = \widehat{f}(\varphi_n(x, y)) - \widehat{f}(\varphi_n(x, 0)), \quad x, y \in [0, 1].$$

Besides, it is well known from the classical theory of branching process that a sequence of Galton-Watson processes, properly rescaled state space converges to the called Jirina process (Jiřina (1958)). In order to obtain a similar result for a Galton-Watson process with rare neutral mutations, in Bertoin (2010) is defined a tree-indexed CSBP, with

reproduction ν, as a process with values in $(0,\infty)$, indexed in the set of finite sequences of positive integers

$$\mathbb{U} = \bigcup_{k\in\mathbb{Z}_+} \mathbb{N}^k,$$

where $\mathbb{N} = \{1,2,...\}$ and $\mathbb{N}^0 = \{\varnothing\}$, such that conditionally on the vertices at level k, the sequence of vertices in the following level is distributed as the family of the atoms of ν, a Poisson random measure on $(0,\infty)$. Recalling the construction of Jirina process from subordinators, we have that vertices at level k represent the sizes of the sub-families at generation k in a Jirina process with reproduction law ν, which descend from a parent at generation $k-1$.

Let $\mathbb{P}_a^p$ be, the probability measure of a Galton-Watson process with neutral mutations started from a ancestors and p, the probability of individuals beget mutants. Interested in to analyze the asymptotic behavior of a sequence of tree of alleles in a more general framework that the studied in Bertoin (2010), where the reproduction law has finite variance, we consider a sequence of Galton-Watson processes with neutral mutations with critical reproduction law such that

$$\bar{\pi}^+(j) := \mathbb{P}(\xi^{(+)} > j) \in RV_\infty^{-\alpha}, \tag{9}$$

where $RV_\infty^{-\alpha}$ denotes the class of functions which are regularly varying with index $-\alpha$ at ∞. This implies that there exists a regular varying function r such that

$$r(n)\mathbb{P}(\xi^+ > ny) \xrightarrow[n\to\infty]{} c_\alpha y^{-\alpha}, \quad \forall y > 0, \tag{10}$$

where $c_\alpha = 1/\Gamma(3-\alpha)$. The notation $f \sim g$ refers to $\lim_{x\to\infty} f(x)/g(x) = 1$. Unlike of Bertoin (2010), where the number of ancestor and the rate of mutations have the following asymptotic behavior

$$a(n) \sim nx \quad \text{and} \quad p(n) \sim cn^{-1}, \quad \text{as } n \to \infty. \tag{11}$$

where c, x are some positive constants, we use the hypothesis below

$$a(n) \sim xr(n)p(n) \quad \text{and} \quad p(n) \sim cn^{-1}, \quad n \to \infty. \tag{12}$$

In this scheme and with the notation $\Longrightarrow$, to refers the convergence in distribution as $n \to \infty$ and $\mathcal{L}\left(\cdot, \mathbb{P}_{a(n)}^{p(n)}\right)$ to the distribution of the process under $\mathbb{P}_{a(n)}^{p(n)}$, one of our main results determines the constants of normalization to get the convergence of a sequence of alleles tree, associated to Galton Watson with neutral mutations defined in the latter framework.

Theorem 3. If (9) and (12) holds. Then, the following convergence holds in the sense of finite dimensional distributions

$$\mathcal{L}\left(((r(n))^{-1}\mathcal{A}_u^{(n)}, (r(n)p(n))^{-1}d_u^{(n)}) : u \in \mathbb{U}), \mathbb{P}_{a(n)}^{p(n)}\right) \Longrightarrow ((\mathcal{Z}_u^{1/\alpha}, \mathcal{Z}_u^{1/\alpha}) : u \in \mathbb{U}),$$

where $\{\mathcal{Z}_u : u \in \mathbb{U}\}$ is a tree-indexed CSBP with reproduction measure

$$\nu^{\alpha}(dy) = c' x^{-1-1/\alpha} dy, \quad y > 0,\ \alpha \in (1,2), \tag{13}$$

where $c' = \alpha^{-1}/\Gamma(1-\alpha^{-1})$.

Unlike classical results of Galton Watson process where the convergence of generations sizes is obtained, with this kind of results we also get the convergence of the population genealogy.

We also establish the convergence of the finite dimensional distributions of the rescaled chain $\{T_n, M_{n+1}\}$, conditioned to non-extinction of mutants, towards a continuous state branching process with immigration in discrete time.

Theorem 4. If the reproduction law is critical, there exist sequences $b_1(n)$ and $b_2(n)$ such that the following joint convergence in the sense of finite dimensional distributions holds:

$$\mathcal{L}\left(\{(b_1(n)T_{k-1}, b_2(n)M_k) : k \in \mathbb{Z}_+\}, \mathbb{P}_{a(n)}^{p(n)\uparrow}\right) \Longrightarrow \{(Y_k, \beta Y_k) : k \in \mathbb{Z}_+\},$$

where $\{Y_k : k \in \mathbb{Z}_+\}$ is a CSBP with immigration, which is characterized by the following conditions:

i) if the reproduction law has finite variance σ^2 and (1.11) holds, then its reproduction measure is given by (1.13) and the immigration measure is $z\nu(dz)$ and $\beta = c$; moreover $b_1(n) = n^{-2}$ and $b_2(n) = n^{-1}$;

ii) if the assumptions (1.19) and (1.21) hold, the reproduction measure is $\nu^{\alpha}(dz)$ as defined in (1.22), the immigration measure is $z\nu^{\alpha}(dz)$ and $\beta = 1$; the normalizing constants are given by $b_1(n) = (r(n)p(n))^{-1}$ and $b_2(n) = (r(n))^{-1}$.

Chapter 2. Gene trees and species trees

The aim of this chapter is to present some of the important ideas in populations genetic, specifying the relationships and differences between gene trees and species, in this sense can be seen as a motivation to the model that we will introduce in Chapter 3, so that, the reader being familiar with this subject can skip it.

According to the modern theory of evolution, all organisms have a common ancestor, this means that all existing and extinct species are related, the phylogeny is the branch of biology that seeks to determine such relationships. The graphical representation of the evolutionary relationships among species of interest resulting phylogenetic trees. Phylogeneticists get their trees from morphological, physiological and molecular characteristics of existing bodies.

In its origins, phylogenetic trees derived from the common features of organisms, the look in regard to morphology, anatomy, embryology, among other characters indicating

the genetic distance between these agencies and therefore its evolution. It was not until 1858, that following the publication of the Origin of Species Darwin, the foundations of the theory of evolution sat through natural selection. The development of evolutionary biology and phylogenetic continued throughout the years, until in the decade of the 60, Emile Zuckerkandl with Linus Pauling discovered that molecules of DNA and proteins they encode are "documents of evolutionary history" given the relative consistency with which accumulate variations (mutations). The proteins are responsible for what is a living being and what you can do in a physical sense, while the nucleic acids encoding the necessary information to produce proteins and are responsible "transmit this recipe" to subsequent generations. So, now molecular evolution is a process based on proteins and nucleic acids, derived under the same principles that Darwin, a molecule varies and some of these variations are transmitted through the generations.

In an attempt to reconstruct the phylogenetic history, mathematical models are indispensable tools for characterizing the evolution process because many aspects are not easily susceptible to direct experimentation. Ideally, a model of evolution should provide a good description of the data and at the same time be configured in a manner that facilitates a biological vision. Given that to date has a large amount of molecular data (even whole sequences of all DNA sequences from an individual or a species known), the statistical models are a particularly important tool in the study of molecular evolution because they use the data to assess rates, processes and constraints on the molecular changes over time, which in turn allows you to assign values to parameters and meet as happened molecular evolution. So they have developed simulation methods and sequences rigorous phylogenetic statistical frameworks, both frequentist and Bayesian, from units of DNA (genes) is known. The weakest point of these models is based on the tree of genes to make inferences about the tree species, however, although both trees keep common features are completely different objects. In this chapter we describe gene trees and tree species, stating the reason why they are different, likewise present the multispecies coalescent, which is the basis of statistical models model, proposing to turn a new attempt to reconstruct the tree species, that allows to calculate the probability that both trees are "consistent".

Chapter 3. Simple nested coalescent

Coalescent theory was first formulated by Kingman (1982a), for the purpose of describing the genealogy of a haploid population where individuals have a binary asexual reproduction, it is therefore improper pattern when the population have significant fluctuations, or when natural selection is present, and can not be ignored. The above described situations correspond to populations where a considerable proportion of the ancestral lines coalesce, and are modeled by Λ-coalescent processes introduced on one side by Pitman (1999), and in the other side by Sagitov (1999). Assuming that at the coalescence times

of Λ-coalescent, different portions of the population merge we get coalescent processes with simultaneous collisions as defined by Möhle and Sagitov (2001), and independently by Schweinsberg (2000).

Besides Bertoin and Le Gall (2003) defined this last family processes using the coagulation operator and called these exchangeable coalescents, in particular the Λ-coalascent is known simple exchangeable coalescent. To be precise, coagulate π by π' writing $\mathrm{Coag}(\pi,\pi')$, means merge the blocks of π according to the called "recipe" partition π'. For instance, if $\pi = (\{1,6,7\},\{2,4,5\},\{3,8\},\{9,10\})$ and $\pi' = (\{1,3\},\{2,4\})$, then

$$\mathrm{Coag}(\pi,\pi') = (\{1,3,6,7,8\},\{2,4,5,9,10\}).$$

Thus an exchangeable coalescent $\Pi(t) := (\Pi(t) : t \geq 0)$ is a Markov process with values in the partitions with semigroup such that for every $t, t' \geq 0$, the condicional distribution of $\Pi(t+t')$ given $\Pi(t) = \pi$ is the law of $\mathrm{Coag}(\pi,\pi')$, where π' is some exchangeable random partition (whose law only depends on t').

The growing demand of population genetics for the development and analysis of models that incorporate more realistic hypothesis has made this theory to expand. For example, a population with immigration is described by coalescing processes distinguished by Foucart (2011) while populations separated by geographical barriers are analyzed by Limic and Sturm (2006), through the spacial Λ-coalescent. In the aim to propose a probabilistic model describing the dynamic gene trees and tree species, presented in Chapter 2, we define a new class of coalescent processes called simple nested exchangeable coalescent, or snec for short. In this sense, the theory of random partitions plays an important role. It is known that for every $n \in \bar{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$, a partition of $B \subset \mathbb{N}$ is a countable collection $\pi = (\pi_i : i \in \mathbb{N})$ of pairwise disjoins subsets of B. In particular, denote by $\mathcal{P}_n$ the space of partitions of $[n] := \{1,2,...,n\}$, we agree that $[\infty] := \mathbb{N}$. The partitions with exactly one non singleton block are called simples, set $\mathcal{P}'_n$ the set of those partitions for the set $[n]$.

The starting point to define a snec process is to observe that the dynamic of a species tree is independent of a gene tree and can be modeled by an exchangeable coalescent process $\mathcal{R}^s := (\mathcal{R}^s(t) : t \geq 0)$, so that, for every n, $\mathcal{R}^s_{|n}$ is a$[n]$-exchangeable coalescent such that $\mathcal{R}^s_{|n}(0) = \pi^s$, for some partition of $[n]$. Then for every n, the block π^s_i is the label associated with the i-th specie and we can think to identify the corresponding genes to the i-th specie with the blocks of this partition. (For an example see Figure 5). Let $\pi = (\pi^s, \pi^g)$ be a partition of $B \subseteq \mathbb{N}^2$ such that π^s models the evolution of species and π^g the evolution of genes. Observe that $i \overset{\pi^g}{\sim} j$ implies $i \overset{\pi^s}{\sim} j$. In this sense let us say that π^g is nested in π^s, we write $\pi^g \subseteq \pi^s$. The set of nested partitions of $[n]^2$ is denoted by $\mathcal{N}_n$. Let $\pi = (\pi^s, \pi^g)$ be a partition of $B \subseteq \mathbb{N}^2$ such that π^s models the evolution of species and π^g the evolution of genes. Observe that $i \overset{\pi^g}{\sim} j$ implies $i \overset{\pi^s}{\sim} j$. In this sense let us say that π^g is nested in π^s, we write $\pi^g \subseteq \pi^s$. The set of nested partitions of $[n]^2$ is denoted by $\mathcal{N}_n$. A consequence, for every $n \in \mathbb{N}$ the genealogy of species tree and the genes tres can be described by a process $\mathcal{R}(t) := ((\mathcal{R}^s(t), \mathcal{R}^g(t)) : t \geq 0)$ with values in the nested partitions,

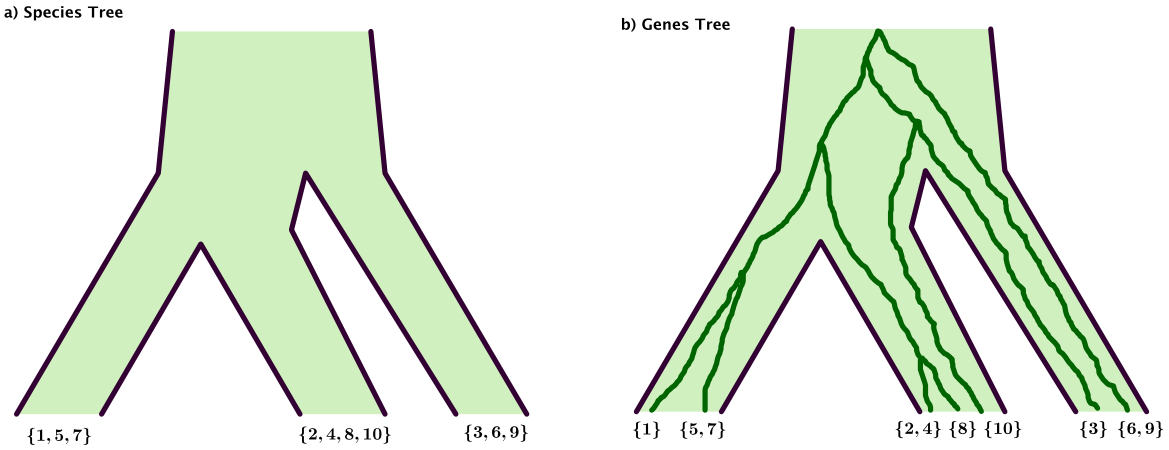

Figure 5: a) Species tree labeled with a partition of $[10]$. b) Genes tree associated to the species tree in (a).

such that the conditional distribution of $\mathcal{R}(t+t')$ given $\mathcal{R}(t) = \pi$ is the law of $\mathrm{Coag}_2(\pi, \tilde{\pi})$, where $\pi = (\pi^s, \pi^g)$, $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ and $\mathrm{Coag}_2(\pi, \tilde{\pi}) = (\mathrm{Coag}(\pi^s, \tilde{\pi}^s), \mathrm{Coag}(\pi^g, \tilde{\pi}^g))$. Given a nested partition π, we define for $n_1 \geq |\pi^s|$ and $n_2 \geq |\pi^g|$, the set $\tilde{\mathcal{P}}_{n_1,n_2}(\pi)$ of *conservative* partitions $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ of $\mathbb{N}^2$ such that $\mathrm{Coag}_2(\pi, \tilde{\pi})$ is nested. In particular, we say that a partition $\tilde{\pi}$ is strictly conservative for π, if $n_1 = |\pi^s|$ and $n_2 = |\pi^g|$, denote the set of these partitions by $\tilde{\mathcal{P}}(\pi)$.

One key property for the coalescents processes is the exchangeability, in order to have this property for the process $\mathcal{R}$, we define the class of permutations preserving the nested property, i.e. the permutation of genes partition is a partition nested into the species partition permuted. To be precise, for every nested partition $\pi = (\pi^s, \pi^g)$ of $[n']$ there exists a unique partition $\bar{\pi}$ in $[|\pi^g|]$ such that $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi})$, the partition $\bar{\pi}$ is called *link*. In terms of the model, the i-th block of the link partition determines the gene blocks nested to the i-th specie. For the nested partition

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

given in Figure 5, $\bar{\pi} = (\{1,4\},\{2,6,7\},\{3,5\})$. From here we can deduce for instance that the blocks 1 and 4 in the genes partition, are nested in the first specie. Now, if $\tilde{\pi}^g$ is a partition such that $\sigma(\tilde{\pi}^g) \overset{\mathcal{L}}{=} \tilde{\pi}^g$, we say that a permutation σ preserves $\bar{\pi}$, if the following implication holds

$$i \overset{\bar{\pi}}{\sim} j \Rightarrow \sigma(i) \overset{\bar{\pi}}{\sim} \sigma(j). \tag{14}$$

Finally we define a snec process as follows:

Definition 1. Fix $n \in \bar{\mathbb{N}}$, for every $t \geq 0$ let $\mathcal{R} := ((\mathcal{R}^s(t), \mathcal{R}^g(t)) : t \geq 0)$ be a Markov process with values in $\mathcal{P}_n^2$. This process is called a simple nested exchangeable coalescent, snec for short, if

i) For any $t \geq 0$, $\mathcal{R}^g(t)$ and $\mathcal{R}^s(t)$ are exchangeable random partitions.

ii) for any $t \geq 0$, $\mathcal{R}^g(t) \subseteq \mathcal{R}^s(t)$;

iii) The process $(\mathcal{R}^s(t) : t \geq 0)$ is a simple exchangeable coalescent process: for any $t, t' \geq 0$, the conditional distribution of $\mathcal{R}^s(t+t')$ given $\mathcal{R}^s(t)$ is the law of $\mathrm{Coag}(\mathcal{R}^s(t), \tilde{\pi}^s)$ where $\tilde{\pi}^s$ is some simple exchangeable random partition independent of $\mathcal{R}^s(t)$, whose law just depends on t'.

iv) 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 $\mathrm{Coag}(\mathcal{R}^g(t), \tilde{\pi}^g)$, where $\tilde{\pi}^g$ is a random partition such that $\sigma(\tilde{\pi}^g) \overset{\mathcal{L}}{=} \tilde{\pi}^g$ for any permutation σ preserving $\bar{\mathcal{R}}(t)$.

Next we analyze the transitions of a snec process $(\mathcal{R}^(t) : t \geq 0)$, in this aim it is enough to consider the rates of jump of its restrictions,

$$q_{\pi,\pi'} := \lim_{t \to 0+} \frac{1}{t} \mathbb{P}(\mathcal{R}_{|n}(t) = \pi' \,|\, \mathcal{R}_{|n}(0) = \pi).$$

Let $\tilde{\pi}$ be the recipe to obtain π' from π, and $\bar{\pi}$ the link partition of π, i.e. $\pi' = \mathrm{Coag}_2(\pi, \tilde{\pi})$ and $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi})$. Observe that if the recipe $\tilde{\pi}$ is not conservative the rate jump is equal to zero, because π' is not an element of $\mathcal{N}_n$. Besides, for all n the transitions of $\mathcal{R}_{|\mathbf{n}}$ only depend on the number of blocks of the species and gene partitions, we also observe that $\bar{\pi}$ is an element of $\mathcal{P}_k$ where $k = |\pi^s|$, and its cardinality is π^s. Thus for every $m > n$, if $\pi'' \in \mathcal{N}_m$ and it is linked by $\bar{\pi}$, then the jump rate of $\mathcal{R}_{|\mathbf{m}}$ from π'' to $\mathrm{Coag}(\pi'', \tilde{\pi})$ is $q_{\pi,\pi'}$. Therefore the family of jump rates and hence the snec $\mathcal{R}$, is fully characterized by the family $(\tilde{q}_{\bar{\pi},\tilde{\pi}} : \bar{\pi} \in \mathcal{P}_\infty, \tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}), \rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n), n \in \mathbb{N})$ where $\tilde{q}_{\bar{\pi},\tilde{\pi}} := q_{\rho_{|\mathbf{n}}, \mathrm{Coag}_2(\rho_{|n}, \tilde{\pi})}$, that is,

$$\tilde{q}_{\bar{\pi},\tilde{\pi}} = \lim_{t \to 0+} \frac{1}{t} \mathbb{P}(\mathcal{R}_{|\mathbf{n}}(t) = \mathrm{Coag}_2(\rho_{|\mathbf{n}}, \tilde{\pi}) \,|\, \mathcal{R}_{|\mathbf{n}}(0) = \rho_{|\mathbf{n}}).$$

One of our results assure that the rates transition $\tilde{q}_{\bar{\pi},\tilde{\pi}}$ can be characterized by a measure with values in $\tilde{\mathcal{P}}(\bar{\pi})$.

Proposition 1. Let $(\tilde{q}_{\bar{\pi},\tilde{\pi}} : \bar{\pi} \in \mathcal{P}_\infty, \tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}), \rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n), n \in \mathbb{N})$ be the family of jump rates of some snec $\mathcal{R}$. Then there exists a unique family $(\mu_{\bar{\pi}}, \bar{\pi} \in \mathcal{P}_\infty)$ where $\mu_{\bar{\pi}}$ is a measure on $\tilde{\mathcal{P}}(\bar{\pi})$ such that for any $\bar{\pi}$, $\mu_{\bar{\pi}}(\mathbf{0}_{\infty^2}) = 0$ and

$$\mu_{\bar{\pi}}(\mathcal{P}_{\infty,\tilde{\pi}}(\rho)) = \tilde{q}_{\bar{\pi},\tilde{\pi}}.$$

Finally, we construct a snec process with jump rates given by a Poisson random measure. In this aim we consider a sequence of independent random variables $\zeta = (\zeta_i)_{i \in \mathbb{N}}$, Bernoulli distributed with parameter $x \in (0,1)$. Let $(\xi^i_j)_{i,j \in \mathbb{N}}$ be an independent array of row wise independent Bernoulli random variables such that $\mathbb{P}(\xi^i_j = 1) = y$. Thus, given a nested partition π linked by $\bar{\pi}$, a conservative recipe $\tilde{\pi}$ for π is constructed, associating

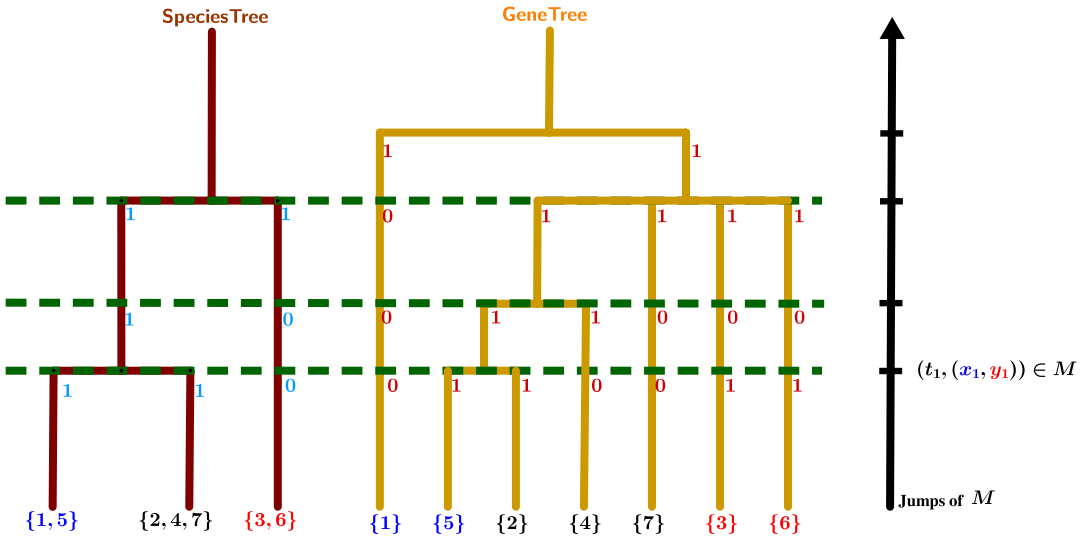

Figure 6: Poisson construction of the snec process $\mathcal{R}_{|\mathbf{n}}$.

the pair (ζ_i, ξ^i_j) to the j-th element in the block $\bar{\pi}_i$, so that species and genes merges when $(\zeta_i, \xi^i_j) = (1,1)$, see figure 6. Let P_{xy} be the law of $\tilde{\pi}$ and ν_{sg}, a sigma-finite measure on $[0,1]^2$ such that

$$\nu_{sg}(\{(0,0)\}) = 0 \quad \text{and} \quad \int_{[0,1]}\int_{[0,1]} (x^2 + xy^2)\nu_{sg}(dx,dy) < \infty. \tag{15}$$

Then

$$\varrho_{\nu_{sg}}(d\tilde{\pi}) = \int_{[0,1]}\int_{[0,1]} \nu_{sg}(dx,dy) P_{xy}(d\tilde{\pi}),$$

is a measure on $(\mathcal{P}'_\infty)^2$ satisfying the following condition

$$\varrho_{\nu_{sg}}(\tilde{\pi} \in \tilde{\mathcal{P}}(\rho) : \tilde{\pi}_{|\mathbf{n}} \neq \mathbf{0}_{|\rho_{|\mathbf{n}}|}) < \infty, \quad \text{for every } n \in \mathbb{N}.$$

We can now construct a snec process $\mathscr{R} = (\mathscr{R}(t) : t \geq 0)$ with transition given by the jumps of a Poisson point process M on $(0,\infty) \times (\mathcal{P}'_\infty)^2$ with intensity $dt \otimes \varrho_{\nu_{sg}}(d\tilde{\pi})$. More precisely, for every $n \in \mathbb{N}$ we consider M_n, the image of M by the map $(t,\tilde{\pi}) \to (t, \tilde{\pi}_{|\mathbf{n}})$, let $(0,\infty) \times ((\mathcal{P}'_\infty)^2 \setminus \mathbf{0}_{|\rho_{|\mathbf{n}}|})$ its atoms ranked in increasing order of their first coordinate. We set $\mathscr{R}_{|\mathbf{n}}(t) = \rho_{|\mathbf{n}}$ for $t \in [0, t_1)$, and recursively

$$\mathscr{R}^{\mathbf{n}}(t_i) = \mathrm{Coag}_2(\mathscr{R}^{\mathbf{n}}(t_i-), \tilde{\pi}^{(i)}(t_i)), \quad \text{for every } t \in [t_i, t_{i+1}).$$

We prove the consistency of this sequence, thus the process $\mathscr{R} = (\mathscr{R}(t), t \geq 0)$ defined for every $n \in \mathbb{N}$ by $\mathscr{R}_{|\mathbf{n}}(t) = \mathscr{R}^{\mathbf{n}}(t)$ exists and moreover is a snec, indeed we prove the following proposition.

Proposition 2. For every $t \geq 0$, the sequence of random bivariate partitions $(\mathscr{R}^{\mathbf{n}}(t), n \in \mathbb{N})$ is consistent. If we denote by $\mathscr{R}(t)$ the unique partition of $\mathcal{N}_\infty$ such that $\mathscr{R}_{|\mathbf{n}}(t) = \mathscr{R}^{\mathbf{n}}(t)$ for every $n \in \mathbb{N}$, then the process $\mathscr{R} = (\mathscr{R}(t), t \geq 0)$ is a snec with jump rate $\mu = \varrho_{\nu_{sg}}$, started from $\rho = (\bar{\pi}, \mathbf{0}_\infty)$. Notice that in this case the jump rate is not state-dependent.

Besides, we observe that thanks to the condition (15), the measures defined as follows

$$u_s(dx) = \int_{y\in[0,1]} \nu_{sg}(dx, dy), \qquad u_g(dy) = \int_{x\in[0,1]} x\nu_{sg}(dx, dy).$$

satisfies

$$u_g(0) = u_s(0) = 0 \quad \text{and} \quad \int_{[0,1]} x^2 u_s(dx), \int_{[0,1]} y^2 u_s(dy) < \infty.$$

Then by Lemma 4.5 of Bertoin (2006), there exists simple exchangeable coalescents Π^s and Π^g with coagulation measure u_s and u_g, respectively. Moreover, $\Pi^s \overset{\mathcal{L}}{=} \mathscr{R}^s$, hence we can think that the properties of the exchangeable coaleascent processes are related to the snec processes $\mathcal{R} = (\mathcal{R}^s, \mathcal{R}^g)$, in particular, if we say that $\mathcal{R}$ comes down from infinity, when

$$\mathbb{P}(|\mathcal{R}^g(t)| < \infty, \text{ for all } t > 0) = 1,$$

we prove the necessary and sufficient conditions in order to the snec before constructed comes down from infinity.

Proposition 3. The simple nested exchangeable coalescent $\mathscr{R}$ with jump rate $\mu = \varrho_{\nu_{sg}}$ CDI if and only if the simple exchangeable coalescent processes Π^s and Π^g, before defined, comes down from infinity.

The Poisson construction of $\mathscr{R}$ is an inspiration to prove that all snec processes can be constructed in the same way, and to get a similar result to Proposition 1, where the measure μ is independent of the present state. This is part of a work in progress, another directions are explained at the end of the Chapter 3.

Chapter 1

On branching process with rare neutral mutations

This is a joint work with Víctor Rivero.

1.1. Model description and main results

A Galton-Watson process models a population where at every generation each individual reproduces according to the same distribution, independently of the others and then dies. For background about branching processes we refer to Athreya and Ney (1972), Lambert (2008) and Li (2011). A number of variants, involving different types of conditioning and limit theorems, are core of branching processes theory. For instance, when the process dies with probability one, Yaglom (1947) proved that the distribution of the process conditioned to non-extinction exists, under some assumptions on the moments of the reproduction law. The proof was simplified and the moment assumptions removed by Joffe (1967) and Seneta and Vere-Jones (1966). More generally Lamperti and Ney (1968) introduced the Q-process.

As a further extension of the Galton-Watson model, Bertoin (2010) studied the so called Galton-Watson process with neutral mutations. This emerges assuming that the mutations modify the genotype of individuals but not the dynamic of the population, which is modeled by a standard Galton-Watson. Since mutations appear in the ancestral lines of the population, each individual begets children that do not necessarily inherit its genetic type (allele). In addition, we suppose that the population has infinity alleles, that is, each mutation event originates a different allele. We denote the size of a typical family by $\xi^{(+)} := \xi^{(c)} + \xi^{(m)}$, where $\xi^{(c)}$, $\xi^{(m)}$ are non-negative random variables which determine respectively, the number of clones and mutants children of a typical individual. We exclude the degenerate cases $\xi^{(c)} \equiv 0$ or $\xi^{(m)} \equiv 0$.

Asymptotic features are established on the genealogy of allelic sub-families in a Galton-Watson process with neutral mutations, by Bertoin (2010). In his development the ge-

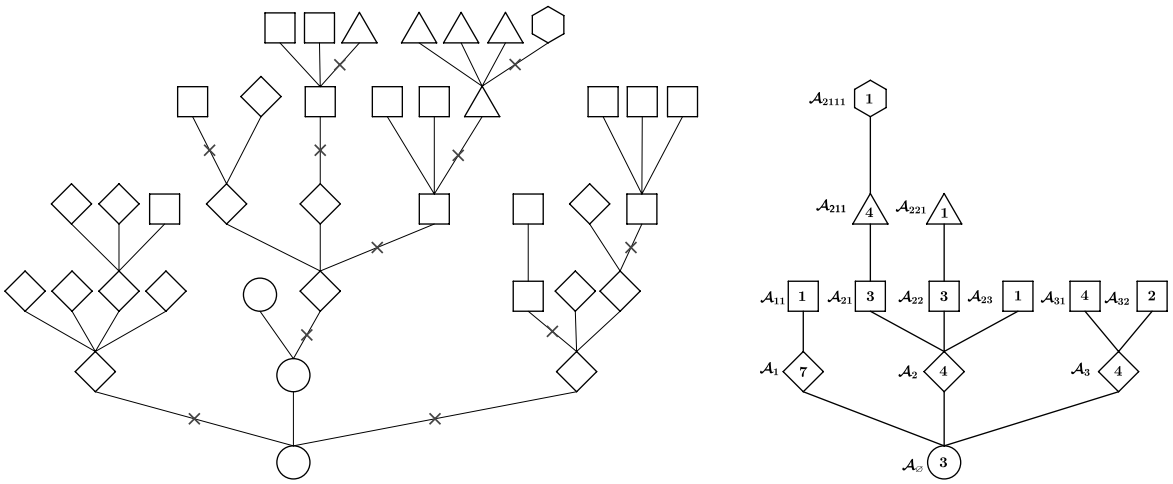

Figure 1.1: The tree on the left, illustrates a Galton Watson process with neutral mutation, the allele type of an individual is represented by the form of its vertex. On the right we have its tree of alleles, from this we can deduce for instance that $M_0 = 1, T_0 = 3, M_1 = 3, T_1 = 15$.

nealogy of the population is described by a planar rooted tree where the mutations are represented by marks in the edges between parents and mutant children. See Figure 1.1 (left). The vertices with n marks in their ancestral line are associated with the called *n-type individuals*. Those individuals with the n-th mark in the edge between them and their parents are known as *mutants of the n-type*. We denote by the T_n the total population of individuals of the n-th type and by M_n the total number of mutants of n-th type. By convention, the individual in the generation 0-th, the ancestors, are consider as mutants of the 0-th type, that is $M_0 = a$, $\mathbb{P}_a$-a.s.

It is well know that the branching property is the most basic property in the analysis of Galton-Watson processes. Then it is natural to expect a branching property for the Galton-Watson process with neutral mutations. Indeed it has the *general branching property*, which states that conditionally on the set of children of a stopping line, the families that those beget are independent copies of the initial tree. The concept of *stopping line* was introduced by Chauvin (1986), where the reader is referred for the formal definition. Roughly, a *line* is a family of edges such that every branch from the root contains at most one edge in that family. A stopping line is a random line such that the event "an edge is in the line", only depends on the marks found on their ancestral line. In particular, the set of edges connecting the mutants of the n type with their parents is a stopping line. Then for every n, each mutant of the n-type begets a sub-family which is independent of the others and has the same distribution as the original tree. Another important consequence of the general branching property is given in the following lemma.

Lemma 1.1 (Bertoin (2010), Lemma 1)**.** *Under $\mathbb{P}_a$, $\{M_n : n \in \mathbb{Z}_+\}$ is a Galton-Watson process with reproduction law $\mathbb{P}_1(M_1 \in \cdot)$. More generally, $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, is a*

Markov chain, with transition probabilities

$$\mathbb{P}_a(T_n = k, M_{n+1} = l \,|\, T_{n-1} = i, M_n = j) = \mathbb{P}_j(T_0 = k, M_1 = l), \quad i, j, k, l \in \mathbb{Z}_+ \text{ and } j \le k. \tag{1.1}$$

Remark 1.2. Since the mutants of the n-th type are also individuals of the n-th type, the transition probabilities in (1.1) are zero when $j > k$.

Let $\{P^n_{(i,j),(k,l)} : i, j, k, l \in \mathbb{Z}_+\}$ denotes the n-step transition probabilities of the process $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, that is

$$P^n_{(i,j),(k,l)} = \mathbb{P}_a(T_{r+n} = k, M_{r+n+1} = l | T_r = i, M_{r+1} = j), \quad i, j, k, l \in \mathbb{Z}_+, n \in \mathbb{N}. \tag{1.2}$$

Then $P^n_{(i,j),(k,l)}$ depends only on the mutants coordinate. Actually, it is not difficult to prove using induction, that the following identity holds

$$P^n_{(i,j),(k,l)} = \sum_{j_{n-1}=1}^{\infty} \mathbf{P}^{n-1}_{(j,j_{n-1})} \mathbb{P}_{j_{n-1}}(T_0 = k, M_1 = l), \tag{1.3}$$

where $j_0 = j$ and $\{\mathbf{P}^n_{(i,j)} : i, j \in \mathbb{Z}_+\}$ denotes the n-step transition probabilities of the Galton Watson process $\{M_n : n \in \mathbb{Z}_+\}$.

We now introduce the space of finite sequence of integers

$$\mathbb{U} := \bigcup_{k \in \mathbb{Z}_+} \mathbb{N}^k,$$

where $\mathbb{N} = \{1, 2, ...\}$ and $\mathbb{N}^0 = \{\varnothing\}$. We recall that this set gives us the label of the vertices in Ulam-Harris-Neveu tree. More precisely the root corresponds to $\{\varnothing\}$, one vertex at level $k > 0$ is $u = (u_1, ..., u_k)$ and $uj = (u_1, ..., u_k, j)$ represents its j-th children. The level of the vertex u is denoted by $|u|$. We shall consider the *tree of alleles* $\mathcal{A} := \{\mathcal{A}_u : u \in \mathbb{U}\}$ constructed recursively in Bertoin (2010). Define $\mathcal{A}_\varnothing = T_0$ and $\mathcal{A}_{uj}$ as the size of the j-th allelic sub-population of the type $|u| + 1$ which descend from the allelic sub-family indexed by the vertex u. In the case of ties, sub-families are ordered by convention uniformly at random. See Figure 1.1 (right). A further consequence of the general branching property is that, the tree of alleles enjoys a branching property. To provide a formal statement we first define the degree of the tree of alleles $\mathcal{A}$ at some vertex $u \in \mathbb{U}$ as

$$d_u := \max\{j \ge 1 : \mathcal{A}_{uj} > 0\},$$

where we agree that $\max \varnothing = 0$. The notation $(d_u \downarrow)$ means that the d_u-tuple has been rearranged in the decreasing order of the first coordinate, by convention, in the case of ties the coordinates are ranked uniformly at random.

Lemma 1.3 (Bertoin (2010), Lemma 2)**.** *For any integers $a \ge 1$ and $k \ge 0$, under $\mathbb{P}_a$ conditionally on $\{(\mathcal{A}_u, d_u) : |u| \le k\}$, for each vertex u at level k with $\mathcal{A}_u > 0$, the family of variables $\{(\mathcal{A}_{uj}, d_{uj}) : 1 \le j \le d_u\}$ are independent with distribution $(T_0, M_1)^{(d_u \downarrow)}$ under $\mathbb{P}_1$.*

It is important to observe the following identities

$$T_k = \sum_{|u|=k} \mathcal{A}_u \quad \text{and} \quad M_{k+1} = \sum_{|u|=k} d_u. \tag{1.4}$$

Hence given a population with neutral mutations, $\{(\mathcal{A}_u, d_u) : |u| \le k\}$ records the genealogy of allelic sub-families together with their sizes. Also, the size of their generations is a Galton-Watson process.

The first goal in this chapter is to construct the version of the chain $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, conditioned on non-extinction of mutants, hence we are interested in the situation where

$$T = \inf\{n \ge 1 : M_n = 0\} < \infty, \tag{1.5}$$

with a strictly positive probability. According to Corollary 1 of Bertoin (2010), this occurs when $\mathbb{E}(\xi^{(c)}) < 1$ and $\mathbb{E}(\xi^{(+)}) \le 1$. This implies that the Galton Watson process of mutants $\{M_n : n \in \mathbb{Z}_+\}$ is critical or subcritical, that is $m := \mathbb{E}_1(M_1) \le 1$.

We can now state our first theorem.

Theorem 1.4. *Let $a \in \mathbb{N}$ and $\{\mathcal{F}_n : n \in \mathbb{Z}_+\}$ be the natural filtration of the process $\{(T_{n-1}, M_n) : n \in \mathbb{N}\}$. Then, there exists a probability measure $\mathbb{P}_a^{\uparrow}$ that is locally absolutely continuous with respect to $\mathbb{P}_a$ with Radom-Nikodim martingale density*

$$Y_n = \frac{M_n q^{M_n - a}}{(f'(q))^n} \mathbf{1}_{\{n<T\}},$$

where $f(y) = \mathbb{E}_1(y^{M_1})$ and $q = \mathbb{P}_1(0 < T < \infty)$, that is

$$d\mathbb{P}_a^{\uparrow}|_{\mathcal{F}_n} = \frac{Y_n}{a}\, d\mathbb{P}_a|_{\mathcal{F}_n}, \quad n \in \mathbb{N}.$$

Furthermore, $\mathbb{P}_a^{\uparrow}$ is the law of a Markov chain $\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\}$ with n-step transition probabilities,

$$Q^n_{(i,j),(k,l)} = \frac{l q^{l-j}}{j(f'(q))^n} P^n_{(i,j),(k,l)}, \quad j, l \ge 1, \tag{1.6}$$

where $\{P^n_{(i,j),(k,l)} : i, j, k, l \in \mathbb{Z}_+\}$ denotes the n-step transition probabilities of the Markov chain $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$.

We next ensure that the process defined in the above theorem is distributed as

$$\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\},$$

conditionally on non-extinction of mutants.

Theorem 1.5. *Suppose that $\mathbb{E}(\xi^{(c)}) < 1$ and $\mathbb{E}(\xi^{(+)}) \le 1$.*

i) Let $a, n \in \mathbb{N}$ with n fixed. The conditional law of the Markov process $\{(T_k, M_{k+1}) : 0 \le k \le n-1\}$ under $\mathbb{P}_a(\cdot\,|n+k<T<\infty)$ converges, as $k \to \infty$, towards the probability measure $\mathbb{P}_a^{\uparrow}$, in the sense that for any n

$$\lim_{k\to\infty} \mathbb{P}_a(A|\, n+k<T<\infty) = \mathbb{P}_a^{\uparrow}(A), \quad \forall A \in \mathcal{F}_n. \tag{1.7}$$

ii) The Yaglom limit

$$\lim_{n\to\infty} \mathbb{P}(T_{n-1} = i, M_n = j|\, n<T<\infty),$$

exists and has a generating function $\widehat{\varphi}(x,y)$ such that for all $n \in \mathbb{N}$,

$$m^n \widehat{\varphi}(x,y) = \widehat{f}(\varphi_n(x,y)) - \widehat{f}(\varphi_n(x,0)), \quad x,y \in [0,1]. \tag{1.8}$$

The proof of the above results is based on classical methods and this due to the fact that the generating function of (T_n, M_{n+1}) can be written in terms of that of M_n, as it is established in Section 1.2.

We now turn to analyze the asymptotic behavior of the tree of alleles. In this purpose we will consider for ever $n \in \mathbb{N}$, a Galton-Watson process $\{Z_k^{(+n)} : k \in \mathbb{Z}_+\}$ such that the reproduction law

$$\pi_k^+ = \mathbb{P}(\xi^{(c)} + \xi^{(m)} = k), \quad k \in \mathbb{Z}_+,$$

is critical (with mean one) and has a finite variance σ^2. We assume that each child is a clone of her mother with probability $1-p(n)$ and a mutant with probability $p(n)$, so the joint law of $(\xi^{(c)}, \xi^{(m)})$, denoted by $\pi = \{\pi_{k,l} : k,l \in \mathbb{Z}_+\}$, that is,

$$\pi_{k,l} = \mathbb{P}(\xi^{(c)} = k, \xi^{(m)} = l), \quad k,l \in \mathbb{Z}_+, \tag{1.9}$$

satisfies

$$\pi_{k,l} = \pi_{k+l}^+ \binom{k+l}{k} (1-p(n))^k p(n)^l, \quad k,l \in \mathbb{Z}_+. \tag{1.10}$$

As usual, in the remainder of this chapter the relation $f \sim g$ refers to $\lim_{x\to\infty} f(x)/g(x) = 1$.

In the work of Bertoin (2010), it has been assumed that the number of ancestors and mutation rate respectively have the following behavior

$$a(n) \sim nx \quad \text{and} \quad p(n) \sim cn^{-1}, \quad \text{as } n \to \infty; \tag{1.11}$$

where c, x are some positive constants. In this setting it has been proved that

$$\mathcal{L}\left(\{(n^{-2}T_k, n^{-1}M_{k+1}) : k \in \mathbb{Z}_+\}, \mathbb{P}_{a(n)}^{p(n)}\right) \Longrightarrow \{(Z_{k+1}, cZ_{k+1}) : k \in \mathbb{Z}_+\}, \tag{1.12}$$

where $\{Z_k : k \in \mathbb{Z}_+\}$ is a (discrete time) continuous state branching process, in short CSBP, with reproduction measure

$$\nu(dy) = \frac{c}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{c^2 y}{2\sigma^2}\right) dy, \quad y > 0, \tag{1.13}$$

and initial population of size x/c. Here and all through the chapter the symbol $\Longrightarrow$ will denote the weak convergence of finite dimensional distributions.

From Lambert (2008), we know that the transition probabilities of any CSBP process $\{Y_k : k \in \mathbb{Z}_+\}$ with reproduction measure ϑ are characterized as follows:

$$\mathbb{E}(e^{-\lambda Y_{k+1}} | Y_k = y) = e^{-y\kappa(\lambda)}, \quad k \in \mathbb{Z}_+,\ \lambda, y \geq 0, \tag{1.14}$$

where κ is the cumulant of a subordinator with Lévy measure ϑ, so that we have the condition $\int_{(0,\infty)}(1 \wedge y)\vartheta(dy) < \infty$. In this work we will consider subordinators without drift, thus

$$\kappa(\lambda) = \int_{(0,\infty)}(1 - e^{-\lambda y})\vartheta(dy), \quad \lambda > 0. \tag{1.15}$$

Applying successively the property (1.14), we obtain

$$\mathbb{E}_x(e^{-s_1 Y_1 \cdots - s_k Y_k}) = e^{-x\kappa(l_{k-1}(s_1))}, \qquad s_i \geq 0,\ i = 1, 2, ..., k, \tag{1.16}$$

where l is defined by induction as follows: $l_0(s) = s$,

$$l_i(s_{n-i}) = s_{n-i} + \kappa(l_{i-1}(s_{n-i+1}))), \quad i \in \mathbb{N}. \tag{1.17}$$

Combining the convergence (1.12) and the identity (1.14), together with the Lévy-Itô decomposition of a subordinator, one could infer that conditionally on $n^{-2}T_k \sim y$ the sequence of the sizes of the sub-population carrying a same allele of the $(k+1)$-type and normalized by a factor n^{-2} should converge in distribution to the sequence of atoms of a Poisson random measure on $\mathbb{R}_+$ with intensity given in (1.13). Thus the limit of a sequence of tree of alleles can be defined as follows.

Definition 1.6 (Bertoin (2010), Definition 1)**.** *Fix $x > 0$ and ϑ a measure on $(0, \infty)$ with $\int_{(0,\infty)}(1 \wedge y)\vartheta(dy) < \infty$. A tree-indexed CSBP with reproduction measure ϑ and initial population of size x, is a process $\{\mathcal{Y}_u : u \in \mathbb{U}\}$ with values in $\mathbb{R}_+$ and indexed by the universal tree, whose distribution is characterized by induction on the levels as follows:*

i) $\mathcal{Y}_\varnothing = x$ a.s.;

ii) for every $k \in \mathbb{Z}_+$ conditionally on $\{\mathcal{Y}_v : v \in \mathbb{U}, |v| \leq k\}$, the sequences $\{\mathcal{Y}_{uj} : j \in \mathbb{N}\}$ for the vertices $u \in \mathbb{U}$ at generation $|u| = k$ are independent, and each sequence is distributed as the family of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $\mathcal{Y}_u\vartheta$, where the atoms are repeated according to their multiplicity, ranked in the decreasing order, and completed by an infinite sequence of 0 if the Poisson measure is finite.

Roughly, the tree-indexed CSBP is a process indexed according to the Ulam-Harris-Neveu tree such that the vertices $u \in \mathbb{U}$ at level $|u| = k > 1$ represent the sizes of the

sub-populations at generation k in the $CSBP$ $\{Y_k : k \in \mathbb{Z}_+\}$, which descent from the same parent at generation $k-1$.

It can be seen that the convergence (1.12) can be written as follows

$$\mathcal{L}\Big(\{(n^{-2}\mathcal{A}_u, n^{-1}d_u) : k \in \mathbb{Z}_+\}, \mathbb{P}^{p(n)}_{a(n)}\Big) \Longrightarrow \{(\mathcal{Z}_u, c\mathcal{Z}_u) : u \in \mathbb{U}\}, \tag{1.18}$$

where $\{\mathcal{Z}_u : u \in \mathbb{U}\}$ is a tree-indexed CSBP with reproduction ν given in (1.13) and random initial population of size x/c. We recall under some assumptions. It is d_u denotes the outer degree at the vertex $u \in \mathbb{U}$ in the tree of alleles. This latter convergence is the main result of Bertoin (2010). It uses an argument on convergence of triangular arrays, described in page 690 therein, that can be extended to a more general context, see e.g. the forthcoming Lemma 1.18.

One main goal of this chapter is to investigate the asymptotic behavior of the population in the same sense of Bertoin (2010) but on a complimentary class of reproduction laws. Instead of assuming that it has finite variance as in Bertoin's paper, we suppose that there exists $\alpha \in (1,2)$ such that,

$$\bar{\pi}^+(j) := \mathbb{P}(\xi^{(+)} > j) \in RV^{-\alpha}_{\infty}, \quad j \in \mathbb{Z}_+, \tag{1.19}$$

where $RV^{-\alpha}_{\infty}$ denotes the class of functions which are regularly varying at ∞ with index $-\alpha$, see Chapter I in Bingham et al. (1987) for background. Note that the case $\alpha \in (0,1)$ is excluded because it contradicts the assumption that π^+ is critical.

In order to extend the main result of Bertoin (2010) to our setting, we prove that there exists a regularly varying function r with index α such that

$$r(n)\mathbb{P}(\xi^+ > ny) \xrightarrow[n\to\infty]{} c_\alpha y^{-\alpha}, \quad \forall y > 0, \tag{1.20}$$

where $c_\alpha = 1/\Gamma(3-\alpha)$. The proof of this fact is given in Proposition 1.13. Moreover, the following behavior will be assumed instead of the hypothesis (1.11),

$$a(n) \sim xr(n)p(n) \quad \text{and} \quad p(n) \sim cn^{-1}, \quad \text{as } n \to \infty. \tag{1.21}$$

The result below extends to our setting the main result in Bertoin (2010).

Theorem 1.7. *If (1.19) and (1.21) holds, then the following convergence holds in the sense of finite dimensional distributions*

$$\mathcal{L}\Big(\{((r(n))^{-1}\mathcal{A}_u, (r(n)p(n))^{-1}d_u) : u \in \mathbb{U}\}, \mathbb{P}^{p(n)}_{a(n)}\Big) \Longrightarrow \{(\mathcal{Z}^{1/\alpha}_u, \mathcal{Z}^{1/\alpha}_u) : u \in \mathbb{U}\},$$

where $\{\mathcal{Z}^{1/\alpha}_u : u \in \mathbb{U}\}$ *is a tree-indexed CSBP with reproduction measure*

$$\nu^\alpha(dy) = c'_\alpha y^{-1-1/\alpha}dy, \quad y > 0,\ \alpha \in (1,2), \tag{1.22}$$

where $c'_\alpha = \alpha^{-1}/\Gamma(1-\alpha^{-1})$.

Finally, we establish the convergence of the finite dimensional distributions of the rescaled chain $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, conditioned to non-extinction of mutants, towards a continuous state branching process with immigration in discrete time.

Theorem 1.8. *If the reproduction law is critical, there exist sequences $b_1(n)$ and $b_2(n)$ such that the following joint convergence in the sense of finite dimensional distributions holds:*

$$\mathcal{L}\left(\{(b_1(n)T_{k-1}, b_2(n)M_k) : k \in \mathbb{Z}_+\}, \mathbb{P}_{a(n)}^{p(n)\uparrow}\right) \Longrightarrow \{(Y_k, \beta Y_k) : k \in \mathbb{Z}_+\},$$

where $\{Y_k : k \in \mathbb{Z}_+\}$ is a CSBP with immigration, which is characterized by the following conditions:

i) if the reproduction law has finite variance σ^2 and (1.11) holds, then its reproduction measure is given by (1.13) and the immigration measure is $z\nu(dz)$ and $\beta = c$; moreover $b_1(n) = n^{-2}$ and $b_2(n) = n^{-1}$;

ii) if the assumptions (1.19) and (1.21) hold, the reproduction measure is $\nu^\alpha(dz)$ as defined in (1.22), the immigration measure is $z\nu^\alpha(dz)$ and $\beta = 1$; the normalizing constants are given by $b_1(n) = (r(n)p(n))^{-1}$ and $b_2(n) = (r(n))^{-1}$.

1.2. Preliminaries

In this section we obtain some useful formulas for the generating function of (T_n, M_{n+1}), denoted for $n \in \mathbb{Z}_+$ by

$$\varphi_n(x, y) := \mathbb{E}_1(x^{T_{n-1}} y^{M_n}), \quad x, y \in [0, 1],$$

where for notational convenience $\varphi_1(x, y) := \varphi(x, y)$. Observe that the generating function of M_n is

$$f_n(y) := \varphi_n(1, y), \quad y \in [0, 1], \tag{1.23}$$

and as before we denote $f_1(y) =: f(y)$.

According with the classical theory of branching processes, the extinction probability of the Galton-Watson process $\{M_n : n \in \mathbb{Z}_+\}$, that we denote by q, is the smallest root of $f(y) = y$, which is less or equal than one depending on whether the mean of the reproduction law, $m := \mathbb{E}_1(M_1)$ is > 1 or ≤ 1, respectively. In order to avoid trivial cases, we assume throughout that

H1) $\mathbb{P}(M_1 = 1) > 0$,

H2) $\mathbb{P}(M_1 = 0) + \mathbb{P}(M_1 = 1) < 1$, and $\mathbb{P}(M_1 = j) \neq 1$, for any j.

We also know that the n-step transition probabilities $\{\mathbf{P}^n_{(i,j)} : i,j \in \mathbb{Z}_+\}$ of the Galton-Watson process $\{M_n : n \in \mathbb{Z}_+\}$ satisfy

$$\sum_{j=0}^{\infty} \mathbf{P}^n_{(i,j)} y^j = \left(f_n(y)\right)^i, \quad i \geq 1. \tag{1.24}$$

For a Galton-Watson process with neutral mutations, let g be the generating function of the reproduction law of a typical individual, that is

$$g(x,y) := \mathbb{E}(x^{\xi^{(c)}} y^{\xi^{(m)}}), \quad x,y \in [0,1].$$

Proposition 1 of Bertoin (2010) ensures that the law of (T_0, M_1) can be obtained applying the Lagrange inversion formula to the equation

$$\varphi(x,y) = xg(\varphi(x,y),y), \quad x,y \in [0,1]. \tag{1.25}$$

Thanks to this latter equality it is possible to deduce that $T \leq \infty$ if $\mathbb{E}(\xi^{(c)}) < 1$ and $\mathbb{E}(\xi^{(+)}) \leq 1$. Similarly we have that $\mathbb{E}(\xi^{(+)2}) < \infty$ if and only if $\mathbb{E}(M_1^2) < \infty$. This equality is also a key tool to establish the following identity

$$\mathbb{P}_a(T_0 = k, M_1 = l) = \frac{a}{k} \pi^{\star k}_{k-a,l}, \quad k \geq a \geq 1 \text{ and } l \geq 0, \tag{1.26}$$

where $\pi^{\star k}$ denotes the k-th convolution of π, as defined in (1.9). Using the previous display we can write the hypothesis (H1) and (H2) in terms of the reproduction distribution of a typical individual.

Moreover, letting $P^n_{(i,j),(k,l)}$ be the n-step transition probabilities of the process $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, for this process we have a equality similar to (1.24), i.e.

$$\sum_{k,l=0}^{\infty} P^n_{(i,j),(k,l)} x^k y^l = \left(\varphi_n(x,y)\right)^j, \quad i,j \geq 1. \tag{1.27}$$

We get the latter equality by induction. Namely, we apply the Chapman-Kolmogorov equation to express the $(n+1)$-step transition probabilities in terms of the transitions in one step and use (1.1).

A simple but key relation for our analysis is

$$\varphi_n(x,y) = f_{n-1}(\varphi(x,y)), \quad x,y \in [0,1]. \tag{1.28}$$

Due to (1.23), the proof of the latter identity is equivalent to establish the following equality

$$\varphi_n(x,y) = \varphi_{n-1}(1, \varphi(x,y)), \quad x,y \in [0,1], \tag{1.29}$$

which follows from the standard calculations:

$$\begin{aligned}
\varphi_n(x,y) &= \mathbb{E}_1(\mathbb{E}_1(x^{T_{n-1}}y^{M_n} \mid T_{n-2}, M_{n-1})) \\
&= \sum_{i,j=0}^{\infty} \mathbb{P}_1(T_{n-2}=i, M_{n-1}=j) \sum_{k=j}^{\infty}\sum_{l=0}^{\infty} x^k y^l \mathbb{P}_1(T_{n-1}=k, M_n=l \mid T_{n-2}=i, M_{n-1}=j) \\
&= \sum_{i,j=0}^{\infty} \mathbb{P}_1(T_{n-2}=i, M_{n-1}=j) \sum_{k=j}^{\infty}\sum_{l=0}^{\infty} x^k y^l \mathbb{P}_j(T_0=k, M_1=l) \\
&= \sum_{i,j=0}^{\infty} \mathbb{P}_1(T_{n-2}=i, M_{n-1}=j)(\varphi(x,y))^j \\
&= \varphi_{n-1}(1, \varphi(x,y));
\end{aligned}$$

where we used the Markov property of $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, Lemma 1.1 and the branching property.

1.3. The process conditioned to non-extinction

This section is devoted to study the process $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$ conditioned to non-extinction.

1.3.1. Construction

Here our aim is to prove Theorem 1.4, which ensures the existence of the law of a Markovian process that we understand as the chain $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, conditioned to non-extinction of mutants in the population.

Proof of Theorem 1.4. An application of the Monotone Convergence Theorem, along with an elementary computation, shows that

$$\frac{d}{ds}\mathbb{E}_a(s^{M_n})\big|_{s=q} = \mathbb{E}_a\left(M_n q^{M_n-1}\right).$$

Moreover, the following identity is deduced from the branching property of the Galton-Watson process $\{M_n : n \in \mathbb{Z}_+\}$ and the properties of its generating function

$$\frac{d}{ds}\mathbb{E}_a(s^{M_n})\big|_{s=q} = aq^{a-1}f_n'(q).$$

The latter and former identities imply in turn that

$$\mathbb{E}_a\left(M_n q^{M_n-1}\right) = aq^{a-1}f_n'(q).$$

Then by the Markov property,

$$\mathbb{E}_a\left(M_{n+k}q^{M_{n+k}-1}|\mathcal{F}_n\right) = M_n q^{M_n-1}f_k'(q).$$

Combining the latter with the fact that $f_k'(q) = [f'(q)]^k$ (see Athreya and Ney (1972), Lemma 3.3), we have that

$$Y_n = \frac{M_n q^{M_n - a}}{(f'(q))^n}, \quad n \geq 0,$$

is a martingale. Now from the theory of h-transforms (see Chapter 11 in Chung and Walsh (2005)), there exists a Markovian process, that we denote by $\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\}$ whose law satisfies

$$\mathbb{P}_a^{\uparrow}(T_0^{\uparrow} = i_0, M_1^{\uparrow} = j_1, \cdots, T_{n-1}^{\uparrow} = i_{n-1}, M_n^{\uparrow} = j_n) := \mathbb{P}_a(A_n)\frac{j_n q^{j_n - a}}{a(f'(q))^n}, \tag{1.30}$$

where for every $n \in \mathbb{N}$

$$A_n = \{T_0 = i_0, M_1 = j_1, \cdots, T_{n-1} = i_{n-1}, M_n = j_n\}, \quad i_0, j_1, ..., i_{n-1}, j_n \in \mathbb{N}. \tag{1.31}$$

■

1.3.2. Conditional laws

Here we will prove Theorem 1.5. The building block in this aim will be the generating function of $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$, hence some of the results given in Section 1.2 will be necessary.

Proof of Theorem 1.5. i) Let A_n be an event of the form given in (1.31). It thus follows from the Markov and branching properties of $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$ that

$$\mathbb{P}_a(n + k < T < \infty) = \mathbb{E}_a(\mathbf{1}_{\{M_{n+k} > 0\}} q^{M_{n+k}}),$$

where as before $q = \mathbb{P}_1(0 < T < \infty)$. We also have that

$$\mathbb{P}_a(A_n, n + k < T < \infty) = \mathbb{E}_a(\mathbf{1}_{A_n} \mathbb{E}_{j_n}(\mathbf{1}_{\{M_k > 0\}} q^{M_k})).$$

Then using (1.24), we get

$$\mathbb{P}_a(A_n \,|n + k < T < \infty) = \mathbb{P}_a(A_n)\frac{\sum_{j=1}^{\infty} \mathbf{P}_{(j_n, j)}^k q^j}{\sum_{j=1}^{\infty} \mathbf{P}_{(1,j)}^{n+k} q^j},$$

we recall that $\mathbf{P}_{(i,j)}^n$ denotes the n-step transition probabilities of $\{M_n : n \in \mathbb{Z}_+\}$. Besides, Theorem 7.4 of Athreya and Ney (1972) establishes that the following limit holds

$$\lim_{k \to \infty} \frac{\mathbf{P}_{(i_1, j)}^{n+k}}{\mathbf{P}_{(i_2, j)}^k} = i_1 i_2^{-1} (f'(q))^k q^{i_1 - i_2}.$$

Finally, thanks to the hypothesis (H2) we can use the previous identiy to obtain

$$\lim_{k\to\infty} \mathbb{P}_a(A_n | n+k<T<\infty) = \mathbb{P}_a(A_n)\frac{j_n q^{j_n-a}}{a(f'(q))^n}, \quad a\in\mathbb{N},$$

which finishes the first part of the proof.

ii) We will first ensure the convergence of the generating function. Since $\{M_n>0\}$ on the event $\{n<T<\infty\}$, we deduce that for all $x,y\in[0,1]$:

$$\begin{aligned}\widehat{\varphi}_n(x,y) &:= \mathbb{E}_1(x^{T_{n-1}}y^{M_n} \mid n<T<\infty)\\ &= \frac{\varphi_n(x,y)-\varphi_n(x,0)}{1-\mathbb{P}(M_n=0)}.\end{aligned}$$

From the identity (1.28) and the fact that $f_n(0)=\mathbb{P}(M_n=0)$, the previous expression can be written as follows

$$\widehat{\varphi}_n(x,y) = \frac{1-f_{n-1}(0)}{1-f_n(0)}\left(\frac{f_{n-1}(\varphi(x,y))-f_{n-1}(0)}{1-f_{n-1}(0)} - \frac{f_{n-1}(\varphi(x,0))-f_{n-1}(0)}{1-f_{n-1}(0)}\right). \quad (1.32)$$

We now take $u=f_{n-1}(0)$ and use $m=f'(1)$ to obtain

$$\lim_{n\to\infty}\frac{1-f_{n-1}(0)}{1-f_n(0)} = \lim_{u\to 1}\frac{1-u}{1-f(u)} = \frac{1}{m}.$$

Observe that the function

$$n \longmapsto \frac{1-f_{n-1}(s)}{1-f_{n-1}(0)},$$

is decreasing for each s. Therefore, as n tends to infinity the expression

$$\frac{f_{n-1}(s)-f_{n-1}(0)}{1-f_{n-1}(0)} = 1-\frac{1-f_{n-1}(s)}{1-f_{n-1}(0)},$$

has a limit, say $1-\widehat{f}(s)$. According to Theorem 1.8.1 in Athreya and Ney (1972), we know that the generating function, $\widehat{f}(s)$, of the Yaglom distribution of $\{M_n : n\in\mathbb{Z}_+\}$ given by the following limit

$$\rho_k = \lim_{n\to\infty}\mathbb{P}(M_n=k \,|\, n<T<\infty), \quad \text{for all } k\in\mathbb{N}.$$

The above cited theorem also ensures that

$$1-\widehat{f}(f(s)) = m(1-\widehat{f}(s)), \quad s\in[0,1]. \quad (1.33)$$

Putting all the pieces together in (1.32), we obtain

$$\widehat{\varphi}(x,y) := \lim_{n\to\infty}\widehat{\varphi}_n(x,y) = \frac{\widehat{f}(\varphi(x,y))-\widehat{f}(\varphi(x,0))}{m}.$$

We now prove by induction (1.8). If $n = 1$, it is the just proved equality. Then suppose (1.8) holds for $n = k$. In order to get the identity for $n = k + 1$ note that by the induction hypothesis

$$m^{k+1}\widehat{\varphi}(x,y) = m\left[1 - \widehat{f}(\varphi_k(x,0))\right] - m\left[1 - \widehat{f}(\varphi_k(x,y))\right], \quad x,y \in [0,1].$$

From the above we deduce the claim using first (1.33) and then (1.28).

■

Remark 1.9. In the previous proof we established the existence of a Yaglom limit when $m \leq 1$, however as in the classical case similar arguments can be used to show existence of a Yaglom limit in the supercritical case.

1.3.3. Interpretation

Motivated by the interpretation of a Galton-Watson process conditioned to non-extinction given in Lambert (2007), our objective in the present subsection is to describe the chain $\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\}$ in terms of immigration of mutants. We start calculating the generating function of the n-step transition probabilities of this process.

Proposition 1.10. *Letting $Q^n_{(i,j),(k,l)}$ be the n-step transition probabilities of the process*

$$\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\},$$

we have

$$\sum_{k,l=1}^{\infty} Q^n_{(i,j),(k,l)} x^k y^l = \frac{yq^{1-j}}{[f'(q)]^n}[\varphi_n(x,qy)]^{j-1}\frac{\partial}{\partial y}\varphi(x,qy)\prod_{i=1}^{n-1} f'(\varphi_i(x,qy)), \quad x,y \in [0,1]. \tag{1.34}$$

Proof. Let us start by pointing out the following formula

$$\sum_{k,l=1}^{\infty} Q^n_{(i,j),(k,l)} x^k y^l = \frac{yq^{1-j}}{[f'(q)]^n}\left[\varphi_n^{j-1}(x,u)\frac{\partial}{\partial u}\varphi_n(x,u)\right]_{u=qy}, \quad x,y \leq 1.$$

This is a consequence of the fact that for each $n \in \mathbb{Z}_+$ the generating function of (T_n, M_{n+1}) is infinitely differentiable in $(x,y) \in [0,1]^2$, the identity (1.27) and some elementary computation. The claimed formula is obtained by applying repeatedly (1.28) and the recursion $f_n(y) = f(f_{n-1}(y))$. ■

Taking $x = 1$ in (1.34) and recalling the fact that the transition probabilities of $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$ depend only on the second coordinate, we can identify a Galton-Watson process with immigration (see Kawazu and Watanabe (1971) for background).

Corollary 1.11. *If $\{M_n : n \in \mathbb{Z}_+\}$ is critical or subcritical, then $\{M_n^{\uparrow} - 1 : n \in \mathbb{Z}_+\}$ is a Galton-Watson process with immigration $[f, \frac{f'}{m}]$.*

Note that $\{M_n^{\uparrow} : n \in \mathbb{Z}_+\}$ is the Q-process associated to $\{M_n : n \in \mathbb{Z}_+\}$ (see for instance Athreya and Ney (1972) or Lambert (2007)). The following corollary is analogous to Proposition 1 in Bertoin (2010).

Corollary 1.12. *If $\{M_n : n \in \mathbb{Z}_+\}$ is critical or subcritical, then the generating function of $(T_0^{\uparrow}, M_1^{\uparrow})$ is determined by the equation*

$$\mathbb{E}_1(x^{T_0^{\uparrow}} y^{M_1^{\uparrow}}) = \frac{xy}{m} \frac{\partial}{\partial y} g(\varphi(x,y), y), \quad x, y \in [0,1].$$

Moreover, the distribution of $(T_0^{\uparrow}, M_1^{\uparrow})$ is given by

$$\mathbb{P}_a^{\uparrow}(T_0^{\uparrow} = k, M_1^{\uparrow} = l) = \frac{l}{mk} \pi_{k-a,l}^{*k}, \quad k \geq a \geq 1 \text{ and } l \geq 0,$$

we recall that π^{*k} denotes the k-th convolution of the law π, defined in (1.9).

Proof. Taking $n = 1$ in the equality (1.34),

$$\mathbb{E}_1(x^{T_0^{\uparrow}} y^{M_1^{\uparrow}}) = \frac{y}{m} \frac{\partial}{\partial y} \varphi(x,y).$$

Then the first identity is obtained using the identity (1.25). To get the second one, recall the definition of $\mathbb{P}^{\uparrow}$ given in (1.30) and (1.26). ∎

We can now give an interpretation to the process $\{(T_n^{\uparrow}, M_{n+1}^{\uparrow}) : n \in \mathbb{Z}_+\}$, in terms of a tree of alleles with immigration $\mathcal{A}^{\uparrow} = \{\mathcal{A}_u^{\uparrow} : u \in \mathbb{U}\}$. This tree will provide a description of the genealogical structure in a population conditioned to non extinction.

We start defining $\mathcal{A}_{\varnothing}^{\uparrow} = T_0^{\uparrow}$ that is, the total number of individuals without mutations into the population, then according to a distribution with generating function f'/m, a random number of individuals of the same genetic type arrive. We enumerate the $M_1^{\uparrow}$ allelic sub-populations of the first type beget by $T_0^{\uparrow}$ in decreasing order, with the convention that in the case of ties, sub-populations of the same size are ranked uniformly at random. Using Corollary 1.11 we choose uniformly at random one of the first type sub-families in the tree of alleles, removing it and replace it by a population of size $T_0^{\uparrow}$ which begets allelic subpopulation according to $M_1^{\uparrow}$, where $(T_0^{\uparrow}, M_1^{\uparrow})$ is given by Corollary 1.12. We continue with the construction by iteration, $\mathcal{A}_{uj}^{\uparrow}$ is the size of the j-th sub-population allelic of type $|u| + 1$ which descend from the allelic sub-family indexed by the vertex u. Then we choose one of the sub-families of type $|u| + 1$ to replace it by one of size $T_0^{\uparrow}$, which begets allelic subpopulation according to $M_1^{\uparrow}$.

1.4. Asymptotic behavior: the α-stable case

Our goal in this section is to prove Theorem 1.7. For that end, until further notice we will consider a sequence of Galton-Watson processes $\{Z_k^{(+n)} : k \in \mathbb{Z}_+\}$ such that the reproduction law $\pi^{(+)}$ is critical with heavy tails; the mutations appear in the population according to (1.10); and the mutation rate, together with the ancestors behavior is given by (1.21).

1.4.1. Approximations for the reproduction law

We start by describing the normalizing sequence appearing in Theorem 1.7.

Lemma 1.13. *If the condition (1.19) holds, then there exists a sequence $\{r(n) : n \geq 0\}$ that is regularily varying at infinity with index α such that*

$$r(n)\pi^+(ndy) \xrightarrow[n\to\infty]{} c_\alpha \frac{dy}{y^{1+\alpha}},$$

in the sense of vague convergence on $(0,\infty)$, where $c_\alpha = 1/\Gamma(3-\alpha)$. In particular

$$\exp\left\{-t\int_{(0,\infty)}(1-e^{-\lambda y}-\lambda y)r(n)\pi^+(ndy)\right\} \xrightarrow[n\to\infty]{} e^{-t\lambda^\alpha}.$$

The proof is an elementary application of standard results from the theory of Regular Variation (see e.g. Bingham et al. (1987) for background). We include a proof in Appendix B for sake of completeness.

In order to link the asymptotic behaviour of the reproduction law of a typical individual with that of the joint distribution of clones and mutants, we first link their Laplace transform. Although in the present setting we use some ideas of the standard Tauberian-Abelian Theorem, we remark that it is not straightforward application of this theorem because we consider sequences of measures indexed by the positive integers changing, unlike to the standard case, where only the normalizing constants change.

Lemma 1.14. *For every positive integer n, let ϕ_n be the Laplace transform of the random vector $\xi^{(n)} = (\xi^{(cn)}, \xi^{(mn)})$ under the measure $\mathbb{P}_1^{p(n)}$. Assume that $\{\lambda(n) : n \in \mathbb{Z}_+\}$ is a positive sequence such that $\lambda(n) \to 0$, as $n \to \infty$. Then*

$$\phi_n(\lambda(n),\theta) \sim \phi^+\left((1-p(n))(1-e^{-\lambda(n)}) + p(n)(1-e^{-\theta})\right), \qquad \text{with } n\to\infty, \forall\theta\geq 0, \quad (1.35)$$

where ϕ^+ is the Laplace transform of $\xi^{(+)}$. In particular ϕ_n^m, respectively ϕ_n^c, the Laplace transform of $\xi^{(mn)}$, respectively $\xi^{(cn)}$, satisfies

$$\phi_n^m(\theta) \sim \phi^+(\,p(n)(1-e^{-\theta})\,), \quad \forall\,\theta\geq 0, \quad (1.36)$$

$$\phi_n^c(\lambda(n)) \sim \phi^+(\,(1-p(n))(1-e^{-\lambda(n)})\,), \qquad \text{as } n\to\infty. \quad (1.37)$$

Proof. According to (1.10), conditionally to $\xi^{(+)} = k$ the distribution of $\xi^{(m)}$ is Binomial with parameter (k, p). This fact implies the following equality in law

$$(\xi^{(c)}, \xi^{(m)}) \overset{\mathcal{L}}{=} \sum_{i=1}^{\xi^{(+)}} (\mathbf{1}_{\{U_i > p\}}, \mathbf{1}_{\{U_i \leq p\}}), \tag{1.38}$$

where $\{U_i : i \in \mathbb{N}\}$ are independent random variables with common distribution that of an uniform random variable in $(0, 1)$. Therefore,

$$\begin{aligned}\phi_n(\lambda(n), \theta) &= \sum_{k=0}^{\infty} \mathbb{P}_1^{p(n)}(\xi^{(+)} = k) \left[(1 - p(n))e^{-\lambda(n)} + p(n)e^{-\theta}\right]^k \\ &= \phi^+(-\log(1 - (1 - (1 - p(n))e^{-\lambda(n)} - p(n)e^{-\theta}))).\end{aligned}$$

We conclude the proof using (1.11) and the elementary asymptotic estimate

$$\frac{\log(1 - y)}{y} \xrightarrow[y \to 0]{} -1. \tag{1.39}$$

■

In the same way it is possible to establish the following estimate.

Corollary 1.15. *For every positive integer n, let ψ_n be the characteristic function of $\xi^{(n)} = (\xi^{(cn)}, \xi^{(mn)})$ under the measure $\mathbb{P}_1^{p(n)}$. Then*

$$\psi_n(\lambda(n), \theta) \sim \phi^+(\,(1 - p(n))(1 - e^{i\lambda(n)}) + p(n)(1 - e^{i\theta})\,), \qquad \text{as } \lambda(n) \xrightarrow[n \to \infty]{} 0, \forall\, \theta \geq 0. \tag{1.40}$$

In particular ψ_n^m, respectively ψ_n^c, the characteristic function of $\xi^{(mn)}$, respectively of $\xi^{(cn)}$, satisfies

$$\begin{aligned}\psi_n^m(\theta) &\sim \phi^+(\,p(n)(1 - e^{i\theta})\,), \\ \psi_n^c(\lambda(n)) &\sim \phi^+(\,(1 - p(n))(1 - e^{i\lambda(n)})\,),\end{aligned}$$

as $n \to \infty$, $\lambda(n) \to 0$ and for all $\theta \geq 0$.

Proof. Similarly to previous lemma, using (1.38) we have

$$\begin{aligned}\psi_n(\lambda(n), \theta) &= \sum_{k=0}^{\infty} \mathbb{P}_1^{p(n)}(\xi^{(+)} = k) \left[(1 - p(n))e^{\mathrm{i}\lambda(n)} + p(n)e^{\mathrm{i}\theta}\right]^k \\ &= \sum_{k=0}^{\infty} \mathbb{P}_1^{p(n)}(\xi^{(+)} = k) \exp\{k \log(1 - ((1 - p(n))(1 - e^{\mathrm{i}\lambda(n)}) - p(n)(1 - e^{\mathrm{i}\theta})))\}.\end{aligned}$$

To conclude we apply the asymptotic estimate (1.39). ■

We can now use the results above to give an estimate for the reproduction measure.

Proposition 1.16. *For every positive integer n, let $\pi^{(cn)}$, $\pi^{(mn)}$ be the reproduction laws of $\xi^{(cn)}$ and $\xi^{(mn)}$, respectively. Assume that $\{y(n) : n \geq 0\}$ is any sequence such that $y(n) \to \infty$ as $n \to \infty$. In the regime (1.19), the asymptotic behavior of the tail distribution of $\xi^{(\cdot)}$ is given by*

$$\bar{\pi}^{(\cdot)}(y(n)) \sim c_\alpha \bar{\pi}^{+}\left(y(n)/\mathbb{E}(\xi^{(\cdot)})\right), \quad \text{as } n \to \infty, \tag{1.41}$$

where $(\cdot) = cn, mn$ and $c_\alpha = 1/\Gamma(3-\alpha)$.

The proof of this proposition is deferred to the Appendix C because it use some elements of the proof of Lemma 1.13, which is included in Appendix B.

1.4.2. Proof of Theorem 1.7

The aim of this section is to prove Theorem 1.7. For that end we required two auxiliares results that we will next state.

Let $n \in \mathbb{N}$ fixed. According with the construction of alleles trees, given a vertex u, at level $k \geq 1$, in $\{\mathcal{A}_u : u \in \mathbb{U}\}$, a vertex uj represents the size of the j-th allelic sub-populations of type $k+1$ begot by u, and this holds for every $j \in \mathbb{N}$. Thus the labels of the vertices at level $k+1$ determine the variable T_{k+1}. Moreover, for all $k \in \mathbb{N}$, the total number of vertices at level k correspond to M_k. Hence, a first step to establish the convergence in Theorem 1.7 will be describe the scaling limit of the process $\{(T_k, M_{k+1}) : k \in \mathbb{Z}_+\}$, towards a CSBP $\{Z_k^{1/\alpha} : k \in \mathbb{Z}_+\}$. That is the purpose of the Proposition 1.17 below, whose proof its deferred to Section 1.4.3.

Proposition 1.17. *Assuming (1.19) and (1.21), we have*

$$\mathcal{L}\left(\left\{\left(\frac{T_k}{r(n)}, \frac{M_{k+1}}{r(n)p(n)}\right) : k \in \mathbb{Z}_+\right\}, \mathbb{P}_{a(n)}^{p(n)}\right) \Longrightarrow \{(Z_{k+1}^{1/\alpha}, Z_{k+1}^{1/\alpha}) : k \in \mathbb{Z}_+\}, \tag{1.42}$$

where $\{Z_k^{1/\alpha} : k \in \mathbb{Z}_+\}$ is a CSBP process with reproduction measure ν^α given in Theorem 1.7.

To obtain from this result the convergence claimed in Theorem 1.7 we will need the Lemma 1.18 below, whose proof is given in Section 1.4.4.

Lemma 1.18. *Let $b(n)$ be a sequence of integers such that $b(n) \sim br(n)p(n)$ for some $b > 0$.*

i) For every $n \in \mathbb{N}$, let $\{\chi_j^{(n)} : 1 \leq j \leq b(n)\}$ be a sequence of independent identically distributed random variables with distribution $\left(\frac{T_0}{r(n)}, \frac{M_1}{r(n)p(n)}\right)$. Defining for every

$n \in \mathbb{N}$, $\gamma_j^{(n)} := \delta_{\chi_j^{(n)}}$ and $\gamma_n = \sum_{j=1}^{\infty} \gamma_j^{(n)}$, the following weak convergence of measures holds

$$\gamma_n \xrightarrow[n\to\infty]{} \gamma, \tag{1.43}$$

where γ is a Poisson point measure with intesity $b\eta$, with η the image of the measure ν^α (given in Theorem 1.7) by the action of the map $x \mapsto (x, x)$.

ii) We have the following convergence, under the measure $\mathbb{P}_1^{p(n)}$

$$\left(\frac{T_0}{r(n)}, \frac{M_1}{r(n)p(n)}\right)^{(b(n)^{\downarrow})} \Longrightarrow (\mathbf{a}_1, \mathbf{a}_2, ...), \tag{1.44}$$

where for all $k \in \mathbb{N}$, $\mathbf{a}_k = (a_k, a_k)$ with $\{a_k : k \in \mathbb{N}\}$, the atoms of a Poisson random measure on $(0, \infty)$ with intensity $b\nu^\alpha$ ranked in decreasing order; the measure ν^α is given in Theorem 1.7.

Taking for granted the above results we can provide a proof to Theorem 1.7

Proof of Theorem 1.7

We will establish

$$\mathcal{L}\left(\left(\left(\left(\frac{\mathcal{A}_u}{r(n)}, \frac{d_u}{r(n)p(n)}\right) : |u| \le k\right) : k \in \mathbb{Z}_+\right), \mathbb{P}_{a(n)}^{p(n)}\right) \Longrightarrow \left(\left(\left(\mathcal{Z}_u^{1/\alpha}, \mathcal{Z}_u^{1/\alpha}\right) : |u| \le k\right) : k \in \mathbb{Z}_+\right). \tag{1.45}$$

Actually by the monotone class theorem, it is enough to show for non-negative measurable continuous functions $f_1, ..., f_k$

$$\mathbb{E}_{a(n)}^{p(n)}\left[\prod_{i=1}^{k} f_i((r(n))^{-1}\mathcal{A}_u, (r(n)p(n))^{-1}d_u : |u| \le i)\right] \xrightarrow[n\to\infty]{} \mathbb{E}^{\mathbb{Q}_x}\left[\prod_{i=1}^{k} f_i(\mathcal{Z}_u^{1/\alpha}, \mathcal{Z}_u^{1/\alpha} : |u| \le i)\right],$$

where $\mathbb{Q}_x$ is the law of a tree-indexed CSBP started with an initial population of size x constructed from the subordinator $\{\tau_t^\alpha : t \ge 0\}$. This will be done by induction on k. The case $k = 1$ is given in the convergence (1.44). Assuming the result holds for$k$, we will prove the convergence for $k + 1$. Conditioning with respect to $\mathcal{G}_k = \sigma(\mathcal{A}_u^{(n)}, d_u^{(n)} : |u| \le k)$ and using Lemma 1.3 we have

$$\begin{aligned}
&\mathbb{E}_{a(n)}^{p(n)}\left[\mathbb{E}_{a(n)}^{p(n)}\left(\prod_{i=1}^{k+1} f_i((r(n))^{-1}\mathcal{A}_u, (r(n)p(n))^{-1}d_u : |u| = i) \,|\, \mathcal{G}_k\right)\right] \\
&\qquad = \mathbb{E}_{a(n)}^{p(n)}\left[\prod_{i=1}^{k} f_i((r(n))^{-1}\mathcal{A}_u, (r(n)p(n))^{-1}d_u : |u| \le i)\right. \\
&\qquad\qquad \left.\times \mathbb{E}_1^{p(n)}\left(f_{k+1}(((r(n))^{-1}T_0, (r(n)p(n))^{-1}M_1)^{d_u\downarrow} : |u| = k)\right)\right].
\end{aligned}$$

Besides, by the induction hypothesis $d_u \sim r(n)p(n)\mathcal{Z}_u^{1/\alpha}$ with $|u| = k$, therefore when $n \to \infty$ in the previous equality we obtain

$$\mathbb{E}^{\mathbb{Q}_x}\left[\prod_{i=1}^{k} f_i(\mathcal{Z}_u^{1/\alpha}, \mathcal{Z}_u^{1/\alpha} : |u| \le i)\mathbb{E}_1(f_{k+1}((\mathbf{a}'_1, \mathbf{a}'_2, ...)))\right],$$

where $\mathbf{a}'_k = (a'_k, a'_k)$ are the atoms of a Poisson random measure on $(0, \infty)$ with intensity $b\mathcal{Z}_u^{1/\alpha}\nu^\alpha$, repeated according to their multiplicity and ranked in the decreasing order. Due to the definition of a tree-indexed CSBP this concludes the proof.

■

1.4.3. Proof of Proposition 1.17

Observe that is enough to show the convergence of the Laplace transforms associated with the finite dimensional distributions of each processes. Since $\{Z_k^{1/\alpha} : k \in \mathbb{Z}_+\}$ is a CSBP with transition probabilities characterized by the subordinator τ^α, we use (1.16) to get its Laplace transform. Therefore we will establish the following convergence:

$$\mathbb{E}_{a(n)}^{p(n)}\left(\prod_{i=1}^{k} e^{-\frac{s_{i-1}}{r(n)}T_{i-1} - \frac{t_i}{r(n)p(n)}M_i}\right) \xrightarrow[n\to\infty]{} \exp\{-x\kappa(l_{k-1}(s_0 + t_1))\}, \quad \text{for all } s_i, t_i \ge 0, i = 1, 2, ..., k. \tag{1.46}$$

According to (1.15), κ denotes the cumulant of the subordinator that characterizes the transition probabilities. In this case we have the $1/\alpha$-stable subordinator τ^α with the Lévy measure ν^α. The function l is given in (1.17).

We will prove the converge (1.46) by induction on k. The aim of the following lemma is to ensure the above claimed result holds for $k = 0$.

Lemma 1.19. *For $\alpha \in (1, 2)$, let τ^α be an $1/\alpha$-stable subordinator with no drift and Lévy measure ν^α. Assuming (1.19) and (1.21)*

i) the following convergences holds:

$$\mathcal{L}\left(\left(\frac{T_0}{r(n)}, \frac{M_1}{r(n)p(n)}\right), \mathbb{P}_{a(n)}^{p(n)}\right) \Longrightarrow (\tau_x^\alpha, \tau_x^\alpha); \tag{1.47}$$

ii) under the measure $\mathbb{P}_1^{p(n)}$, the behavior of the joint tail distribution of T_0 and M_1 is given by

$$\lim_{n\to\infty} r(n)p(n)\mathbb{P}_1^{p(n)}\left(\frac{T_0}{r(n)} > s, \frac{M_1}{r(n)p(n)} > t\right) = \bar{\nu}^\alpha(s \wedge t), \tag{1.48}$$

where $\bar{\nu}^\alpha$ denotes the tail function of the Lévy measure ν^α.

A key tool to establish Lemma 1.19 is the following:

Lemma 1.20. *In the regime (1.19) and (1.21) the normalized random walk defined by*

$$\bar{\mathbf{S}}^{(n)}_{\lfloor r(n)t\rfloor} = (a(n)/r(n)p(n), 0) + \sum_{i=1}^{\lfloor r(n)t\rfloor} \left((\xi_i^{(cn)} - 1)/n, \xi_i^{(mn)}/r(n)p(n)\right), \quad t \geq 0,$$

converges weakly

$$\left\{\bar{\mathbf{S}}^{(n)}_{\lfloor r(n)t\rfloor} : t \geq 0\right\} \Longrightarrow \{(x + X_t, t) : t \geq 0\},$$

where $\{X_t : t \geq 0\}$ is an α-stable process with no-negative jumps with and characteristic exponent $c_\alpha|\lambda|^\alpha$.

Given that this result is similar to other existing results in the literature we prefer to postpone its proof to the Appendix D and focus in the proof of Lemma 1.19.

Proof of Lemma 1.19 i). From Lemma 3 of Bertoin (2010), we know that the first passage time below 0 for the centered random walk $\mathbf{S}^{(n)}_{1,k} = a(n) + \sum_{i=1}^{k}(\xi_k^{(cn)} - 1)$ has the same distribution as T_0. Let $(\varsigma^{(n)}(0), \Sigma^{(n)}(0))$ be the random variables

$$\varsigma^{(n)}(0) = \inf\{k \in \mathbb{Z}_+ : \mathbf{S}^{(n)}_{1,k} = 0\} \quad \text{and} \quad \Sigma^{(n)}(0) := \sum_{i=1}^{\varsigma^{(n)}(0)} \xi_i^{(mn)}.$$

According to Lemma 3 of Bertoin (2010), we have that $(\varsigma^{(n)}(0), \Sigma^{(n)}(0))$ has the same distribution as (T_0, M_1) under $\mathbb{P}^{p(n)}_{a(n)}$. On the other hand, we also have the following two identities

$$\frac{\varsigma^{(n)}(0)}{r(n)} = \frac{1}{r(n)} \inf\{k \in \mathbb{Z}_+ : \mathbf{S}^{(n)}_{1,k} = 0\} = \inf\{t \geq 0 : \bar{\mathbf{S}}^{(n)}_{1,\lfloor r(n)t\rfloor} = 0\},$$

and

$$\left(\frac{1}{r(n)}\varsigma^{(n)}(0), \bar{\mathbf{S}}^{(n)}_{\varsigma^{(n)}(0)}\right) = \left(\frac{1}{r(n)}\varsigma^{(n)}(0), \left(\bar{\mathbf{S}}^{(n)}_{1,\lfloor\varsigma^{(n)}(0)\rfloor}, \frac{1}{r(n)p(n)}\Sigma^{(n)}(0)\right)\right). \tag{1.49}$$

From the Lemma 1.20 we have the weak convergence

$$\left\{\bar{\mathbf{S}}^{(n)}_{\lfloor r(n)t\rfloor} : t \geq 0\right\} \Longrightarrow \{(x + X_t, t) : t \geq 0\},$$

and in fact the convergence holds in the sense of Skorohod's topology, see Chapter IV of Whitt (2002). Since X is an α-stable process, Theorem 1 in Chapter VII of Bertoin (1996) ensures that the first passage time below $-x$ for the process X

$$\tau^\alpha_x = \inf\{t \geq 0 : X_t \leq -x\}, \quad x \geq 0.$$

is a stable subordinator of parameter $1/\alpha$. We will conclude from these facts that the claimed convergence holds as soon as

$$\left(\frac{1}{r(n)}\varsigma^{(n)}(0), \bar{\mathbf{S}}^{(n)}_{\varsigma^{(n)}(0)}\right) \Longrightarrow \left(\tau^\alpha_x, (x + X_t, t)|_{t=\tau^\alpha_x}\right). \tag{1.50}$$

But according to Theorem 13.6.5 of Whitt (2002) about weak convergence of first passage times and undershoots and overshoots, when there is convergence in Skorohod's topology, we have

$$\left(\frac{1}{r(n)}\varsigma^{(n)}(0), \bar{\mathbf{S}}^{(n)}_{1,\varsigma^{(n)}(0)}\right) \Longrightarrow \left(\tau^{\alpha}_{x}, X_{\tau^{\alpha}_{x}} + x\right).$$

Moreover since we have the joint convergence

$$\left\{\left(\bar{\mathbf{S}}^{(n)}_{1,\lfloor\varsigma^{(n)}(0)t\rfloor}, \bar{\mathbf{S}}^{(n)}_{2,\lfloor\varsigma^{(n)}(0)t\rfloor}\right) : t \geq 0\right\} \Longrightarrow \left\{(x + X_t, t) : t \geq 0\right\},$$

in the sense of Skorohod's topology, and the second coordinate is a determinist linear function, it is an elementary exercise to extend the above mentioned result of Whitt (2002) to get that the convergence in (1.50) holds. ■

Proof of Lemma 1.19 ii). We will apply the same techniques used in the proof of statement (ii) in Lemma 4 of Bertoin (2010). Let us start observing that for every $x, y \in \mathbb{R}$

$$e^{-sx-ty} = st\int_0^\infty \int_0^\infty e^{-sx-ty}\mathbf{1}_{\{x<u,y<v\}}dudv, \quad s,t \geq 0.$$

Thus Fubini's Theorem implies that for any random vector (X, Y) the following identity holds.

$$1 - \mathbb{E}(e^{-sX-tY}) = st\int_0^\infty \int_0^\infty e^{-sx-ty}\mathbb{P}(X \geq u \text{ or } Y \geq v)dudv, \quad s,t \geq 0.$$

In particular,

$$1 - \mathbb{E}_1^{p(n)}(e^{-\frac{s}{r(n)}T_0 - \frac{t}{r(n)p(n)}M_1}) = st\int_0^\infty \int_0^\infty e^{-su-tv}\bar{\mu}_n(r(n)u, r(n)p(n)v)dudv, \quad s,t \geq 0,$$

where $\bar{\mu}_n(x,y) := \mathbb{P}_1^{p(n)}(T_0 > x \text{ or } M_1 > y)$. Hence by the branching property,

$$\mathbb{E}^{p(n)}_{a(n)}(e^{-\frac{s}{r(n)}T_0 - \frac{t}{r(n)p(n)}M_1}) = \left(1 - st\int_0^\infty \int_0^\infty e^{-su-tv}\bar{\mu}_n(r(n)u, r(n)p(n)v)dudv\right)^{a(n)}. \tag{1.51}$$

According to the first part of this lemma together with (1.21), the previous display converges as $n \to \infty$ towards

$$\mathbb{E}\left(e^{-(s+t)\tau^{\alpha}_{x}}\right) = \exp\left(-x\int_0^\infty (1 - e^{-(s+t)y})\nu^{\alpha}(dy)\right).$$

Taking logarithms in the last two identities we obtain that

$$\lim_{n\to\infty} sta(n)\int_0^\infty \int_0^\infty e^{-su-tv}\bar{\mu}_n(r(n)u, r(n)p(n)v)dudv = x\int_0^\infty (1 - e^{-(s+t)y})\nu^{\alpha}(dy).$$

Hence it only remains to see that the line above is equal to

$$xst\int_0^\infty\int_0^\infty e^{-sy-tz}\bar{\nu}^\alpha(y\wedge z)dydz.$$

For that end we observe the equality

$$\int_0^\infty\int_0^\infty e^{-sy-tz}\bar{\nu}^\alpha(y\wedge z)dydz=\int_0^\infty\nu^\alpha(du)\int_0^\infty\int_0^\infty e^{-sy-tz}\mathbf{1}_{\{u>y \text{ or } u>z\}}dydz,$$

and we obtain the claimed identity by uniqueness of the Laplace transform. ∎

Continuing with the proof of Proposition 1.17, we assume that (1.46) holds for $n=k$ and prove the convergence for $n=k+1$. To this end, we use the Markov property of $\{(T_n,M_{n+1}):n\in\mathbb{Z}_+\}$ and the fact that, conditionally to $M_n=j$, the pair (T_n,M_{n+1}) has the same distribution as (T_0,M_1) under $\mathbb{P}_j$, to obtain

$$\begin{aligned}&\mathbb{E}^{p(n)}_{a(n)}\left(\prod_{i=1}^{k+1}e^{-\frac{s_{i-1}}{r(n)}T_{i-1}-\frac{t_i}{r(n)p(n)}M_i}\right)\\&\quad=\mathbb{E}^{p(n)}_{a(n)}\left(e^{-\frac{s_0}{r(n)}T_0-\frac{t_1}{r(n)p(n)}M_1}\cdots e^{-\frac{s_{k-1}}{r(n)}T_{k-1}-\left(\frac{t_k}{r(n)p(n)}-\frac{1}{r(n)p(n)}\log\mathbb{E}^{p(n)}_{r(n)p(n)}\left(e^{-\frac{s_k}{r(n)}T_0-\frac{t_{k+1}}{r(n)p(n)}M_1}\right)\right)M_k}\right).\end{aligned}$$

Due to the assumption $r(n)p(n)\sim a(n)x$ in hypothesis (1.21), we obtain as a consequence of Lemma 1.19 (i), that

$$\mathbb{E}^{p(n)}_{a(n)}\left(\prod_{i=1}^{k+1}e^{-\frac{s_{i-1}}{r(n)}T_{i-1}-\frac{t_i}{r(n)p(n)}M_i}\right)\sim\mathbb{E}^{p(n)}_{a(n)}\left(e^{-\frac{s_0}{r(n)}T_0-\frac{t_1}{r(n)p(n)}M_1}\cdots e^{-\frac{s_{k-1}}{r(n)}T_{k-1}-\frac{1}{r(n)p(n)}(t_k+\kappa(s_k+t_{k+1}))M_k}\right).$$

Then using the induction hypothesis with

$$s'_{i-1}+t'_i=\begin{cases}s_{i-1}+t_i & i<k,\\ l(s_{i-1}+t_i) & i=k,\end{cases}\tag{1.52}$$

we get

$$\mathbb{E}^{p(n)}_{a(n)}\left(e^{-\frac{s_0}{r(n)}T_0-\frac{t_1}{r(n)p(n)}M_1}\cdots e^{-\frac{s_{k-1}}{r(n)}T_{k-1}-\frac{1}{r(n)p(n)}(t_k+\kappa(s_k+t_{k+1}))M_k}\right)\xrightarrow[n\to\infty]{}\exp\{-x\kappa(l_{k-1}(s'_0+t'_1))\}.$$

This concludes the proof because of the recursive definition of l_i given in (1.17) together with the choice of $s'_{i-1}+t'_i$,

$$\kappa(l_{k-2}(s'_1+t'_2))=\kappa(l_{k-1}(s_1+t_2)),$$

as consequence of $l_{k-1}(s'_0+ct'_1)=l_k(s_0+t_1)$.

1.4.4. Proof of Lemma 1.18

By the construction $\{\gamma_j^{(n)} : 1 \le j \le b(n)\}$ is a sequence of independent random variables. Besides, the convergence (1.48) in Lemma 1.19 implies $\gamma_j^{(n)} \xrightarrow{P} 0$ as $n \to \infty$, uniformly in j. Then

$$\sup_j \mathbb{E}(|\gamma_j^{(n)} \wedge 1|) \to 0.$$

Hence, according with the definition given in Chapter 4 of Kallenberg (2002), we have that $\{\gamma_j^{(n)} : 1 \le j \le b(n)\}$ is a null array. Thus we will get the convergence (1.43) as an application of Theorem 16.18 of Kallenberg (2002), once we verify the following conditions:

a) $\sum_j \mathbb{P}(\gamma_j^{(n)}(B) > 0) \to \eta(B)$, as $n \to \infty$, for all $B \in \hat{\mathcal{B}}$, where η is as defined in the statement of the lemma,

b) $\sum_j \mathbb{P}(\gamma_j^{(n)}(B) > 1) \to 0$, as $n \to \infty$, for all $B \in \mathcal{B}$.

Here $\mathcal{B}$ is the Borel σ-algebra of $[0,\infty)^2$, $\hat{\mathcal{B}} =: \{B \in \mathcal{B} : \gamma(\partial B) = 0 \text{ c.s.}\}$, with γ the measure defined in the statement of this lemma and the symbol ∂ denotes the boundary of B. Observe that the class of sets $B = ((b,\infty) \times \mathbb{R}_+) \cup (\mathbb{R}_+ \times (b',\infty))$ is a π-system which generates a λ-system that coincides with $\mathcal{B}$. Then, by Dynkin's Theorem it is enough to establish the conditions above for sets of the latter form B. In this setting, the condition (b) holds because $\gamma_j^{(n)}(\cdot)$ takes only the values 0 or 1, for any j and n. To establish (a), observe the following identities

$$\begin{aligned}\sum_{j=1}^{b(n)} \mathbb{P}(\gamma_j^{(n)}(B) > 0) &= \sum_{j=1}^{b(n)} \mathbb{P}(\gamma_j^{(n)}(B) = 1) \\ &= \sum_{j=1}^{b(n)} \bar{\mu}_n((r(n))^{-1}s, (r(n)p(n))^{-1}t) \\ &= b(n)\bar{\mu}_n((r(n))^{-1}s, (r(n)p(n))^{-1}t),\end{aligned}$$

here we recall the identity $\bar{\mu}_n(x,y) = \mathbb{P}_1^{p(n)}(T_0 > x \text{ or } M_1 > y)$. Assuming the behavior $b(n) \sim br(n)p(n)$ for some $b > 0$, from Lemma 1.19 and the last equality we have

$$\sum_{j=1}^{b(n)} \mathbb{P}(\gamma_j^{(n)}(B) > 0) \xrightarrow[n\to\infty]{} b\bar{\nu}^\alpha(s \wedge t)).$$

To get the first convergence in the lemma, it remains to observe that $b\bar{\nu}^\alpha(s \wedge t) = \eta(B)$ holds. But this follows from the equalities,

$$\int_B \eta(dx,dy) = b\int_{(x,x)\in B} \nu^\alpha(dx) = b\int_{(s\wedge t,\infty)} \nu^\alpha(dx).$$

We will now prove the convergence (ii). For $i = 1, 2$, $\chi_{ij}^{(n)}$ denotes the i-th coordinate of the sequence $\chi_j^{(n)}$ that appears in the statement (i). Assuming that $\chi_i^{(n)} := (\chi_{1i}^{(n)}, \chi_{2i}^{(n)}) \leq \chi_j^{(n)} := (\chi_{1j}^{(n)}, \chi_{2j}^{(n)})$ if and only if $\chi_{1i}^{(n)} \leq \chi_{1j}^{(n)}$ or $\chi_{2j}^{(n)} \leq \chi_{2j}^{(n)}$, let us define j_1 as the index where the maximum of the sequence $\{\chi_i^{(n)} : 1 \leq i \leq b(n)\}$ is reached.

$$\chi_{j_1}^{(n)} = \max_{1 \leq i \leq b(n)} \chi_i.$$

Similarly for $k = 2, ..., b(n)$, let j_k be the index of the k-order statistic

$$\chi_{j_k}^{(n)} = \max_{i \in J_k} \chi_i,$$

where $J_k = \{1, ..., b(n)\} \setminus \{j_1, ..., j_{k-1}\}$. Then observe that

$$\mathbb{P}(\chi_{j_1}^{(n)} \geq \mathbf{c}_1, \chi_{j_2}^{(n)} \geq \mathbf{c}_2, ..., \chi_{j_k}^{(n)} \geq \mathbf{c}_k) = \mathbb{P}(\gamma_n(C_1) \geq 1, \gamma_n(C_2) \geq 2, ..., \gamma_n(C_k) \geq k),$$

if $\mathbf{c}_i = (c_i, c_i)$, $C_i = (0,1) \times (0,1) \setminus (0, c_i) \times (0, c_i)$ and $c_1 > \cdots > c_k$. Taking now the limit as $n \to \infty$ in the equality below and using the convergence in (1.48), we have

$$\mathbb{P}(\chi_{j_1} \geq \mathbf{c}_1, \chi_{j_2} \geq \mathbf{c}_2, ..., \chi_{j_k} \geq \mathbf{c}_k) \xrightarrow[n \to \infty]{} \mathbb{P}(\gamma(C_1) \geq 1, \gamma(C_2) \geq 2, ..., \gamma(C_k) \geq k).$$

This implies the desired convergence because

$$\mathbb{P}(\gamma(C_1) \geq 1, ..., \gamma(C_k) \geq k) = \mathbb{P}(\mathbf{a}_{j_1} \geq \mathbf{c}_1, \mathbf{a}_{j_2} \geq \mathbf{c}_2, ..., \mathbf{a}_{j_k} \geq \mathbf{c}_k),$$

where $\mathbf{a}_k = (a_k, a_k)$ with $\{a_k : k \in \mathbb{N}\}$, the atoms of a Poisson random measure on $(0, \infty)$ with intensity $b\nu^\alpha$ ranked in decreasing order; the measure ν^α is given in Theorem 1.7. As before we used the indices j_k to rank in decreasing order the sequence $\mathbf{a}_k$.

■

1.5. Asymptotic behavior: the conditioned to non-extinction case

This section is devoted to establish Theorem 1.8. Following the same strategy of Proposition 1.17, we shall establish by induction the convergence of Laplace transforms of the finite dimensional distributions associated with the processes involved. With this aim, we firstly deduce the Laplace transform of the finite dimensional distributions of a CSBP with immigration $\{Z_n^I : n \in \mathbb{N}\}$, with mechanism (ϑ, ι). We recall that it is defined for every $n \in \mathbb{N}$ as follows

$$Z_{n+1}^I = \tau_n(Z_n^I) + I_n,$$

where $\{I_n : n \in \mathbb{N}\}$ is a sequence of nonnegative random variables with common probability measure ω, which determine the distribution of individual immigrants arriving in the population. Let us denote its Laplace transform of ω by ι, i.e.

$$\iota(\lambda) = \int_0^\infty e^{-\lambda x}\omega(dx), \quad \lambda \geq 0. \tag{1.53}$$

Let $\{T^{(n)}(t) : t \geq 0\}_{n\geq 0}$ be a sequence of independent subordinators (without drift) and also independents of I_n, with the same distribution and Laplace transform given in (1.15). Thereby

$$\mathbb{E}(e^{-\lambda T^{(n)}(Z_n^I)}|Z_n^I) = e^{-Z_n^I\kappa(\lambda)}.$$

This previous equality together with the Markov property imply that

$$\mathbb{E}_x(e^{-s_1Z_1^I-\cdots-s_kZ_k^I}) = \prod_{i=0}^{k-1}\iota(l_i(s_{k-i}))e^{-x\kappa(l_{k-1}(s_1))}, \quad \text{for all } s_i \geq 0, i = 1, 2, ..., k;$$

where the functions $l_\cdot(s_{k-\cdot})$ are defined in (1.17). Thus the proof of statement (i) of Theorem 1.8 require to establish that for all $s_i, t_i \geq 0$, $i = 1, 2, ..., k$, the convergence below holds

$$\mathbb{E}_{a_n}^{p(n)\uparrow}\left(\prod_{i=1}^{k} e^{-\frac{s_{i-1}}{n^2}T_{i-1}-\frac{t_i}{n}M_i}\right) \xrightarrow[n\to\infty]{} \exp\{-x\kappa(l_{k-1}(s_0 + c_\vartheta t_1))\}\prod_{i=0}^{k-1}\iota(l_{k-i}(s_{i-1} + \tilde{c}_\vartheta t_i)), \tag{1.54}$$

where κ and l are respectively defined in (1.15) and (1.17), taking in particular $\vartheta = \nu$ given in (1.13). To obtain (ii) of Theorem 1.8, the previous convergence is proved with $\vartheta = \nu^\alpha$ defined in (1.22). The following lemma establishes the above convergence in the case $k = 1$. In its proof we use the reference Bertoin (2010) to justify the statement corresponding to $\vartheta = \nu$ and previous results here obtained to establish the case where $\vartheta = \nu^\alpha$.

Lemma 1.21. *If (1.11) holds, then we have the following convergence*

$$\mathcal{L}\left((b_1(n)T_0, b_2(n)M_1), \mathbb{P}_{a(n)}^{p(n)\uparrow}\right) \Longrightarrow (\tau, c_\vartheta\tau),$$

where τ is a random variable with Laplace transform $e^{-\kappa(s)}\iota(s)$, where $\kappa(s)$, $\iota(s)$ are given in (1.15) and (1.53), according to

i) if the reproduction law has finite variance σ^2, $\vartheta = c^{-1}\nu$, where ν is the measure in (1.13). Moreover $b_1(n) = n^{-2}$, $b_2(n) = n^{-1}$ and $c_\vartheta = c$;

ii) otherwise, under the assumptions (1.19) and (1.21), we have $b_1(n) = (r(n))^{-1}$, $b_2(n) = (r(n)p(n))^{-1}$, $\vartheta = \nu^\alpha$ is given by (1.22) and $c_\vartheta = 1$.

Proof. We will only prove the claim in the setting (i). The proof in the other case is fully analogue. To simplify the notation we just write b_1 and b_2. We prove the convergence of Laplace transform (b_1T_0, b_2M_1) under the measure $\mathbb{P}^{p(n)\uparrow}_{a(n)}$. First, recalling the definition of the conditional measure given in (1.30), an elementary calculation using the branching property shows

$$\mathbb{E}^{p(n)\uparrow}_{a(n)}\left(e^{-sT_0-tM_1}\right) = \mathbb{E}^{p(n)}_{a(n)-1}\left(e^{-sT_0-tM_1}\right)\mathbb{E}^{p(n)}_{1}\left(e^{-sT_0-tM_1}M_1\right). \tag{1.55}$$

Thanks to Lemma 4 of Bertoin (2010) and Lemma 1.19.

$$\mathbb{E}^{p(n)}_{a(n)}\left(\exp\left(-sb_1T_0 - tb_2M_1\right)\right) \xrightarrow[n\to\infty]{} \exp\left(-x\int_0^\infty (1-e^{-(s+c_\vartheta t)y})c_\vartheta^{-1}\vartheta(dy)\right), \tag{1.56}$$

so it remains to calculate the limit of the second factor in (1.55). Using again Bertoin (2010) together with the equality (1.51), we have

$$\mathbb{E}^{p(n)}_{1}\left(\exp\left(-sb_1T_0 - tb_2M_1\right)\right) = 1 - st\int_0^\infty\int_0^\infty e^{-sx-ty}\bar{\mu}_n(b_1^{-1}x, b_2^{-1}y)dxdy,$$

where as before $\bar{\mu}_n(x,y) := \mathbb{P}^{p(n)}_1(T_0 > x \text{ or } M_1 > y)$. Due to Lemma 4 (ii) in Bertoin (2010) and Lemma 1.19 (ii) the latter implies

$$\begin{aligned}\mathbb{E}^{p(n)}_{1}\left(\exp\left(-sb_1T_0 - tb_2M_1\right)M_1\right) \xrightarrow[n\to\infty]{} & s\int_0^\infty\int_0^\infty e^{-sx-ty}\bar{\vartheta}\left(x\wedge\frac{y}{c_\vartheta}\right)dxdy \\ & - st\int_0^\infty\int_0^\infty ye^{-sx-ty}\bar{\vartheta}\left(x\wedge\frac{y}{c_\vartheta}\right)dxdy.\end{aligned}$$

Computing the integrals we get

$$\mathbb{E}^{p(n)}_{1}\left(\exp\left(-sb_1T_0 - tb_2M_1\right)M_1\right) \xrightarrow[n\to\infty]{} \int_0^\infty e^{-(s+c_\vartheta t)z}z\vartheta(dz). \tag{1.57}$$

This finishes the proof. ∎

We can now continue with the proof of Theorem 1.8. We assume that (1.54) holds for k and verify it also holds for $k+1$. Let $\mathcal{F}_k = \sigma((M_{j-1}, T_j), j \le k)$ and (T_0', M_1') be an independent copy of (T_0, M_1). Recalling the definition of the measure $\mathbb{P}^{\uparrow}_a$ in Theorem 1.4 we have for any $a \in \mathbb{N}$ and $p > 0$ that,

$$\begin{aligned}&\mathbb{E}^{p\uparrow}_a\left(e^{-\lambda_0T_0-\theta_1M_1}\cdots e^{-\lambda_kT_k-\theta_{k+1}M_{k+1}}\right) \\ &\qquad = \mathbb{E}^{p}_a\left(e^{-\lambda_0T_0-\theta_1M_1}\cdots e^{-\lambda_{k-1}T_{k-1}-\theta_kM_k}\frac{1}{a}\mathbb{E}^{p}_{M_k}\left(e^{-\lambda_kT_0'-\theta_{k+1}M_1'}M_1'\right)\right).\end{aligned}$$

Then applying the identity (1.55) we get

$$\begin{aligned}&\mathbb{E}^{p\uparrow}_a\left(e^{-\lambda_0T_0-\theta_1M_1}\cdots e^{-\lambda_kT_k-\theta_{k+1}M_{k+1}}\right) \\ &\quad = \mathbb{E}^{p}_a\left(e^{-\lambda_0T_0-\theta_1M_1}\cdots e^{-\lambda_{k-1}T_{k-1}-\theta_kM_k}\frac{M_k}{a}\left(\mathbb{E}^{p}_1\left(e^{-\lambda_kT_0'-\theta_{k+1}M_1'}\right)\right)^{M_k-1}\mathbb{E}^{p}_1\left(e^{-\lambda_kT_0'-\theta_{k+1}M_1'}M_1'\right)\right).\end{aligned}$$

Using the Markov property of $\{(T_n, M_{n+1}) : n \in \mathbb{Z}_+\}$ and writing the terms suitably, we get

$$\begin{aligned}
&\mathbb{E}_a^{p\uparrow}\left(e^{-\lambda_0 T_0-\theta_1 M_1}\ldots e^{-\lambda_k T_k-\theta_{k+1}M_{k+1}}\right)\\
&\quad= e^{-\frac{1}{n}\log \mathbb{E}_n^p\left(e^{-\lambda_k T_0'-\theta_{k+1}M_1'}\right)}\mathbb{E}_1^p\left(e^{-\lambda_k T_0'-\theta_{k+1}M_1'}M_1'\right)\\
&\qquad\qquad\times \mathbb{E}_a^p\left(e^{-\lambda_0 T_0-\theta_1 M_1}\ldots e^{-\lambda_{k-1}T_{k-1}}e^{-\left(\theta_k-b_2\log \mathbb{E}_{b_2^{-1}}^{p(n)}\left(e^{-\lambda_k T_0'-\theta_{k+1}M_1'}\right)\right)M_k}\frac{M_k}{a}\right).
\end{aligned}$$

In the previous equality, $(\lambda_{i-1},\theta_i) = (b_1 s_{i-1}, b_2 t_i)$, $i = 1,\ldots,k+1$ and $(a,p) = (a(n),p(n))$, and we use hypotheses (1.11) and (1.21) to obtain

$$\begin{aligned}
&\mathbb{E}_{a(n)}^{p(n)\uparrow}\left(\prod_{i=1}^{k+1}e^{-b_1 s_{i-1}T_{i-1}-b_2 t_i M_i}\right)\\
&\qquad\sim e^{-b_2\log \mathbb{E}_{b_2^{-1}}^{p(n)}\left(e^{-b_1 s_k T_0'-b_2 t_{k+1}M_1'}\right)}\mathbb{E}_1^{p(n)}\left(e^{-b_1 s_k T_0'-b_2 t_{k+1}M_1'}M_1'\right)\\
&\qquad\qquad\times \mathbb{E}_{a(n)}^{p(n)\uparrow}\left(e^{-b_1 s_0 T_0-b_2 t_1 M_1}\ldots e^{-b_1 s_{k-1}T_{k-1}-b_2(t_k+\kappa(s_k+ct_{k+1}))M_k}\right).
\end{aligned}$$

Now we have to calculate the limit of each factor. The first one converges towards to 1 thanks to (1.56). Besides, to get

$$\mathbb{E}_1^{p(n)}\left(e^{-b_1 s_k T_0'-b_2 t_{k+1}M_1'}M_1'\right)\xrightarrow[n\to\infty]{}\iota\big(\kappa_0(s_k+c_\vartheta t_{k+1})\big), \tag{1.58}$$

we use the convergence (1.57) together with the convention $\kappa_0(s) = s$. As in Proposition 1.17, in order to conclude we use the induction hypothesis with $s'_{i-1}+t'_i$, $1\le i\le k$ as defined in (1.52).

Chapter 2

Gene trees and species trees

The aim of this chapter is to motivate the family of coalescent processes that will be introduced in Chapter 3. In this sense we provide an overview of the models and techniques that have been used in molecular biology to describe the relationships between gene trees and species trees. We emphasize that this chapter is a compilation in this subject based on Liu et al. (2009) and Szöllősi et al. (2014). We suggest to the reader being familiar with these works to skip this part and continue the reading on Chapter 3.

2.1. Introduction

Phylogeny is used to represent the evolutionary history of species observed through time, and is one of the most important entities in evolutionary biology. It assumes that all species arise from a common ancestor and genetic material is transmitted from ancestors to descendants along the branches of the phylogenetic tree. Phylogenetic information is encoded in the genetic material of contemporary species in a manner that allows the information from data such as DNA sequences, to be used to trace the history back to the most recent common ancestor of the species. This approach has been extremely fruitful indeed, the improvement in the accuracy and resolution of phylogenetic reconstruction together with our understanding of evolutionary processes at the molecular level, are the most significant contribuitions. However, the reconstructed trees describe the history of fragments of genomic sequence, but never the history of species. Gene trees are not species trees, Maddison (1997).

"Each gene tree reflects a unique story, which is linked to species history, but often significantly differs from it. Gene tree reflects the process of replication at a local level: a copy of a gene at a locus in the genome, for example a protein, coding gene, replicates and its copies are passed on from parent to offspring, generating branching points in the gene tree. Because each gene has a single ancestral copy, barring recombination, the resulting history is a branching tree. Recombination however, breaks up the genomic history into a series of partially independent stories that is, into gene trees along the

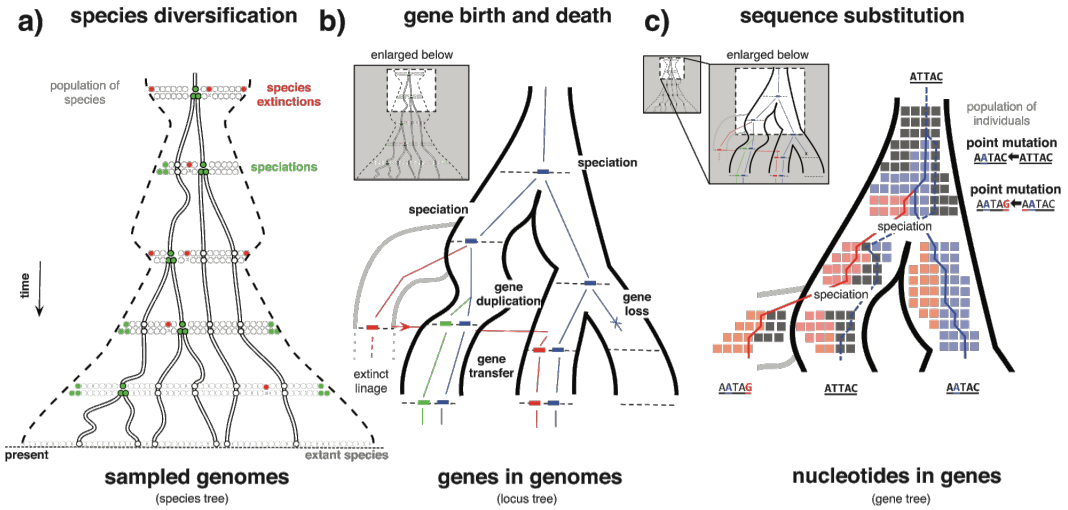

Figure 2.1: "A hierarchy of evolutionary processes contribute to sequence evolution. a) Individual species (circles) and their genomes evolve among a population of species, according to a diversification process consisting of speciation (light gray, green online) and extinction (dark gray, red online) events. The variation in the number of species existing at any given time is indicated by the dashed contour. When attempting to infer the species tree typically only a fraction of existing species (gray and black circles on dashed line) are sampled (black circles). b) Inside each genome, each gene evolves according to gene duplication, loss, and transfer events. c) Individual sites evolve through point mutations. Processes at the gene and site level are played out at the population level, where changes fix or are lost", Figure 1 in Szöllősi et al. (2014).

genomes of species. Starting from an individual site in a genome up, to the species level, a hierarchy of evolutionary processes generate genomic sequences. Individual sites evolve as a result of point mutations. The fate of individuals carrying each mutation is played out at the population level, and determines whether a mutation is fixed in the population as a substitution, or is ultimately lost. The birth and death of stretches of sequence, e.g. of single sites or even of entire genes, occurs as a result of insertions and deletions in individual genomes, the fate of which, similar to point mutations, is played out at the population level. The source of the inserted sequence differentiates between duplication events, wherein a sequence from the same genome is inserted, and lateral transfer events, wherein a sequence from an external source is inserted. Finally populations of genomes, evolve through speciation and extinction events. As illustrated in Figure 2.1 each level of the hierarchy contributes to generating phylogenetic signal that can lead to differences between reconstructed gene trees. Segregating mutations that cross speciation events (a process called incomplete lineage sorting) leave topological signatures in gene trees (see Fig. 2.1 (c))", Szöllősi et al. (2014).

2.2. The multispecies coalescent

The fact that a gene tree is the evolutionary history of alleles randomly chosen from species provides, from a biological perspective, a reasonable explanation for the relation-

ship between gene trees and the phylogeny of species (Pamilo and Nei (1988), Maddison (1997)). It indicates that a gene tree is a random tree generated within the phylogeny of species and phylogenies of species should be studied in the framework of probabilistic models that incorporate the probability distribution of gene trees given the phylogeny of species. Although a few techniques have been developed to specify this probability distribution in the context of a variety of biological phenomena such as horizontal gene transfer (HGT) and gene duplication/loss (Arvestad et al. (2003); Linz et al. (2007)), here we will focus on approaches that assume that the conflicts between gene trees and the species tree are exclusively due to deep coalescence (Maddison (1997); Maddison and Knowles (2006)). To be precise one may assume that the lineage dynamics within populations (or species) are well described by the conventional Wright-Fisher model and the distribution of the gene trees within each population is approximated by the coalescent process (or Kingman coalescent, see Chapter 3 or Felsenstein and Felenstein (2004); Hein et al. (2005); Wakeley (2008) for a more biological approach).

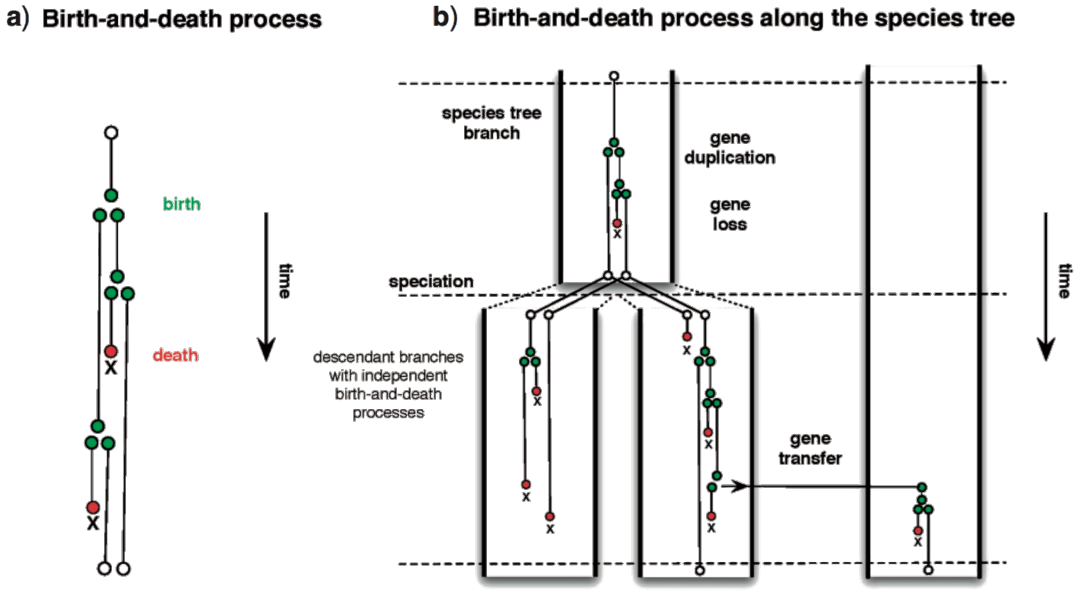

Figure 2.2: "Birth death processes for generating species trees and gene trees. Death events (species extinctions and gene loss) are in dark gray (red online), birth events (speciation, duplication and transfer events) are light gray (green online). a) Birth death processes modeling speciation and extinction. b) Birth death process modeling gene family evolution inside a species tree", Figure 2 in Szöllősi et al. (2014).

"Coalescent models aimed at modeling the discordance between gene tree and species tree arising from population-level processes have enjoyed increasing popularity in the last 10 years. Here birth events correspond to the appearance of a new allele, and death events to the disappearance of an allele, without any change in the locus of the gene. At any given time in a species, for a given locus in the genome, there may be several alleles, Figure 2.2. These alleles have their own history, some alleles being more closely related

than others. When speciation occurs, most alleles will be sorted randomly between the two incipient species: in some cases both species will receive copies of all alleles, in others, each will receive only a subset of the alleles present in the parent population. In all cases, the history reconstructed from the allele sequences will be the allele history, and not the species history. The allele history and the species history always differ in the timing of the bifurcation events: assuming no hybridization has occurred between the lineages, alleles have necessarily split before species split. They can also differ in their topology, especially if only a brief interval of time passes between successive speciation events, and/or the effective population size of the parent species is large (Rosenberg and Nordborg (2002)). Given the coalescent model, the amount of discordance in topology and divergence times between the trees of several loci and a species tree can therefore be used to estimate effective population sizes along the species tree (Rannala and Yang (2003); Liu and Pearl (2007); Heled and Drummond (2008); Minin et al. (2008); Kubatko et al. (2009)). Such a model where the population size is assumed to differ between the branches of the species tree has been called the *multispecies coalescent*", Szöllősi et al. (2014).

"The multispecies coalescent determines the probability distribution of gene trees and their branch lengths. The parameters of the distribution are the shape of the species tree, the divergence times within the species tree, and the population sizes along the branches of the species tree (one parameter for each branch). We bundle these parameters into the single composite parameter S, so that the probability of a gene tree $\mathbf{G}$ given the species tree is $f(\mathbf{G}|S)$, we treat this quantity as a density rather than a discrete probability because of the continuous branch lengths of $\mathbf{G}$.

The probability distribution $f(\mathbf{G}|S)$ can be used to infer species phylogenies when gene trees $\mathbf{G}$ are known. However, gene trees are generally unknown and phylogenies must be estimated from multilocus data. A probabilistic model for estimating species phylogenies consists of three components; multilocus data ($\mathbf{D}$), gene trees ($\mathbf{G}$), and the phylogeny of species (S) (Liu and Pearl (2007)). The data commonly be nucleotide sequence or amino acid data, but any data that contain phylogenetic information for a gene may be used. The data for a particular gene are the result of a process of descent with modification along the branches of the gene tree, while the gene tree itself is a random tree sampled from a probability distribution dependent on the phylogeny of species. Although gene trees are random conditional on the species tree, this need not mean that gene trees are heterogeneous in topology; depending on the species tree, random gene trees may be highly constrained and thus highly uniform in topology and branch lengths. The sequences are generated from the phylogeny S through two random processes; the process that generates gene trees from the phylogeny of species, which has probability distribution $f(\mathbf{G}|S)$, and the process that generates data from gene trees, which has probability distribution $f(\mathbf{D}|\mathbf{G})$. Putting $f(\mathbf{D}|\mathbf{G})$ and $f(\mathbf{X}|\mathbf{G})$ together, we obtain the joint probability (or density) of the alignment $\mathbf{D}$ and the gene tree $\mathbf{G}$:

$$\mathbb{P}(\mathbf{D}, \mathbf{G}|S) = \mathbb{P}(\mathbf{D}|\mathbf{G})\mathbb{P}(\mathbf{G}|S)$$

The gene tree G is not observed directly and it can be difficult to estimate. Since our focus is on the species tree and the features of the species tree, we work with the marginal probability of the data. Let Ψ be the set of all possible genealogies for the individuals incorporating both the topologies and branch lengths. The marginal probability for the data is then found by integrating over Ψ:

$$\mathbb{P}(\mathbf{D}|S) = \int_{\Psi} \mathbb{P}(\mathbf{D}|\mathbf{G})\mathbb{P}(\mathbf{G}|S)d\mathbf{G} \tag{2.1}$$

Generally, we have nucleotide and amino acid sequence data, for which the mutation process describes how the nucleotides in the sequences change through time along the branches of gene trees. For multilocus data, we use D_i and G_i to denote the aligned nucleotides or amino acids of all individuals sampled from the species under study and the gene tree (topology and branch lengths) for locus i, respectively. The probability distribution $f(D_i|G_i)$ of the alignment D_i given the gene tree G_i is the likelihood function traditionally used in the maximum likelihood method for estimating gene trees (Jukes and Cantor (1969); Felsenstein (1981); Hasegawa et al. (1985); Whelan and Goldman (2001); Sullivan (2005)). Assuming independence among loci in a multilocus data set, the likelihood function $f(\mathbf{D}|\mathbf{G})$ is the product of functions $f(D_i|G_i)$ across loci, which is used to measure the fit of gene trees to the multilocus sequence data.

An explicit mathematical formulation of the model described above for multilocus sequences $\mathbf{D}$, gene trees G, and the phylogeny of species S, then, is as follows:

$$\begin{aligned} f(\mathbf{D}|\mathbf{G}) &= \prod_{i=1}^{K} f(D_i|G_i) \\ f(\mathbf{G}|S) &= \prod_{i=1}^{K} f(G_i|S) \\ f(D_i|G_i, S) &= f(D_i|G_i) \quad \text{for } i = 1, ...K, \end{aligned}$$

where K is the number of genes. The last equation indicates that se- quences Di are conditionally independent of the phylogeny S when the gene tree G_i is given. The function $f(D_i|G_i)$ is the likelihood function derived from nucleotide substitution models (Jukes and Cantor (1969); Felsenstein (1981), Felsenstein and Felenstein (2004); Hasegawa et al. (1985)), while $f(G_i|S)$ is Rannala and Yang's gene tree density given a species tree (Rannala and Yang (2003)).

At this point, it is appropriate to reflect on what exactly is required when applying the equation (2.1) to large numbers of multilocus sequences. As in the likelihood analysis for gene tree estimation, one must evaluate a large number of phylogenetic trees, in this case a large number of species trees, in order to find the maximum likelihood estimate of the phylogeny. But in addition to having to evaluate a large number of species trees, the large number of gene trees for any given species tree means that the above likelihood is impractical to calculate directly for all but the smallest species trees. Even this sobering

conclusion does not quite hold if one has sampled many alleles per species, which necessarily vastly increases the number of gene trees to be evaluated, as in traditional phylogenetic analysis when one has sampled a large number of species (Felsenstein (1988))", Liu et al. (2009).

2.3. Estimating phylogenies of species

The phylogeny S can be estimated from multilocus sequence data from a multilocus sequence $\mathbf{D}$ using the likelihood function $f(\mathbf{D}|S)$. For example the maximum likelihood estimate (MLE) of the phylogeny S is given by

$$\hat{S} = \arg\max_{S}\{f(\mathbf{D}|S)\}.$$

Bayesian approaches assume a prior distribution for the phylogeny S and use the posterior distribution (the combination of likelihood and prior distributions) to infer phylogenies. The posterior distribution of the phylogeny S is

$$f(S|\mathbf{D}) = \frac{f(\mathbf{D}|S)f(S)}{f(\mathbf{D})}$$

Unlike the maximum likelihood and Bayesian approaches, which utilize the full data $\mathbf{D}$ and the likelihood function $f(\mathbf{D}|S)$ to infer the phylogeny of species (as well as prior distributions in the case of Bayesian methods), methods based on summary statistics seek to estimate the phylogeny S by summarizing the gene trees estimated from multilocus sequences. To summarize, one can see a phylogenetic pipeline as a series of statistical inferences, starting from raw sequences coming out of sequencing machines, and finishing with the inference of a species tree. Necessary steps include sequencing error correction, assembly of reads into contins and scaffolds, gene annotation, gene family clustering, alignment, and tree reconstruction. Most of these steps are done sequentially, so that later steps in the pipeline entirely disregard any estimate of uncertainty from the previous steps, and do not provide any feedback to these. Gene tree-species tree models take a step toward a more principled approach, by allowing communication between two steps of this pipeline, the construction of gene trees, and the construction of a species tree.

"Figure 2.3 places the above discussed models and associated phylogenetic software in the context of the complete phylogenetic inference pipeline. Gray nodes are considered known, and white nodes are inferred. This figure shows that a large diversity of inferential problems have been addressed, considering gene alignments, gene trees, species trees, or several of these as data. We refer to for a review of some popular methods and algorithms that have been used to address these inferential problems, simulating gene trees under the multispecies coalescent. In typical phylogenetic studies of individual genes, the estimated gene tree topology is used as the estimate of the species tree topology. When many loci are studied, the species tree topology is often estimated using the most frequently

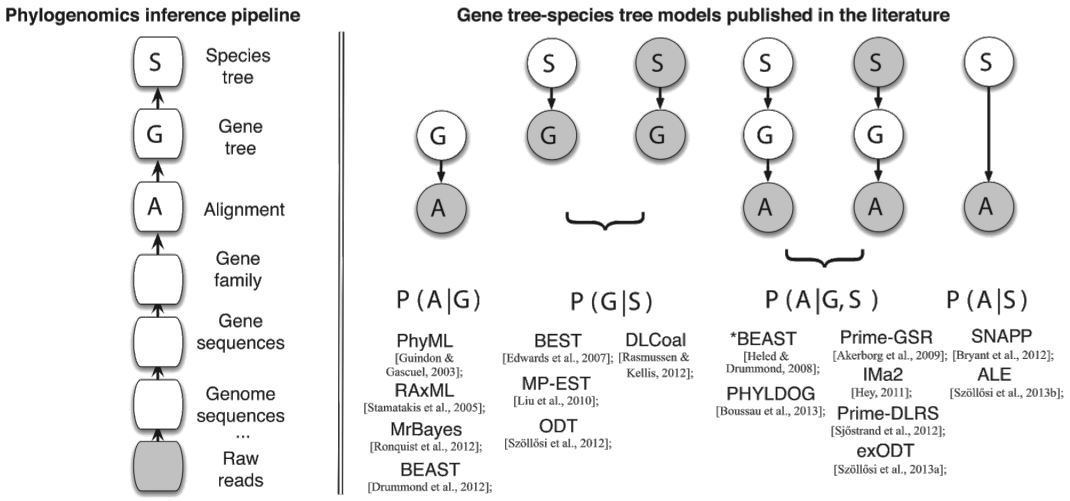

Figure 2.3: "Gene tree-species tree models in the context of the phylogenomics inference pipeline. Left: the inference pipeline (some steps are not represented, such as sequencing error correction). Right: graphical representation of the inferential problem for a selection of the models and associated phylogenetic software discussed in the main text. The sequence of steps in the graphical model representations correspond to the hierarchical sequence of evolutionary process generating genomic sequences. The likelihood that must be computed is also shown. Graphical model conventions are observed: stochastic nodes, nodes corresponding to data considered as known are gray, and nodes whose states are inferred are in white. The models have been simplified, and parameters others than the gene tree and the species tree have not been represented", Figure 5 in Szöllősi et al. (2014).

inferred gene tree topology. Although it is well-known that the sorting of gene lineages at speciation can cause gene trees to differ in topology from species trees, the assumption that the most probable gene tree topology to be produced by this sorting is the same as the species tree topology -the implicit premise that makes it sensible to estimate a species tree using a single gene tree or the most common among several gene trees- has remained unquestioned. Besides, recent advances in genealogical modeling suggest that resolving close species relationships is not quite as simple as applying more data into inference of species trees, because in general, neither gene trees nor species trees are known, indeed, during the last 30 years, the probability of concordance of gene trees and species trees has been a frequent source of discussion, this is the aim of the following section", Szöllősi et al. (2014).

2.4. The probability of topological concordance of gene trees and species trees

In the framework of multispecies coalescent, early theoretical work included the analytical derivation of the probabilities for different gene tree topologies relating four individuals from two different species, and showed that when the two populations diverged

only recently an incorrect tree is not the exception but a common occurrence (Tajima (1983)). Analytical results are also known for three individuals from three species (Nei (1987)).

We start off specifying the terminology. Takahata (1989) uses the term "consistent" to refer to those gene trees and species trees where the most recent interspecific coalescence occurred between the pair of sister species in the phylogeny, and this event took place later than the first bifurcation of the ancestral group to all three species. Besides Rosenberg (2002) uses the term "concordant" to call the event of a gene tree and species trees having the same topology (see Figure 2.4) which occurs if and only if the collapsed gene tree is congruent to the species tree. The collapsed gene tree from a gene tree is constructed, proceeding backwards in time until a coalescence of lineages occurs between two species. Group the two species involved in this coalescence into a clade, where clade is understood to subsume species as a special case. Continue backwards in time until another coalescence occurs between two clades. If both clades involved in this coalescence have already experienced inter-clade coalescences, ignore the event. If one or neither of the clades has already had interlude coalescences, group these two clades into a larger clade. Proceed backwards in time until all species have been involved in interclade coalescences, see Figure 2.5. Another definition of concordance was proposed in Pamilo and Nei (1988) herein, the mean coalescence times of two lineages, one from each species, are taken to define a distance, the values are kept in a matrix, roughly speaking the sense of concordance depends on the distance measurement between a pairwise of the matrix. With three species this decision is straightforward, however with more species the results may depend on which algorithm for constructing the topology from the matrix is used, thus to asses analytically that genes and species trees have the same topology is difficult. Hereafter we will be interested in the concordance of gene tree and species tree in the sense considered by Rosenberg (2002), because this topological definition allows gene trees to be partitioned into the topological classes and this property is useful for phylogenetic applications.

By treating the species tree as fixed (including branch lengths), in the framework of multispecies coalescent, gene lineages from separate species can only occur more anciently than the splitting times of the species to which they belong. Thus, the probability that i lineages coalesce into j lineages, within an amount of time where the length of the branch is t, was early derived in Tavaré (1984), using dynamics of Kingman coalescent. Namely, denoting this probability by $g_{ij}(t)$, the expression is the following,

$$g_{ij}(t) = \sum_{k=j}^{i} e^{-k(k-1)t/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)! i_{(k)}}, \tag{2.2}$$

where $a_{(k)} = a(a+1)\cdots(a+k-1)$ for $k \geq 1$ with $a_{(0)} = 1$ and $a_{[k]} = a(a-1)\cdots(a-k+1)$ for $k \geq 1$ with $a_{[0]} = 1$. $g_{ij}(t) = 0$, except when $1 \leq j \leq i$. To find this probability one have to consider the probability that a pair of genes have found a common ancestor or not after time t and the probability that a sample of i has j ancestors at time t. In

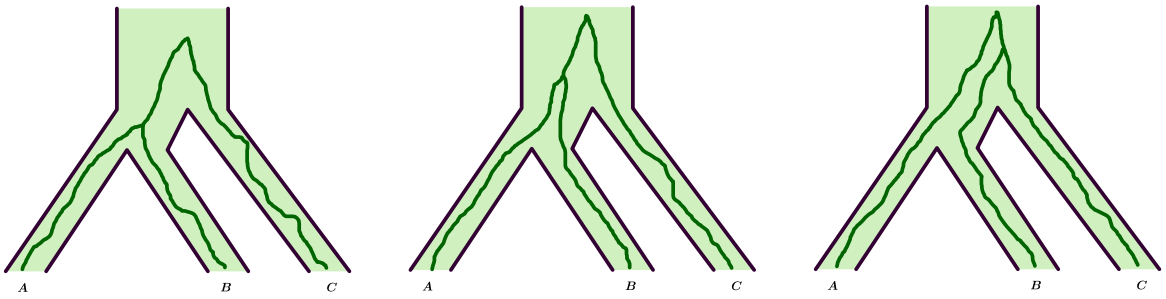

Figure 2.4: Congruence of gene trees and species trees. A, B, and C. (Left) Gene tree that is both congruent and Takahata-congruent to the species tree. (Middle) Gene tree that is congruent but not Takahata-congruent to the species tree. (Right) Gene tree that is neither congruent nor Takahata-congruent to the species tree.

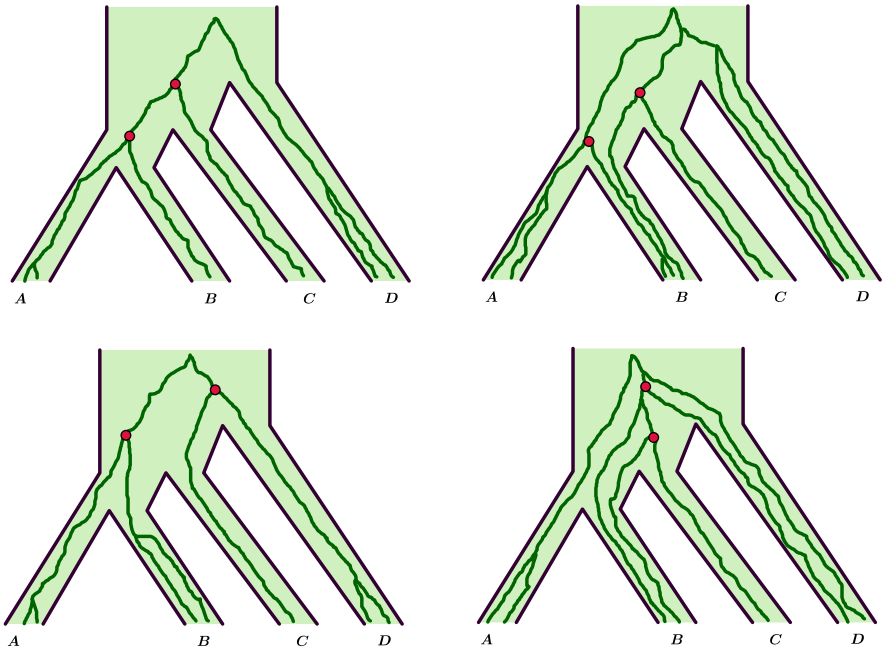

Figure 2.5: Conrcordance of genes and species trees. A B, C and D are present day species. Circules indicate interspecifc coalescences that are used in determining the collapsed gene tree. The collapsed gene tree for (up left) and (up right) have topology (((A,B)C)D). For (down left) the collapsed gene tree has topology ((A,B)(C,D)), and for (down right) the topology is (((B,C)D)A).

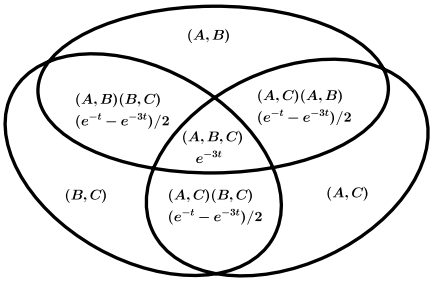

Figure 2.6: The probability that a sample of three genes have two ancestors at time t. An area is indicated with a set of sequences, these sequences have not found any common ancestors.

particular $g_{3,2}(t) = \frac{3}{2}(e^{-t} - e^{-3t})$, to explain this case, we include Figure 2.6, where each circle is the event where the two genes have not found an ancestor after time t and has probability e^{-t}. The intersection of all three circles has probability e^{-3t}, since it is the event that no genes have found common ancestors. The three intersections of two circles not including the third circle have probability $(e^{-t} - e^{-3t})/2$. Note that the figure is misleading because the part of a circle outside the intersections is empty: If two pairs of genes have found common ancestors at time t then all three pairs have. Therefore, all probabilities of the areas are defined and it is a simple addition-subtraction exercise to obtain the probabilities of interest. The event where the three genes have two ancestors are the areas that are contained in two circles. The area outside any circle are the events where all have found a common ancestor. Obviously for a large i and j book-keeping will be much more complicated, but would follow the same principles. For details of this deduction see Hein et al. (2005).

Observe now that gene and species trees are concordant, according to Takahata, if and only if the collapsed tree is congruent to the species tree, and the collapsed gene tree contains no coalescences prior to the most ancient species divergence, so a first step to compute the probability of topological concordance (in the sense of Rosenberg), is to determine the Takahata concordance probability, in this direction consider the species tree $((A, B)C)$. Assume that the species A, B and C have equal and constant haploid population sizes (all equal to N), equal generation times, and r, s, q are respectively, the lineages within species at the present and let T_i be the time of the i-th coalescence (Figure 2.7). If species tree topology and gene genealogy are known exactly, the probability that species A and B are respectively represented by m and n ancestral lineages at time T_1 is $g_{rm}(T_1)g_{sn}(T_1)$. As well, the probability that $m + n$ lineages in the ancestral species at time T_1 coalesce to k lineages at time $T_1 + T_2$ is $g_{m+n,k}(T_2)$.

Besides, let $F_k^{A,B}(r, s, q)$ be the probability that in coalescing from r, s, and q lineages from species A, B and C, respectively, to k total lineages, an interspecific coalescence occurs and the most recent interspecific coalescence links lineages of species A and B. In

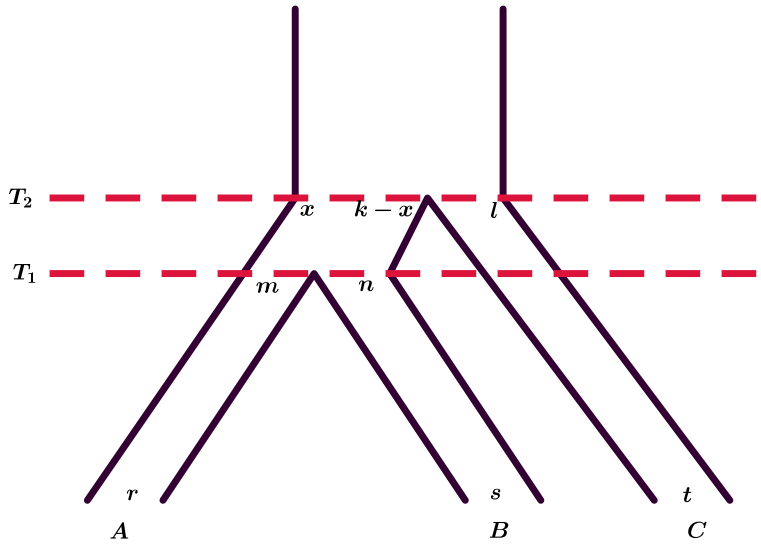

Figure 2.7: Three-species divergence model. The quantities r, s, and q are number of sampled lineages. The remaining variables, m, n, l, x and $k-x$, all represent number of ancestral lineages.

particular $F_k^{A,B}(m,n,0)$, denotes the probability that an interspecific coalescence occurs between a lineage of species A and a lineage of B, during the process of coalescence of these $m+n$ lineages to k lineages in the two-species phase. Finally, the conditional probability of Takahata concordance given configurations of lineages throughout the history of the three species, summed over possible configurations is

$$P_T(r,s,q,T_1,T_2) = \sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{k=1}^{m+n} g_{rm}(T_1)g_{sn}(T_1)g_{m+n,k}(T_2)F_k^{A,B}(m,n,0).$$

Now a term must be added to the latter expression to determine the probability of concordance. Namely, it is necessary to take into account the probability of all the following events: (a) no interspecific coalescences happen in the two-species phase; (b) the most recent interspecific coalescence happens in the one-species phase; and (c) this coalescence joins ancestral lineages of species A and B. In this aim observe that if m and n lineages from species A and B are present at time T_1, and these lineages have k total ancestors at time T_1+T_2, then the probability of the event (a) is $1-F_k^{A,B}(m,n,0)$. All the coalescences are intraspecific, at time T_1+T_2, there are, say, X_1 and X_2 ancestral lineages to species A and B, respectively. Since k total lineages are present at time T_1+T_2, we have that $X_1+X_2=k$. Also, $1\leq X_1, X_2 \leq k-1$ because each species is represented by at least one lineage. In order to determine probabilities of events in the one-species phase, we will need to consider all possible values of X_1 and X_2. Thus, $\mathbb{P}(X_1=x, X_2=k-x|X_1+X_2=k)$

denoted by $W_{(m,n)(x,k-x)}(T_2)$ is obtained using the Bayes's theorem,

$$W_{(m,n)(x,k-x)}(T_2) = \frac{g_{mx}(T_2)g_{n,k-x}(T_2)}{\sum_{i=1}^{k-1} g_{mi}(T_2)g_{n,k-i}(T_2)}.$$

Simultaneous to the entry of lineages from A and B into the one species phase, lineages ancestral to species C also enter the one-species phase. The probability that species C is represented by l ancestral lineages at time T_1+T_2 is $g_{ql}(T_1+T_2)$. The last quantity needed for the calculation is the probability $F_1^{A,B}(a,b,c)$ that for a, b, and c lineages from species A, B, and C present at the ancestral divergence, the most recent interspecific coalescence occurs between lines ancestral to species A and B. This probability is necessary because if the most recent interspecific coalescence involves a lineage from species C, the collapsed gene tree will be discordant with the species tree. Combining the various components, the topological concordance probability is

$$\begin{aligned} P_C(r,s,q,T_3,T_2) = \sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{k=1}^{m+n} & g_{rm}(T_3)g_{sn}(T_3)g_{m+n,k}(T_2) \\ & \times F_k^{A,B}(m,n,0) + [1 - F_k^{A,B}(m,n,0)] \sum_{x=1}^{k-1} W_{(m,n)(x,k-x)}(T_2) \\ & \times \sum_{l=1}^{q} g_{ql}(T_3+T_2) F_1^{A,B}(x,k-x,l) \;]. \end{aligned}$$

Observe that the key determinants of the topological concordance probability are the number of ancestral lineages to the samples of species A and B at time T_1, and the amount of time that these ancestral lineages have to coalesce (that is, T_2). The behavior of the topological concordance probability can be determined by considering several cases. Namely, standard coalescent simulation shows that if T_1 and T_2 are both small, topological concordance probability can be increased by enlarging samples. The increase in sample size needed for achieving a desired topological concordance probability depends on T_1 and T_2. The topological concordance is nearly guaranteed when T_2 large. If T_2 is small and T_3 is large, topological concordance is not likely; this result is little affected by sample size.

The deduction above presented was made by Rosenberg under simplify assumptions about equality and stability of population sizes and absence of population structure, in general, the effect of geographic structure within species is to decrease the topological concordance probability. Recent studies have investigated the probability distribution of random gene tree topologies under the multispecies coalescent Degnan and Rosenberg (2006), Degnan and Rosenberg (2009), Liu et al. (2009). In particular, treating a species tree as a parameter consisting of a fixed labeled topology and fixed branch lengths, Degnan and Salter (2005) obtained a probability distribution under the model for the labeled topology of a random gene tree evolving on the species tree, under a general method which implemented in the computer program COAL, which is available at

http://www.coaltree.net or by request from the authors. The gene tree topologies examined in the probability of distribution of Degnan and Salter are unranked, in that they consider only the topological relationship among gene lineages, and not the sequence in which the lineages coalesce. The additional information contained in the coalescence sequence or labeled history of a gene tree, however, can potentially lead to a novel method of summarizing gene tree distributions using ranked rather than unranked trees, thereby facilitating new approaches both in problems. Degnan et al. (2012) developed the analogous theory for a random ranked gene tree topologies, i.e. derived the probability distribution of ranked gene tree topologies conditional on a fixed species tree, considering both the topology and the sequence of coalescences for a random gene tree.

2.5. Another perspective

Assuming that gene and species trees evolve according to a coalescent process, we would like to compute the probability that these trees have the same topology. Unlike the multispecies coalescent, we do not assume that neither species nor tree species are fixed. In our aim, the strategy is to find all the posible species trees topologies τ^s with the same topology that a given gene tree topology τ^g and then take all the possible gene trees. Hence in contrast with previous studies, we consider τ^g in the set $\mathcal{T}$ of all posible topologies. Since phylogenetic trees are characterized by its coalescence times, we introduce T_i^s, the i-th coalescence time in a specie tree as well as T_i^g, for the gene tree. Recalling that coalescence times have exponential distribution with parameter say c, the computation of the probability of concordance for three genes is given by the following display. We use Figure 2.8 to show the two possible scenarios where we have species and gene trees topologically concordant, as well as, to determine the integration regions for the coalescence times.

$$
\begin{aligned}
\mathbb{P}(\tau^s = \tau^g) &= \sum_{\tau \epsilon \mathcal{T}} \sum_{i=1,2} \mathbb{P}(\tau_i^s = \tau^g | \tau^g) \mathbb{P}(\tau^g = \tau) \\
&= \int_0^T \int_0^{t_1^g} \int_0^{t_1^g} \int_{t_1^g}^{t_2^g} c^2 e^{ct_1^s + ct_2^s - ct_1^g - ct_2^g} dt_2^s dt_1^s dt_2^g dt_1^g \\
&\qquad + \int_0^T \int_0^{t_1^g} \int_0^{t_1^g} \int_{t_1^s}^{t_1^g} c^2 e^{ct_1^s + 2ct_2^s - 2ct_1^g - ct_2^g} dt_2^s dt_1^s dt_2^g dt_1^g \\
&= \frac{1}{36} e^{-3cT} (2 - 3e^{cT} - 24e^{4cT} + 18e^{2cT}(3 + 2cT) + e^{3cT}(-29 + 42cT + 18c^2T^2)),
\end{aligned}
$$

where T is the time to the most recent common ancestor. In the aim to find a recurrence relation for the desired probability we also consider the case with four genes, where we have two topologies for the gene trees $(((A, B), C)D)$ and $((A, B), (C, D))$. (Figure 2.9 illustrates examples of gene trees nested in the first topology). Once analyzed each

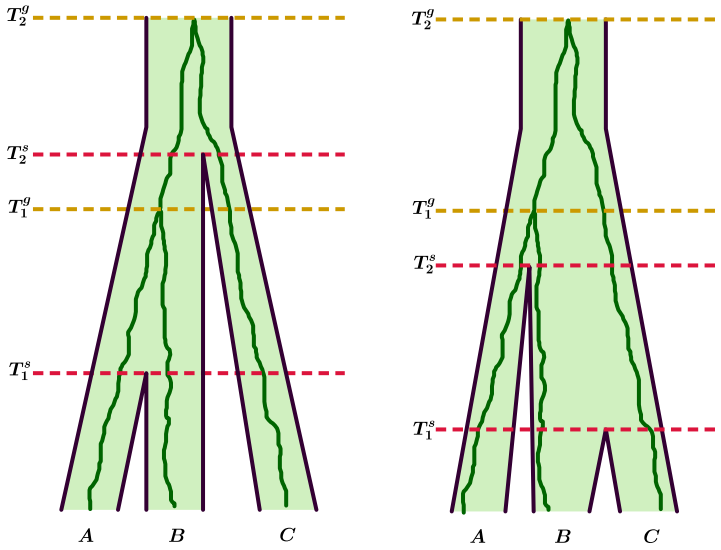

Figure 2.8: A three-taxon gene tree within the two equivalent species tree topologies

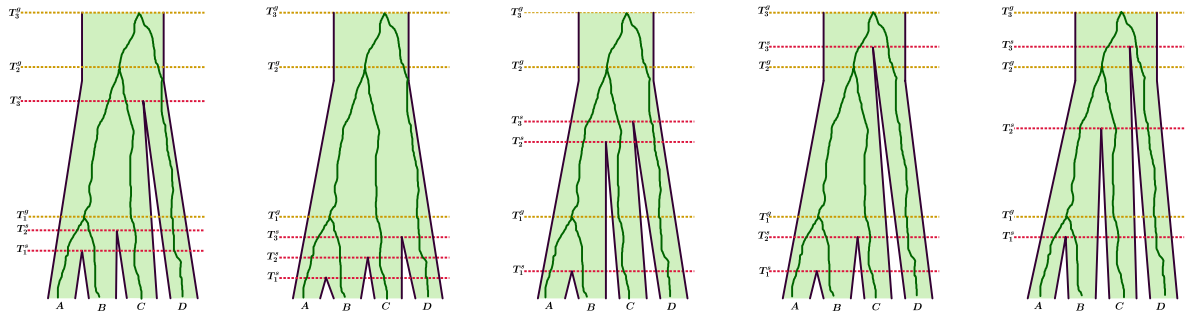

Figure 2.9: A four-taxon gene with the topology $(((A, B), C)D)$, sometimes called comb, within equivalent species tree.

topology, we use Mathematica to obtain:

$$\begin{aligned}\mathbb{P}(\tau_5^s = \tau^g) &= \sum_{i=1,2}\sum_{j=1}^{5}\mathbb{P}(\tau_j^s = \tau_i^g|\tau_i^g)\\ &= bc^{-3}\big[P_3(T) + P_2(T)e^{-1} + P_1(T)(e^{-3}+e^{-2}+e) + b_6e^{-6} + b_5e^{-5} + b_2e^2 + b_3e^3\big].\end{aligned}$$

where $P_i(T) = \sum_{k=1}^{i} b_{k_i}(cT)^k$ with b, b_k, b_{k_i} constants. Therefore we can assure that the probability of concordance in the situation studied satisfies a recurrence relation.

In the past 15 years, the relationship between gene trees and species tree has been greatly clarified. This conceptual advance has been accompanied by methodological developments in models of gene family evolution and in the algorithms needed for statistical inference. These rely heavily on coalescent and birth-death processes and dynamic programming. Alternatively, in the next chapter we propose a probabilistic model to study jointly the species and gene trees.

Chapter 3

Simple nested coalescent

This chapter is based on a work with Amaury Lambert and Arno Siri-Jégousse.

3.1. Background on coalescent processes

3.1.1. A short overview

In molecular biology, gene trees appear to provide a diagrammatical representation of evolutionary relationships. In this case the branching points are generated because a gene replicates, and its copies are passed on to more than one offspring. Therefore, one can think that the genes are the individuals of an haploid population which reproduces in non-overlapping discrete generations. The properties of gene trees have been investigated under various models of mathematical population genetics. Seminal works include: Ewens (1972), Watterson (1975), Kingman (1982b), Tajima (1983) and Hudson (1991). A key tool in this analysis is Kingman coalescent (Kingman (1982a)), which emerges as the scaling limit of the genealogical trees of populations described by the Cannings model. It is asymptotically related to Wright-Fisher diffusion which is commonly used to study the evolution of the frequencies of two types population.

More precisely, the Cannings model describes the dynamics of a haploid population with constant size equal to $N \geq 1$ where the individuals are randomly labeled as $\{1, ..., N\}$. Let ν_i^R represent the number of children of individual i of generation R. Define now $\nu^R = (\nu_1^R, ..., \nu_N^R)$ and suppose that $(\nu^R : R \geq 1)$ is a family of independent copies of a random variable $\nu = (\nu_1, ..., \nu_N)$. Assume that ν is exchangeable, i.e.

$$(\nu_1, ..., \nu_N) \stackrel{\mathcal{L}}{=} (\nu_{\sigma(1)}, ..., \nu_{\sigma(N)}),$$

for any permutation σ of $[n] := \{1, ..., N\}$. The latter hypothesis happens because there are no spatial effects or natural selection inside the population, so that every individual is treated equally.

Imagine we pick $n < N$ individuals at random without replacement from the present generation of a population of size N which is governed by the Cannings model. According to Kingman's formulation, looking at the genealogy backwards in time, the common ancestry of the sample is given through the concept of equivalence classes of $[n]$. More precisely, one identifies the individuals in the original sample with the trivial partition $(\{1\},\{2\},...,\{n\})$, and for any positive integer r by the equivalence relation: $i \overset{r}{\sim} j$ if and only if i and j have the same ancestor, r generations back in the past. Let us denote the collection of the equivalent classes generated by this relation by $R_r^{(N,n)}$. Obviously, $R_0^{(N,n)}$ is the trivial equivalence relation. Observe now that the probability that two individuals (chosen at random and without replacement from any generation) fall among the ν_1 individuals that make up the first family is just $\nu_1(\nu_1-1)/N(N-1)$. Now average over the distribution of the vector $(\nu_1,\nu_2,...,\nu_N)$. This gives the probability that both offsprings are in the first family. Using the exchangeability property we have that the probability that both belong to the same family is just N times this probability, that is

$$c_N := \mathbb{E}\left(\frac{\nu_1(\nu_1-1)}{N-1}\right).$$

In particular,

$$\mathbb{P}(1 \overset{1}{\sim} 2) = c_N.$$

and assuming that $c_N \to 0$ as $N \to \infty$,

$$\mathbb{P}(1 \overset{r}{\sim} 2) = 1-(1-c_N)^r \sim 1-e^{-rc_N}.$$

Moreover, notice that taking $r = \lfloor t/c_N \rfloor$, for all $t \geq 0$, we can infer that the right time scaling to obtain from $R^{(N,n)}$ a nontrivial continuous-time Markov chain as a limit process, as $N \to \infty$, is $1/c_N$.

Besides, the probability for three individuals of a given generation to share a common parent is

$$d_N = \mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\sim} 3) = \frac{\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]}{(N-1)(N-2)}.$$

Then taking

$$\frac{d_N}{c_N} \to 0, \quad N \to \infty, \tag{3.1}$$

we will only ever see pairwise mergers in the limit. The process obtained is the Kingman n-coalescent, sometimes referred simply as n-coalescent.

Theorem 3.1 (Möhle (2000))**.** *Consider a Cannings model defined by independent copies $(\nu_i : i = 1,...,N)$. For $t \geq 0$, denote $\mathcal{R}_t^{(N,n)} := R_{\lfloor t/c_N \rfloor}^{(N,n)}$. As $N \to \infty$, the following convergence in law*

$$\left(\mathcal{R}_t^{(N,n)} : t \geq 0\right) \to \left(\mathcal{R}_t^{(n)} : t \geq 0\right),$$

holds in the space of càdlàg process on $\mathbb{R}_+$ taking values in the space of partitions of $[n]$, if and only if $c_N \to 0$ and (3.1) holds.

To summarize, Kingman coalescent corresponds to the dynamics where each pair of blocks merges at rate 1. In particular, when there are n blocks present in the configuration, the total number of blocks decreases by 1 at rate $\binom{n}{2}$. We refer to Siri-Jégousse (2009) for details of the above construction of Kingman coalescent.

We remark that Kingman n-coalescent has a consistency property: the $(k+l)$-coalescent restricted to $[k]$ is a k-coalescent. This consistency property is an essential part from the viewpoint of modeling a sample from a very large population. The consistency property allowed Kingman (1982a) to take a projective limit and to define the Kingman coalescent valued on the equivalence relations of $\mathbb{N}$ with the property that, for each k, its restriction to $[k]$ is a k-coalescent. By convention, we take the initial state to be the trivial partition into singletons.

Coalescence theory was initially formulated in Kingman (1982b), with the purpose of describing the genealogy of haploid populations with binary reproduction, indeed, its genealogy can be approximated by this process. However it is not a suitable model for modeling the evolution of populations where a large proportion of individuals have the same parent. This situation corresponds for instance, to populations such that there is a large impact of selection: individuals who get a beneficial mutation will quickly recolonize an important fraction of the population, hence we will observe multiple collisions when tracing the ancestral lineages of individuals. Large variability in offspring distribution such as certain marine organisms also leads to the same property that many lineages may coalesce at once.

In the framework above, the processes known as Λ-coalescents were introduced by Sagitov (1999). Independently, Pitman (1999) defined these as processes with values among partitions of $\mathbb{N}$ and for each $n \in \mathbb{N}$. Similarly to Kingman coalescent their law can be described specifying the restriction to partitions of $[n]$. More precisely, a Λ-coalescent $(\Pi(t) : t \geq 0)$ is a Markov process with transitions probabilities determined by a σ-finite measure Λ on $[0,1]$, as follows: for all $b, j \in \mathbb{N}$ such that $2 \leq j \leq b$, if there are currently b blocks in $\Pi(t)$ then each transition involving j of the blocks merging into one happens at rate

$$\lambda_{b,j} = \int_{[0,1]} x^{j-2}(1-x)^{b-j}\Lambda(dx),$$

and these are the only possible transitions. Observe that if $\Lambda(\{0\}) = \Lambda([0,1]) > 0$, the only transitions are mergers of pairs of blocks, then the Λ-coalescent corresponds to the Kingman coalescent. Besides, if $\Lambda(\{0\}) = 0$, one may construct the Λ-coalescent through a Poisson point process

$$N(\cdot) = \sum_{i \in \mathbb{N}} \delta_{t_i, x_i}(\cdot)$$

on $\mathbb{R}_+ \times (0,1)$ with intensity $dt \otimes \nu(dx)$ where $\nu(dx) = x^{-2}\Lambda(dx)$, as follows. First, label independently each atom (t,x) with a sequence $\xi^{(t,x)} := (\xi_1^{(t,x)}, \xi_2^{(t,x)}, ...)$ of independent Bernoulli trials such that $P(\xi_i^{(t,x)} = 1) = x$, for all i. Next, given an arbitrary partition π of $\mathbb{N}$, let $\Pi_n(0)$ be the restriction of π to $[n]$, and let the process Π_n be allowed the

possibility of jumping only at the times t of points (t,x) of N such that $\sum_{i=1}^{n}\xi_i^{(t,x)} \geq 2$. For the times t in this set, if $\Pi_n(t-) = \{A_1, ..., A_b\}$ say, where the A_i are in the order of their least elements, let $\Pi_n(t)$ be derived from $\Pi_n(t-)$ by merging those A_i with i such that $\xi_i^{(t,x)} = 1$.

A large family of coalescent processes where infinitely many individuals coalesce and different merging may take place simultaneously, was first considered by Möhle and Sagitov (2001), and Schweinsberg (2000). Bertoin and Le Gall (2003) called this family exchangeable coalescents. Thereby the Λ-coalescent is a special class of exchangeable coalescent sometimes called simple exchangeable coalescent.

In other direction the notion of Λ-coalescent is extended to the spatial setting in Limic and Sturm (2006). Namely, the partition elements migrate in a geographical space and may only coalesce while sharing the same location, according to the mechanism of a multiple merger coalescent. More precisely, the model is defined as follows. Given a graph G and a set of particles:

i) Particles follow the trajectory of independent simple random walks in continuous time on $\mathbb{Z}^d$, with a fixed jump rate $\rho > 0$.

ii) Particles that are on the same sites coalesce according to the dynamics of a Λ-coalescent.

Observe that the Λ-coalescent corresponds to the setting with one vertex. Earlier works on variants of spatial coalescents, sometimes also referred to as structured coalescents, have all assumed Kingman coalescent-like behavior, and include Notohara (1990), Herbots (1997), and more recently Barton et al. (2004) in the case of finite initial configurations, and Greven et al. (2005) with infinite initial states. A related model has been studied by Zähle et al. (2005) on two-dimensional tori. Finally, spatial coalescents are related to coalescing random walks, the difference being that for coalescing random walks blocks coalesce instantaneously when they enter the same site.

Coalescence theory has been expanded because of the rising demand from population geneticists to develop and to analyze models which incorporate more realistic features. As we already mention in the previous chapter, molecular phylogeny has focused to build models describing the relationship between gene and species trees because it can improve the reconstruction of gene trees when a species tree is known, and viceversa. Motivated by this issue our goal is to develop a new class of coalescent processes such that gene lineages are allowed to coalesce while they are in the same branch of the species trees. Namely, in the following section we will define the simple nested coalescet, a Markov process with values in the space of nested bivariate partitions of $\mathbb{N}$.

3.1.2. Exchangeable coalescents

In order to define an exchangeable coalescent we firstly introduce the theory of exchangeable random partitions, which is a basic building block on our study. We refer to

Chapter 4 of Bertoin (2006) for further details.

Random partition

For every $n \in \bar{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$, a partition of $B \subset \mathbb{N}$ is a countable collection $(\pi_i : i \in \mathbb{N})$ of pairwise disjoint subsets of B, called *blocks*, such that

$$B = \bigcup_{i \in \mathbb{N}} \pi_i.$$

By convention, the blocks of π are ranked according to the increasing order of their least element, that is, for every $i \leq j$, $\min \pi_i \leq \min \pi_j$, with the convention of $\min \varnothing = \infty$. The number of nonempty blocks of π is denoted by $|\pi|$. The space of all partitions of $[n] = \{1, 2, ..., n\}$ is denoted by $\mathcal{P}_n$, where by convention $[\infty] := \mathbb{N}$ for $n = \infty$. Under the same convention, for every $n \in \bar{\mathbb{N}}$, $\mathbf{0}_n$ denotes the partition into singletons of $[n]$. For all $\pi \in \mathcal{P}_\infty$ and $n \in \mathbb{N}$, $\pi_{|n} \in \mathcal{P}_n$ is by definition the restriction of π to $[n]$.

According to the biological standpoint the blocks of a partition correspond to the labels of individuals in a population without selection, i.e. the individuals have the same reproductive capacities at every generation so, it must be possible to randomly reassign the labels without effect. Hence it will be convenient to consider exchangeable random partitions, those with distribution invariant under the action of permutations. To be formal, for every $n \in \mathbb{N}$, a permutation of $[n]$ is a bijection $\sigma : [n] \to [n]$; whereas a permutation of $\mathbb{N}$, is any bijection $\sigma : \mathbb{N} \to \mathbb{N}$ such that $\sigma(k) = k$ when k is large enough. Then for all $n \in \bar{\mathbb{N}}$, a random partition π of $[n]$ is *exchangeable* if for every permutation σ of $[n]$, π has the same law of a partition $\sigma(\pi)$, induced by the following relation

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

Observe that the blocks of $\sigma(\pi)$ are the images of those of π by the action of the inverse mapping σ^{-1}.

If π is an exchangeable partition, the asymptotic frequency

$$|B| := \lim_{n \to \infty} \frac{1}{n} |B \cap [n]| = \lim_{n \to \infty} \frac{1}{n} |\{k \in B : k \leq n\}|,$$

exists for every block B of π a.s. Moreover the law of π is completely determined by that of its ranked sequence of (nonzero) frequencies, which are exactly the random elements of the set of mass-partitions

$$\mathcal{S} := \{\mathbf{s} = (s_1, s_2, ...) : s_1 \geq s_2 \geq \cdots \geq 0 \text{ and } \sum_{i=1}^{\infty} s_i \leq 1\},$$

whose nonzero values. Conversely, given a probability distribution on $\mathcal{S}$, there exists an exchangeable partition of $\mathbb{N}$ whose ranked sequence of frequencies has the given distribution. This correspondence between the laws of exchangeable random partitions and

the laws of previous mass partitions is an important result about exchangeable partitions (Kingman (1978b)). Indeed, this bijection is induced by the so called paint-box, an exchangeable partition that may be constructed by Kingman's paintbox scheme (Kingman (1978a)), as follows.

Consider an element $s \in \mathcal{S}$ and the decomposition of the interval $[0,1]$,

$$A_1 = [0, s_1],\ A_i = \left(\sum_{j=1}^{i-1} s_j, \sum_{j=1}^{i} s_j\right),\ i \geq 1 \text{ and } A_0 = \left(\sum_{j=1}^{\infty} s_j, 1\right).$$

Let $(U_i : i \geq 1)$ be a sequence of independent random variables with uniform distribution on $[0,1]$, a s-paintbox is a partition π of $\mathbb{N}$ induced by the following equivalence relation:

- Integers i, j are in the same block of π if and only if $U_i, U_j \in A_k$ for some $k \in \mathbb{Z}_+$.

- If U_i belong to $A_0 = \left(\sum_{j=1}^{\infty} s_j, 1\right)$, then i is a singleton.

The law of a paint-box based on $\mathbf{s} \in \mathcal{S}$ will be denoted by $\varrho_{\mathbf{s}}$.

Coagulation of partitions

For all $k, k' \in \bar{\mathbb{N}}$ a pair of partitions $(\pi, \pi') \in \mathcal{P}_k \times \mathcal{P}_{k'}$ is called *admissible* if the number of non-empty blocks of π is $|\pi| \leq k'$. For every admissible pair of partitions (π, π'), the *coagulation operation* of π by π' denoted with $\mathrm{Coag}(\pi, \pi')$, results in a partition $\pi'' = (\pi''_j : j \in \mathbb{N})$ of $[k]$ given by

$$\pi''_j := \bigcup_{i \in \pi'_j} \pi_i, \quad j \in \mathbb{N}.$$

Plainly the partition into singletons $\mathbf{0}_\infty := (\{1\}, \{2\}, ...)$, is the neutral element for the coagulation operator, that is, for each partition π

$$\mathrm{Coag}(\pi, \mathbf{0}_\infty) = \mathrm{Coag}(\mathbf{0}_\infty, \pi) = \pi.$$

Since the labels of the blocks of a partition are assigned according to the order of their least element yields that for every $n \in \mathbb{N}$ and $\pi, \pi' \in \mathcal{P}_\infty$

$$\mathrm{Coag}(\pi, \pi')_{|n} = \mathrm{Coag}(\pi_{|n}, \pi') = \mathrm{Coag}(\pi_{|n}, \pi'_{|n}). \tag{3.2}$$

If π and π' are two independent exchangeable random partitions. Then the random partition $\mathrm{Coag}(\pi, \pi')$ is also exchangeable (see Lemma 3.2 in Bertoin (2006)).

Definition and important features

A Markov process $\Pi = (\Pi(t) : t \geq 0)$ continuous in probability with values in $\mathcal{P}_n$, $n \in \bar{\mathbb{N}}$, such that its semigroup can be described as follows. For every $t, t' \geq 0$, the conditional distribution of $\Pi(t+t')$ given $\Pi(t) = \pi$ is the law of $\mathrm{Coag}(\pi, \pi')$, where π' is some exchangeable random partition (whose law only depends on t'). Such a process Π is called an *exchangeable coalescent*. If additionally $\Pi(0) = \mathbf{0}_n$, we will say that Π is a *standard* exchangeable coalescent.

The Poisson construction for Λ-coalescents given in the previous section, leads naturally a Poisson construction for exchangeable coalescents processes (Proposition 4.5 of Bertoin (2006)). In this direction, let us consider μ, an exchangeable measure on $\mathcal{P}_\infty$ (invariant by the action of permutations) such that

$$\mu(\{\mathbf{0}_\infty\}) = 0 \quad \text{and} \quad \mu(\pi \in \mathcal{P}_\infty : \pi_{|n} \neq \mathbf{0}_n) < \infty, \quad \text{for every } n \in \mathbb{N},$$

and a Poisson point process M on $[0,\infty) \times \mathcal{P}_\infty$ with intensity $dt \otimes \mu(d\pi)$.

Explicitly we have the following construction of Π: let M_n be the image of M by the map $(t,\pi) \to (t, \pi_{|n})$, i.e. M_n is a Poisson random measure on $(0,\infty) \times \mathcal{P}_n$ with intensity $dt \otimes \mu_n(dx)$, where μ_n denotes the measure on $\mathcal{P}_n$, obtained as the image of μ by the restriction map $\pi \to \pi_{|n}$. Consider $\{(t_i, \pi^{(i)}), i \in \mathbb{N}\}$ the family of atoms of M_n on $(0,\infty) \times (\mathcal{P}_n \setminus \{\mathbf{0}_n\})$ ranked in increasing order of their first coordinate. We set $\Pi^{(n)}(t) = \mathbf{0}_n$ for $t \in (0, t_1)$, then define recursively

$$\Pi^{(n)}(t_i) = \mathrm{Coag}(\Pi^{(n)}(t_i-), \pi^{(n)}(t_i)), \quad \text{for every } t \in [t_i, t_{i+1}).$$

It is known (Theorem 4.2 of Bertoin (2006)) that μ can be characterized as follows,

$$\mu(d\pi) = c \sum_{1 \leq i < j} \delta_{K(i,j)}(d\pi) + \int_{\mathcal{S}} \varrho_{\mathbf{s}}(d\pi) \nu(ds),$$

where $\varrho_{\mathbf{s}}$ is the paint-box based on $\mathbf{s} \in \mathcal{S}$, $K(i,j)$ stands for the partition of $\mathbb{N}$ whose block consist of the pair $\{i,j\}$ and the singletons $\{k\}$ for $k \neq i, j$, $c \geq 0$ the coefficient of binary coagulation and ν a unique measure on $\mathcal{S}$ such that $\nu(\{\mathbf{0}\}) = 0$ and $\int_{\mathcal{S}} \left(\sum_{i=1}^{\infty} s_i^2\right) \nu(d\mathbf{s}) < \infty$, we stress that ν determines the multiple coagulations of the exchangeable coalescent.

3.2. Simple nested exchangeable coalescent

3.2.1. Nested partitions

Hereafter we consider the sub-family of simple partitions of $[n]$. A partition $\pi \in \mathcal{P}_n$ is called *simple* if and only if at most one of its blocks is neither empty nor reduced to a singleton. For $n \in \bar{\mathbb{N}}$, we denote the set of simple partitions of $[n]$ by $\mathcal{P}'_n$, that is,

$$\mathcal{P}'_n = \{\pi \in \mathcal{P}_n : \mathrm{Card}\{i : |\pi_i| > 1\} \leq 1\}.$$

For all $n \in \bar{\mathbb{N}}$. If $\mathcal{P}([n])$ denotes the space of all subsets of $[n]$ then we can define the function φ_n defined from $\mathcal{P}'_n$ to $\mathcal{P}([n])$ giving the only non singleton block of a simple partition, where by convention $\varphi_\infty = \varphi$ and $\varphi(\mathbf{0}_\infty) = \varnothing$.

Recalling 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; we next define a nested partition.

Definition 3.2. *Let $n \in \bar{\mathbb{N}}$ and π^g, π^s partitions of $[n]$, we will say that the partition π^g is nested in π^s or $\pi = (\pi^s, \pi^g)$ is nested, and we write $\pi^g \subseteq \pi^s$ when $i \overset{\pi^g}{\sim} j$ implies $i \overset{\pi^s}{\sim} j$.*

Example 3.3. An example of nested partition of $[10]$ is $\pi = (\pi^s, \pi^g)$, where

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

From now on, A^2 denotes the cartesian product of the set A. For every $n \in \bar{\mathbb{N}}$, $\mathcal{N}_n$ is the subset of $\mathcal{P}_n^2$ made of the nested partitions. A typical element of this set will be written by $\pi = (\pi^s, \pi^g)$. We use this notation in the aim to model the evolution of genes inside species in a population, namely π^s will represent the species partition and π^g will represent the gene partition. Since each specie have genetic information, in this work we only consider nested partitions $\pi = (\pi^s, \pi^g)$ such that $\text{Card}\{i : \pi_i^g \subseteq \pi_j^s\} \geq 1$, for all $j \leq |\pi^s|$.

The notation and properties of $\mathcal{P}_n$ can be naturally extended to the framework of bivariate partitions. For sake of completeness we specify here those that we will use constantly. The number of non-empty blocks of a partition $\pi = (\pi_1, \pi_2) \in \mathcal{P}_n^2$ is merely $|\pi| := (|\pi^s|, |\pi^g|)$. If $m_1, m_2 < n$, $\pi_{|m_1 \times m_2}$ denotes the restriction of π to $\mathcal{P}_{m_1} \times \mathcal{P}_{m_2}$, that is, $\pi_{|m_1 \times m_2} = (\pi^s \cap [m_1], \pi^g \cap [m_2])$. For the sake of a simple notation, we will write $\pi_{|\mathbf{m}}$ for the restriction of π to $\mathcal{P}_m^2$. A sequence $\pi^{(1)}, \pi^{(2)}, \ldots$ of elements of $\mathcal{P}_1^2, \mathcal{P}_2^2, \ldots$ 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 $\{\pi^{(n)} : \pi^{(n)} \in \mathcal{P}_n^2 \text{ and } n \in \mathbb{N}\}$ is consistent if and only if there exists $\pi \in \mathcal{P}_\infty^2$, such that $\pi_{|\mathbf{n}} = \pi^{(n)}$ for every $n \in \mathbb{N}$. In particular we have the notion of consistency for $\mathcal{N}_n$.

Remark 3.4. If π^g is nested in π^s then every block of π^g belongs to some block of π^s. One may identify the specie block in which is contained the i-th gene block by a function η called *nest*, that is, $\eta(\pi_i^g) = j$ if $\pi_i^g \subseteq \pi_j^s$.

Given a nested partition we use the coagulation operator to write the partition of the species in terms of the labels of genes partition, as we establish in the following proposition.

Proposition 3.5. *For every $n \in \bar{\mathbb{N}}$, let $\pi = (\pi^s, \pi^g)$ be an element of $\mathcal{N}_n$ and write $m = |\pi^g|$. There exists a unique partition $\bar{\pi} \in \mathcal{P}_m$ such that $\pi^s = Coag(\pi^g, \bar{\pi})$. In this sense, we shall say that a partition π is nested with link partition $\bar{\pi}$, or π is linked by $\bar{\pi}$.*

Proof. Define a partition $\bar{\pi}$ by

$$i \in \bar{\pi}_k \Leftrightarrow \eta(\pi_i^g) = k, \quad \text{for all } k \leq |\pi^s|, i \leq m.$$

According to this $\bar{\pi} \in \mathcal{P}_m$, and if $i, j \in \bar{\pi}_k$ then $\pi_i^g \cup \pi_j^g \subseteq \pi_k^s$, where $\eta(\pi_i^g) = \eta(\pi_j^g) = k$. Since π^g and π^s are ranked partitions of the same subset of $\mathbb{N}$, for every $k \leq |\pi^s|$,

$$\pi_k^s = \bigcup_{i \in \bar{\pi}_k} \pi_i^g. \tag{3.3}$$

This means that $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi})$. To get the uniqueness, suppose $\bar{\pi}' \in \mathcal{P}_m \backslash \{\bar{\pi}\}$ such that $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi}')$. The latter is equivalent to ask that

$$\pi_k^s = \bigcup_{i \in \bar{\pi}_k} \pi_i^g = \bigcup_{i \in \bar{\pi}'_k} \pi_i^g, \quad \text{for every } k \leq |\pi^s|.$$

Since $\bar{\pi} \neq \bar{\pi}'$, there exists one block $\bar{\pi}_k$ containing an element j such that $j \in \bar{\pi}_k$ and $j \notin \bar{\pi}'_k$, or there exists one block $\bar{\pi}'_k$ containing an element i such that $i \notin \bar{\pi}_k$ and $i \in \bar{\pi}'_k$. This is a clear contradiction as the latter equality cannot hold. ∎

To illustrate the above proposition observe that the nested partition defined in Example 3.3 has link partition $\bar{\pi} = (\{1,4\}, \{2,6,7\}, \{3,5\})$.

We can next get a partition of $\mathcal{P}_n \times \mathcal{P}_m$ through the coagulation of two pairs of admissible partitions. More precisely, for $n, n', m, m' \in \bar{\mathbb{N}}$, if $(\pi^1, \tilde{\pi}^1) \in \mathcal{P}_n \times \mathcal{P}_{n'}$ and $(\pi^2, \tilde{\pi}^2) \in \mathcal{P}_m \times \mathcal{P}_{m'}$ are admissible partitions, then $(\mathrm{Coag}(\pi^1, \tilde{\pi}^1), \mathrm{Coag}(\pi^2, \tilde{\pi}^2))$ is an element of $\mathcal{P}_n \times \mathcal{P}_m$. If we denote $\pi = (\pi^1, \pi^2)$ and $\tilde{\pi} = (\tilde{\pi}^1, \tilde{\pi}^2)$ we will say that the couple $(\pi, \tilde{\pi})$ is *admissible* and denote the latter operation by $\mathrm{Coag}_2(\pi, \tilde{\pi})$. In the following we will sometimes call the partition π as the *ingredients* and the partition $\tilde{\pi}$ as the *recipe*.

Lemma 3.6. *Let $(\pi, \tilde{\pi})$ be an admissible couple of bivariate partitions. Define the partition $\pi' = Coag_2(\pi, \tilde{\pi})$ and consider $\tilde{\pi}'$ such that the couple $(\pi', \tilde{\pi}')$ is admissible. Define now $\pi'' = Coag_2(\pi', \tilde{\pi}')$. Then $\pi'' = Coag_2(\pi, \tilde{\pi}'')$ where $\tilde{\pi}'' = Coag_2(\tilde{\pi}, \tilde{\pi}')$.*

Proof. Observe that is enough to prove the statement separately for species and genes coordinates with the operator Coag. To simplify the notation we shall omit the species and gene exponents. According to the definition of the coagulation operator, we have

$$\pi'_j = \bigcup_{i \in \tilde{\pi}_j} \pi_i, \quad \text{and} \quad \pi''_j = \bigcup_{i \in \tilde{\pi}'_j} \pi'_i.$$

Combining these latter equations to obtain

$$\pi''_j = \bigcup_{i \in \tilde{\pi}'_j} \bigcup_{k \in \tilde{\pi}_i} \pi_k = \bigcup_{k \in \tilde{\pi}''_j} \pi_k$$

where

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

∎

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, $\mathrm{Coag}_2(\pi, \tilde{\pi})$ is not necessarily nested. For instance, if we coagulate the partition π of Example 3.3, with the partition $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ where $\tilde{\pi}^s = (\{1,2\},\{3\})$, and $\tilde{\pi}^g = (\{1,3\},\{2\},\{4\},\{5\},\{6\},\{7\})$ then $\mathrm{Coag}(\pi^g, \tilde{\pi}^g)$ is not nested in $\mathrm{Coag}(\pi^s, \tilde{\pi}^s)$. 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 let us define the set

$$\tilde{\mathcal{P}}_{\infty,\infty}(\pi) = \left\{ \tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g) \in \mathcal{P}'_\infty \times \mathcal{P}'_\infty : \varphi(\tilde{\pi}^g) \subseteq \bigcup_{j \in \varphi(\tilde{\pi}^s)} \bar{\pi}_j \right\},$$

where as before φ is the function that allow us to obtain the non-singleton block in a simple partition. We will say that a two-dimensional partition $\tilde{\pi}$ is *conservative* for π, if for some n_1 and n_2 possibly infinity such that $n_1 \geq |\pi^s|, n_2 \geq |\pi^g|$ this pair of partitions takes values in the following set:

$$\tilde{\mathcal{P}}_{n_1,n_2}(\pi) = \left\{ \tilde{\pi}_{|n_1 \times n_2} = (\tilde{\pi}^s_{|n_1}, \tilde{\pi}^g_{|n_2}) \in \mathcal{P}'_{n_1} \times \mathcal{P}'_{n_2} : \tilde{\pi} \in \tilde{\mathcal{P}}_{\infty,\infty}(\pi) \right\}$$

In particular, two-dimension partitions $\tilde{\pi}$ will be called *strictly conservative* for π if they take values in the space

$$\tilde{\mathcal{P}}(\pi) := \tilde{\mathcal{P}}_{|\pi|}(\pi),$$

where we recall the notation $|\pi| = (|\pi^s|, |\pi^g|)$.

Remark 3.7. i) If π is in $\mathcal{N}_n$, for some n, and $\tilde{\pi} \in \tilde{\mathcal{P}}(\pi)$, then $\mathrm{Coag}_2(\pi, \tilde{\pi}) \in \mathcal{N}_n$.

ii) If $\pi = (\pi^s, \pi^g)$ with $|\pi^s| = \infty$ and $\tilde{\pi} \in \tilde{\mathcal{P}}(\pi) \backslash \mathbf{0}_\infty$, then $\mathrm{Coag}_2(\pi, \tilde{\pi}) = \pi'$ with $\pi'^{,s} \neq \pi^s$. Nevertheless, if $\pi \in \mathcal{N}_n$ and $\tilde{\pi} \in \tilde{\mathcal{P}}_{\infty,\infty}(\pi) \backslash \mathbf{0}_\infty$, it may happen that $\mathrm{Coag}_2(\pi, \tilde{\pi}) = \pi'$ with $\pi'^{,s} = \pi^s$. Take for instance $\tilde{\pi}$ such that $\inf \varphi(\tilde{\pi}^s) = |\pi^s|$, in this case $\tilde{\pi}^s|_{|\pi^s|} = \mathbf{0}_{|\pi^s|}$.

3.2.2. Simple nested coalescent

In the aim to describe genealogical trees of species and genes, we will define in this section a coagulation process with values in the nested partitions of $\mathcal{N}_\infty$, whose first coordinate describes the evolution of species and second coordinate describes evolution of genes. As a first rule, lineages of genes will be permitted to merge only when they are part of the same specie. However simultaneous coalescent events (species and genes) will be also permitted. This rule will be firstly expressed considering processes with values in $\mathcal{N}_n$.

Definition 3.8. *Fix $n \in \bar{\mathbb{N}}$, for every $t \geq 0$ let $\mathcal{R} := ((\mathcal{R}^s(t), \mathcal{R}^g(t)) : t \geq 0)$ be a Markov process with values in $\mathcal{P}^2_n$. This process is called a simple nested exchangeable coalescent, snec for short, if*

i) *For any $t \geq 0$, $\mathcal{R}^g(t)$ and $\mathcal{R}^s(t)$ are exchangeable random partitions.*

ii) *for any $t \geq 0$, $\mathcal{R}^g(t) \subseteq \mathcal{R}^s(t)$;*

iii) *The process $(\mathcal{R}^s(t) : t \geq 0)$ is a simple exchangeable coalescent process: for any $t, t' \geq 0$, the conditional distribution of $\mathcal{R}^s(t+t')$ given $\mathcal{R}^s(t)$ is the law of $Coag(\mathcal{R}^s(t), \tilde{\pi}^s)$ where $\tilde{\pi}^s$ is some simple exchangeable random partition independent of $\mathcal{R}^s(t)$, whose law just depends on t'.*

iv) *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 $\sigma(\tilde{\pi}^g) \overset{\mathcal{L}}{=} \tilde{\pi}^g$ for any permutation σ preserving $\bar{\mathcal{R}}(t)$ i.e.,*

$$i \overset{\bar{\mathcal{R}}(t)}{\sim} j \Rightarrow \sigma(i) \overset{\bar{\mathcal{R}}(t)}{\sim} \sigma(j). \tag{3.4}$$

In the framework of this definition, let us introduce the following notation and terminology. For fixed $n \in \bar{\mathbb{N}}$ and $\pi \in \mathcal{N}_n$ with link partition $\bar{\pi}$, a permutation σ of $[n]$ that satisfies the condition (3.4) will be said to *preserve* $\bar{\pi}$, writing this class of permutations $\Sigma(\bar{\pi})$. A random partition $\tilde{\pi}^g$ will be called *weakly exchangeable with respect to* $\bar{\pi}$ if its law is invariant under the action of any permutation $\sigma \in \Sigma(\bar{\pi})$, as in index *iv)* of Definition 3.8.

Remark 3.9. For the Example 3.3, the σ_1, σ_2 permutations defined as follows

$$\sigma_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 7 & 5 & 1 & 3 & 2 & 6 \end{pmatrix} \qquad \sigma_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 5 & 7 & 4 & 3 & 1 & 2 & 6 \end{pmatrix},$$

fulfill the condition (3.4) in the Definition 3.8.

An important observation is given in the following lemma.

Lemma 3.10. *For any $n \in \mathbb{N}$, let $\pi^1 = (\pi^{1,s}, \pi^{1,g})$ be a random variable of $\mathcal{N}_n$ linked by $\bar{\pi}_1$. Let $\tilde{\pi}^1 = (\tilde{\pi}^{1,s}, \tilde{\pi}^{1,g}) \in \tilde{\mathcal{P}}(\pi_1)$ such that $\tilde{\pi}^{1,g}$ is weakly exchangeable with respect to $\bar{\pi}^1$. Consider a random partition $\pi^2 = (\pi^{2,s}, \pi^{2,g})$ distributed as $Coag_2(\pi^1, \tilde{\pi}^1)$ and call $\bar{\pi}^2$ for the link partition of π^2. In a recursive way, let $\tilde{\pi}^2 = (\tilde{\pi}^{2,s}, \tilde{\pi}^{2,g})$ be a random partition in $\tilde{\mathcal{P}}(\pi^2)$ such that $\tilde{\pi}^{2,g}$ is weakly exchangeable with respect to $\bar{\pi}^2$ and define a random partition $\pi^3 = (\pi^{3,s}, \pi^{3,g})$ with the same law as $Coag_2(\pi^2, \tilde{\pi}^2)$. Then $\pi^3 \overset{\mathcal{L}}{=} Coag_2(\pi^1, \tilde{\pi}^3)$ where $\tilde{\pi}^3 \overset{\mathcal{L}}{=} Coag_2(\tilde{\pi}^1, \tilde{\pi}^2)$ and $\tilde{\pi}^{3,g}$ is weakly exchangeble with respect to $\bar{\pi}^1$.*

Proof. The fact that $\tilde{\pi}^3 = \mathrm{Coag}_2(\tilde{\pi}^1, \tilde{\pi}^2)$ has already been proved in Lemma 3.6. Now consider $\sigma \in \Sigma(\bar{\pi}^1)$, and observe that the blocks of $\sigma(\tilde{\pi}^{3,g})$ are the images of those of $\tilde{\pi}^{3,g}$ by the action of σ^{-1}:

$$\sigma^{-1}(\tilde{\pi}_j^{3,g}) = \sigma^{-1}\left(\bigcup_{i \in \tilde{\pi}_j^{2,g}} \tilde{\pi}_i^{1,g} \right) = \bigcup_{i \in \tilde{\pi}_j^{2,g}} \sigma^{-1}(\tilde{\pi}_i^{1,g}), \quad j \in \mathbb{N}.$$

The blocks $\sigma^{-1}(\tilde{\pi}_i^{1,g})$ form the partition $\sigma(\tilde{\pi}^{1,g})$ which is distributed as $\tilde{\pi}^{1,g}$, since $\tilde{\pi}^{1,g}$ is by hypothesis, weakly exchangeable with respect to $\bar{\pi}^1$. However, we can not conclude that $\sigma(\tilde{\pi}^{3,g}) = \mathrm{Coag}(\sigma(\tilde{\pi}^{1,g}), \tilde{\pi}^{2,g})$ as the permutation σ will in general affect the order of the blocks. Nevertheless, there exists a permutation σ' such that

$$\sigma(\tilde{\pi}^{3,g}) = \mathrm{Coag}(\sigma(\tilde{\pi}^{1,g}), \sigma'(\tilde{\pi}^{2,g})).$$

The partition σ' is necessarily in $\Sigma(\bar{\pi}^2)$ as the partition σ is in $\Sigma(\bar{\pi}^1)$ and species blocks of π^2 are obtained by merging species blocks of π^1. Since $\tilde{\pi}^{2,g}$ is weakly exchangeable with respect to $\bar{\pi}^2$, we can conclude that $\sigma(\tilde{\pi}^{3,g}) \overset{\mathcal{L}}{=} \mathrm{Coag}(\tilde{\pi}^{1,g}, \tilde{\pi}^{2,g}) = \tilde{\pi}^{3,g}$. ∎

To start the analysis of a snec we would like to make some observations related to Definition 3.8. First note that $\mathcal{R}$ is a $\mathcal{N}_n$-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 $\tilde{\pi} \in \tilde{\mathcal{P}}(\pi)$, hence the law of $\tilde{\pi}$ depends on t' but also on π.

From statement *iii)*, $(\mathcal{R}^s(t) : t \geq 0)$ is an exchangeable coalescent, however $(\mathcal{R}^g(t) : t \geq 0)$ is not a Markov process in general. This is due to statement (iv), making the distribution of $\mathcal{R}^g(t+t')$ depend on $\mathcal{R}^s(t)$.

We now turn to investigate the transitions of the restrictions of a snec to finite patitions, this relies in the following lemma, which is consequence of the consistency property of nested partition.

Lemma 3.11. *Let $\mathcal{R} = (\mathcal{R}(t) : t \geq 0)$ be a process with values in $\mathcal{N}_\infty$ and for every integer n, write $\mathcal{R}^{\mathbf{n}} = (\mathcal{R}^{\mathbf{n}}(t) : t \geq 0)$ for its restriction to $\mathcal{N}_n$, i.e., for any $t \geq 0$, $\mathcal{R}^{\mathbf{n}}(t) = \mathcal{R}_{|\mathbf{n}}(t)$. Then $\mathcal{R}$ is a snec in $\mathcal{N}_\infty$ if and only if $\mathcal{R}^{\mathbf{n}}$ is a snec in $\mathcal{N}_n$ for all $n \in \mathbb{N}$.*

Observe that for every $n \in \mathbb{N}$, $\mathcal{R}_{|\mathbf{n}}$ is a Markov chain. Then, its distribution is characterized by its jump rates. In this direction we denote the jump rate of $\mathcal{R}_{|n}$ from π to π' by

$$q_{\pi,\pi'} := \lim_{t \to 0+} \frac{1}{t} \mathbb{P}(\mathcal{R}_{|n}(t) = \pi' \,|\, \mathcal{R}_{|n}(0) = \pi).$$

As a first remark, $q_{\pi,\pi'}$ will be zero if π' is not obtained from π by coagulating according by a conservative partition. Next it is important to observe that if $\bar{\pi}$ is the link partition of π, i.e. $\pi^s = \mathrm{Coag}(\pi^g, \bar{\pi})$, and $\tilde{\pi}$ is the recipe to obtain π' from π, i.e. $\pi' = \mathrm{Coag}_2(\pi, \tilde{\pi})$, then for every $n_1 > n$, $q_{\pi,\pi'}$ is also the jump rate of $\mathcal{R}_{|\mathbf{n_1}}$ from $\pi_1 \in \mathcal{N}_{n_1}$ to π_1', whenever π_1 is linked by $\bar{\pi}$ and π_1' is obtained from π_1 by applying the recipe $\tilde{\pi}$. Indeed, the latter remark allows to ensure that we do not need to know the values of the gene blocks. This implies that the family of jump rates and hence the snec $\mathcal{R}$, is fully characterized by the family $(\tilde{q}_{\bar{\pi},\tilde{\pi}} : \bar{\pi} \in \mathcal{P}_\infty, \tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}), \rho_{|\mathbf{n}} = ((\bar{\pi}_{|n}, \mathbf{0}_n) : n \in \mathbb{N})$ where

$$\tilde{q}_{\bar{\pi},\tilde{\pi}} := q_{\rho_{|\mathbf{n}}, \mathrm{Coag}_2(\rho_{|n},\tilde{\pi})} = \lim_{t \to 0+} \frac{1}{t} \mathbb{P}(\mathcal{R}_{|\mathbf{n}}(t) = \mathrm{Coag}_2(\rho_{|\mathbf{n}}, \tilde{\pi}) \,|\, \mathcal{R}_{|\mathbf{n}}(0) = \rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n)). \quad (3.5)$$

Observe that it could appear more natural to consider $\bar{\pi} \in \mathcal{P}_n$ instead of $\mathcal{P}_\infty$ but this framework will make things easier in the next proof.

Now let $\bar{\pi} \in \mathcal{P}_\infty$, $\rho = (\bar{\pi}, \mathbf{0}_\infty)$ and fix $n \in \bar{\mathbb{N}}$. Our first goal is to represent the jump rates from a starting state $\rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n)$ by a single measure $\mu_{\bar{\pi}}$ (or μ_ρ) on $\tilde{\mathcal{P}}(\rho)$. To this end, if $\tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|n})$ and $n' \geq n$, let us define

$$\mathcal{P}_{n',\tilde{\pi}}(\rho) = \left\{ \tilde{\pi}' \in \tilde{\mathcal{P}}(\rho_{|n'}) : \tilde{\pi}'|_{|\rho_{|n}|} = \tilde{\pi} \text{ and } \forall n+1 \leq k \leq n', \tilde{\pi}'|_{|\rho_{|k}|} \in \tilde{\mathcal{P}}(\rho_{|k}) \right\}.$$

Proposition 3.12. *Let $(\tilde{q}_{\tilde{\pi},\tilde{\pi}} : \bar{\pi} \in \mathcal{P}_\infty, \tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}), \rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n), n \in \mathbb{N})$ be the family of jump rates of some snec $\mathcal{R}$. Then there exists a unique family $(\mu_{\bar{\pi}} : \bar{\pi} \in \mathcal{P}_\infty)$ where $\mu_{\bar{\pi}}$ is a measure on $\tilde{\mathcal{P}}(\bar{\pi})$ such that for any $\bar{\pi}$, $\mu_{\bar{\pi}}(\mathbf{0}_{\infty^2}) = 0$ and*

$$\mu_{\bar{\pi}}(\mathcal{P}_{\infty,\tilde{\pi}}(\rho)) = \tilde{q}_{\bar{\pi},\tilde{\pi}}.$$

Proof. First, we have the following identity

$$\mathcal{P}_{\infty,\tilde{\pi}}(\rho) = \bigcup_{\tilde{\pi}' \in \mathcal{P}_{n',\tilde{\pi}}(\rho)} \mathcal{P}_{\infty,\tilde{\pi}'}(\rho),$$

where the union is a union of disjoint subsets of $(\mathcal{P}'_\infty)^2$.

Besides, the Markov chain $\mathcal{R}_{|\mathbf{n}}$ can be obtained as the restriction of $\mathcal{R}_{|\mathbf{n}'}$ to $\mathcal{N}_n$, thus its jump rate $\tilde{q}_{\tilde{\pi},\tilde{\pi}}$ from $\rho_{|\mathbf{n}}$ to $\mathrm{Coag}_2(\rho_{|\mathbf{n}}, \tilde{\pi})$ coincides with the total jump rate for $\mathcal{R}_{|\mathbf{n}'}$ from $\rho_{|\mathbf{n}'}$ to $\tilde{\mathcal{P}}_{n',,\tilde{\pi}}(\rho)$. Therefore,

$$q_{\bar{\pi},\tilde{\pi}} = \sum_{\tilde{\pi}' \in \mathcal{P}_{n',\tilde{\pi}}(\rho)} q_{\bar{\pi},\tilde{\pi}'}.$$

The previous equality implies that the function

$$\mu_{\bar{\pi}} : \mathcal{P}_{\infty,\tilde{\pi}}(\rho) \to \tilde{q}_{\bar{\pi},\tilde{\pi}}$$

is additive. Since the class of finite unions of sets $\tilde{\mathcal{P}}_{\infty,\tilde{\pi}}(\rho)$ is a ring which generates the Borelian σ-field of $\tilde{\mathcal{P}}(\rho)$, an application of Caratheodory's extension theorem shows that μ_ρ has an unique extension on $\tilde{\mathcal{P}}(\rho)$. Moreover, according to definition of $\tilde{\mathcal{P}}(\rho)$, the partition $\mathbf{0}_{\infty^2}$ has μ_ρ-measure zero. ∎

3.2.3. An example of Poissonian construction

The goal of the present section is to give an illustrative example of a simple nested exchangeable coalescent constructed from a Poisson point process. In this aim consider a sequence of independent random variables $\zeta = (\zeta_i)_{i \in \mathbb{N}}$, with Bernoulli distribution of parameter $x \in (0,1)$. Let $(\xi^i_j)_{i,j \in \mathbb{N}}$ be an independent array of row wise independent Bernoulli random variables such that $\mathbb{P}(\xi^i_j = 1) = y$. Denote by $\mathcal{E}$ the set of matrices X

with entries $X_{ij} = (\zeta_i, \xi^i_j)$. The values of X_{ij} will determine if the j-th gene block of the i-th species will coalesce or not.

Given a partition $\pi \in \mathcal{N}_\infty$ linked by $\bar{\pi}$, we will need a different labelization of the gene blocks of π. For this define a function $l(k) = (s(k), g(k))$ applying on the indexes of the blocks of π^g. In words $l(k) = (i, j)$ if π^g_k is the j-th gene block of the i-th species block. Formally $k \in \bar{\pi}_{s(k)}$ and $g(k)$ is the position of k in the block $\bar{\pi}_{s(k)}$. Plainly, $l(\min \bar{\pi}_i) = (i, 1)$ and $l(\max \bar{\pi}_i) = (i, |\bar{\pi}_i|)$, for all $i \leq |\bar{\pi}|$. Moreover $l_1(k) = (\eta(\pi^g_k))$ where l_1 denotes the first coordinate of l and η is the nest function defined in Remark 3.4.

Let us now associate to $k \leq |\pi^g|$ the pair $(\zeta_{s(k)}, \xi^{s(k)}_{g(k)})$. A success of ζ_i and ζ_j means that species i and j merge whereas $\xi^{s(k)}_{g(k)} = \xi^{s(l)}_{g(l)} = 1$ means that genes k and l are available to coalesce. More precisely, from the sequence (ζ_i, ξ^i_j), we can obtain a simple random partition $\tilde{\pi} = (\tilde{\pi}^s, \tilde{\pi}^g)$ of $\mathbb{N}^2$ by the following equivalence relation:

$$\begin{aligned} i \overset{\tilde{\pi}^s}{\sim} j &\Leftrightarrow \zeta_i = \zeta_j = 1 \\ i \overset{\tilde{\pi}^g}{\sim} j &\Leftrightarrow (\zeta_{s(i)}, \xi^{s(i)}_{g(i)}) = (\zeta_{s(j)}, \xi^{s(j)}_{g(j)}) = (1, 1). \end{aligned} \tag{3.6}$$

Observe that as it is built the recipe partition $\tilde{\pi}$ is conservative with respect to π. Moreover, if $\rho_{|\mathbf{n}} = (\bar{\pi}_{|n}, \mathbf{0}_n)$, from any matrix of the set

$$E_{|\rho_{|\mathbf{n}}|} = \{X \in \mathcal{E} : \sum_{i=1}^{|\bar{\pi}_{|n}|} \zeta_i \geq 2 \text{ or } (\sum_{i=1}^{|\bar{\pi}_{|n}|} \zeta_i = 1 \text{ and } \sum_{i=1}^{|\bar{\pi}_{|n}|} \zeta_i \sum_{j=1}^{n} \xi^i_j \geq 2)\},$$

we will build a simple partition $\tilde{\pi} \in \tilde{\mathcal{P}}(\rho_{|\mathbf{n}}) \setminus \mathbf{0}_{|\rho_{|\mathbf{n}}|}$, which means that $\mathrm{Coag}_2(\rho_{|\mathbf{n}}, \tilde{\pi})$ is nested.

Denote by P_{xy} the distribution of the simple partition defined in (3.6). Consider a sigma-finite measure ν_{sg} on $[0,1]^2$ that satisfies

$$\nu_{sg}(\{(0,0)\}) = 0 \quad \text{and} \quad \int_{[0,1]} \int_{[0,1]} (x^2 + xy^2) \nu_{sg}(dx, dy) < \infty. \tag{3.7}$$

Define the following measure on $(\mathcal{P}'_\infty)^2$:

$$\varrho_{\nu_{sg}}(d\tilde{\pi}) = \int_{[0,1]} \int_{[0,1]} \nu_{sg}(dx, dy) P_{xy}(d\tilde{\pi}),$$

Notice that as a consequence of (3.6) we have that, for any $\rho = (\bar{\pi}, \mathbf{0}_\infty)$, $\varrho_{\nu_{sg}}((\mathcal{P}'_\infty)^2 \setminus \tilde{\mathcal{P}}(\rho)) = 0$ meanwhile as a consequence of (3.7) we have:

$$\varrho_{\nu_{sg}}(\tilde{\pi} \in \tilde{\mathcal{P}}(\rho) : \tilde{\pi}_{|\mathbf{n}} \neq \mathbf{0}_{|\rho_{|\mathbf{n}}|}) < \infty, \quad \text{for every } n \in \mathbb{N}.$$

With the latter measure we are ready to construct a snec process $\mathscr{R} = (\mathscr{R}(t) : t \geq 0)$.

To start, consider a partition $\pi \in \mathcal{N}_\infty$ linked by $\bar{\pi}$, define $\rho = (\bar{\pi}, \mathbf{0}_\infty)$ and for all $n \in \mathbb{N}$, $\mathscr{R}^{\mathbf{n}}(0) = \rho_{|\mathbf{n}}$. To determine the jumps of the process we will use the measure $\varrho_{\nu_{sg}}$

introduced above. Let M be a Poisson point process on $(0,\infty)\times(\mathcal{P}'_\infty)^2$ with intensity $dt\otimes\varrho_{\nu_{sg}}(d\tilde{\pi})$. For each $n\in\mathbb{N}$, M_n denotes the image of M by the map $(t,\tilde{\pi})\to(t,\tilde{\pi}_{|\mathbf{n}})$. So M_n is a Poisson measure on $(0,\infty)\times(\mathcal{P}'_n)^2$ with intensity $dt\otimes\varrho_n(d\tilde{\pi})$, where ϱ_n denotes the measure on $(\mathcal{P}'_\infty)^2$ obtained as the image of $\varrho_{\nu_{sg}}$ by the restriction map $\tilde{\pi}\to\tilde{\pi}_{|\mathbf{n}}$. Let $\{(t_i,\tilde{\pi}^{(i)}),i\in\mathbb{N}\}$ be the family of atoms of M_n on $(0,\infty)\times((\mathcal{P}'_\infty)^2\setminus\mathbf{0}_{|\rho_{|\mathbf{n}|}})$ ranked in increasing order of their first coordinate. We set $\mathscr{R}_{|\mathbf{n}}(t)=\rho_{|\mathbf{n}}$ for $t\in[0,t_1)$. Then define recursively

$$\mathscr{R}^{\mathbf{n}}(t_i)=\mathrm{Coag}_2(\mathscr{R}^{\mathbf{n}}(t_i-),\tilde{\pi}^{(i)}(t_i)),\quad \text{for every } t\in[t_i,t_{i+1}).$$

This process is consistent, indeed we prove.

Proposition 3.13. *For every $t\geq 0$, the sequence of random bivariate partitions $(\mathscr{R}^{\mathbf{n}}(t),n\in\mathbb{N})$ is consistent. If we denote by $\mathscr{R}(t)$ the unique partition of $\mathcal{N}_\infty$ such that $\mathscr{R}_{|\mathbf{n}}(t)=\mathscr{R}^{\mathbf{n}}(t)$ for every $n\in\mathbb{N}$, then the process $\mathscr{R}=(\mathscr{R}(t),t\geq 0)$ is a snec with jump rate $\mu=\varrho_{\nu_{sg}}$, started from $\rho=(\bar{\pi},\mathbf{0}_\infty)$. Notice that in this case the jump rate is not state-dependent.*

The proof uses similar arguments as in the proof of consistency of exchangeable coalescents given in Proposition 4.5 of Bertoin (2006), we include it here for sake of completeness.

Proof. Fix $n\geq 2$ and write $(t_1,\tilde{\pi}^{(1)})$ for the first atom of M_n on $(0,\infty)\times((\mathcal{P}'_n)^2\setminus\mathbf{0}_{|\rho_{|\mathbf{n}|}})$. Plainly, $\mathscr{R}^{\mathbf{n-1}}(t)=\mathscr{R}^{\mathbf{n}}_{|\mathbf{n-1}}(t)$ for every $t\in[0,t_1)$. Consider first the case when $\tilde{\pi}^{(1)}_{|\mathbf{n-1}}\neq\mathbf{0}_{|\rho_{|\mathbf{n-1}|}}$. Then $\tilde{\pi}^{(1)}_{|\mathbf{n-1}}$ is the first atom of M_{n-1} on $(0,\infty)\times((\mathcal{P}'_{n-1})^2\setminus\mathbf{0}_{|\rho_{|\mathbf{n-1}|}})$, and it follows from (3.2) that $\mathscr{R}^{\mathbf{n-1}}(t)=\mathscr{R}^{\mathbf{n}}_{|\mathbf{n-1}}(t)$ for every $t\in[t_1,t_2)$. Next, consider the case $\tilde{\pi}^{(1)}_{|\mathbf{n-1}}=\mathbf{0}_{|\rho_{|\mathbf{n-1}|}}$. Then M_{n-1} has no atoms on $[0,t_2)\times((\mathcal{P}'_{n-1})^2\setminus\mathbf{0}_{|\rho_{|\mathbf{n-1}|}})$, and it follows again from (3.2) that $\mathscr{R}^{\mathbf{n-1}}(t)=\mathscr{R}^{\mathbf{n}}_{|\mathbf{n-1}}(t)=(\bar{\pi}_{|n-1},\mathbf{0}_{n-1})$ for every $t\in[0,t_2)$. By iteration, this shows that the restriction of $\mathscr{R}^{\mathbf{n}}$ to $[n-1]^2$ coincides with $\mathscr{R}^{\mathbf{n-1}}$.

From this Poissonian construction $\mathscr{R}^{\mathbf{n}}$ is a Markov process, moreover it take values in $\mathcal{N}_n$ since it is defined by the operation Coag_2 with conservative partitions $\tilde{\pi}$. Is immediate also form the construction that $(\mathscr{R}^{\mathbf{n}})^s$ is a simple exchangeable coalescent process. Property *iv)* in Definition 3.8 is a consequence of Lemma 3.10. ∎

One of the most striking features of coalescent process, is the so called property of comes down from infinity. An exchangeable coalescent Π comes down from infinity if $|\Pi(t)|<\infty$ a.s. for every $t>0$. In this sense, we say that the simple nested exchangeable coalescent $\mathcal{R}=(\mathcal{R}^s,\mathcal{R}^g)$ CDI, if $\mathbb{P}(|\mathcal{R}^g(t)|<\infty,\ \text{for all } t>0)=1$.

Remark 3.14. If $\mathbb{P}(|\mathcal{R}^g(t)|<\infty,\ \text{for all } t>0)=1$ then $\mathbb{P}(|\mathcal{R}^s(t)|<\infty,\ \text{for all } t>0)=1$, because otherwise for some $t>0$, $\mathrm{Card}\{i:\mathcal{R}^g_i(t)\subseteq\mathcal{R}^s_j(t)\}=0$, for at least one $j\in\mathbb{N}$, which is imposible according with our framework.

In order to establish the necessary and sufficient conditions for the snec $\mathscr{R}$, constructed in previous section, consider the measures on $[0,1]$ defined as follows

$$u_s(dx)=\int_{y\in[0,1]}\nu_{sg}(dx,dy),\qquad u_g(dy)=\int_{x\in[0,1]}x\nu_{sg}(dx,dy). \tag{3.8}$$

From (3.7), we get

$$u_g(\{0\}) = u_s(\{0\}) = 0 \quad \text{and} \quad \int_{[0,1]} x^2 u_s(dx), \int_{[0,1]} y^2 u_s(dy) < \infty.$$

Using Lemma 4.5 of Bertoin (2006), there exists simple exchangeable coalescent processes that we will denote by Π^s and Π^g with coagulations measures u_s and u_g, respectively. Namely, $\Pi^s \overset{\mathcal{L}}{=} \mathscr{R}^s$.

Proposition 3.15. *The simple nested exchangeable coalescent $\mathscr{R}$ with jump rate $\mu = \varrho_{\nu_{sg}}$ CDI if and only if the simple exchangeable coalescent processes Π^s and Π^g, above defined, CDI.*

Proof. Suppose that $\mathscr{R}(0) = \mathbf{0}_\infty$ and assume that Π^s and Π^g come down from infinity. It is clear that $\mathscr{R}^s$ comes down from infinity. Therefore $\tau^s := \inf\{t \geq 0 : |\mathscr{R}^s(t)| < \infty\} = 0^+$ a.s.. This means that at any time $t > 0$ gene blocks are nested in a finite number of species blocks. Thus, to get that $\mathscr{R}$ comes down from infinity, we will prove that the time at which the number of gene blocks inside each species block is finite a.s is 0+. In this aim consider for every $i \leq |\mathscr{R}^s(\tau^s)|$, the process $\mathscr{R}^{g,i}$ defined for each $t > 0$ by $\mathscr{R}^{g,i}(t) = \{\mathscr{R}^g_j(t) : \mathscr{R}^g_j(t) \subseteq \mathscr{R}^s_i(t)\}$. Then for every i, the block-counting process of $\mathscr{R}^{g,i}$ is almost surely bounded from above by the block-counting process of a simple exchangeable coalescent with jump rates determined by the measure u_g. Then $\mathscr{R}^{g,i}$ comes down from infinity. Therefore $T^{g,i} := \inf\{t \geq 0 : |\mathscr{R}^{g,i}| < \infty\} = 0^+$ a.s. and $\mathscr{R}$ comes down from infinity because

$$\tau^g := \inf\{t \geq 0 : |\mathscr{R}^g(t)| < \infty\} = \max_{1 \leq i \leq |\mathscr{R}^s(\tau^s)|} T^{g,i}$$

Suppose now that Π^s or Π^g stays infinite. Observe that when Π^s stays infinite, $\mathscr{R}^s$ also stays infinite. From here, the hypothesis $\text{Card}\{i : \mathscr{R}^g_i(t) \subseteq \mathscr{R}^s_j(t)\} \geq 1$ for all j implies that $\mathscr{R}$ stays infinite. Assume now that Π^g stays infinite. Observe that $|\Pi^g|$ is almost surely bounded from above by $|\mathscr{R}^g|$ because there is no structure restriction for Π^g. This ends the proof. ∎

3.2.4. Future work

In the latter section we gave an example of a snec built using a poissonian construction. The interest of this construction is that it is entirely characterized by a measure on $\mathcal{P}^2_\infty$ and bot by a kernel as in Proposition 3.12. We would like to prove that any snec could be built with this type of construction. This would mean in particular that we could characterize this class of processes by a measure, independent of the state of the process at any time.

Finally, we recall that the motivation for introducing Kingman coalescent was to study the genealogy in the Wright-Fisher model in the limit when the size N of the population tends to ∞, and in the regime when one unit of time corresponds to N generations. The

Fleming-Viot process arises in the limit of rescaled Wright-Fisher in the same regime, and can be viewed in some sense (which has a rigorous mathematical interpretation) as the dual of Kingman coalescent. We would like to study infinite population models which are dual to a snec process. In this aim we should analyze the construction due to Donnelly and Kurtz (1999) of a population model whose genealogy can then be interpreted in terms of a simple exchangeable coalescent process. We can as well consider the point of view of Bertoin and Le Gall (2003), where a stochastic flow of bridges which encodes simultaneously an exchangeable coalescent process and a continuous population model. Partition flows introduced by Foucart (2012) also seem to be generalized. Another outline is to find some forward in time evolutionary models with (limit) genealogies being snec processes, as in Möhle and Sagitov (2001) where exchangeable coalescents arise as genealogies of Cannings models.

Appendix A

The space $D[0,\infty)$

Let $D[0,\infty)$ be the space of real-valued functions X on $[0,\infty)$ that are right-continuous and have left-hand limits, i.e.

i) $X_{t+} = \lim_{s\downarrow t} X_s$ exists,

ii) $X_{t+} = X_t$ for all $t \geq 0$,

iii) $X_{t-} = \lim_{s\uparrow t} X_s$ exists for all $t \geq 0$.

Functions having these properties are called càdlàg (a french acronym for 'continue à droite, limites à gauche'). Introducing the uniform metric on $D[0,\infty)$ by setting,

$$\|X\|_N = \sup_{s\leq N}\{|X_s|\}, \quad \text{for all } N \in \bar{\mathbb{N}} := \mathbb{N}\cup\{\infty\},$$

it becomes a Banach space but it is non-separable. This non-separability causes well-known problems of measurability in the theory of weak convergence of measures on the space. To overcome this inconvenience, A. Skorokhod Skorokhod (1956) introduced in his seminal paper four different topologies on $D[0,\infty)$; the so-called J_1, J_2, M_1 and M_2 topologies.

Being Jacod and Shiryaev (1987) and Whitt (2002) basic references, firstly we would like to introduce the metric d_{J_1} that generates the topology J_1, the most famous topology on $D[0,\infty)$. Although the original metric introduced by him has a drawback in the sense that the metric obtained is not complete, it turned out that it is possible to construct an equivalent metric d_{J_1} (i.e. giving the same topology) under which $D[0,\infty)$ becomes a Polish space. We start off with $C[0,\infty)$ the space of all continuous functions: $\mathbb{R}_+ \to \mathbb{R}$, a particularly important subspace of $D[0,\infty)$ nicely topologized by the local uniform metric,

$$d(X,Y) = \sum_{n=1}^{\infty} 2^{-n}(1 \wedge \|X-Y\|_n).$$

Namely, $X, Y \in C[0,\infty)$ are near one another in the uniform topology if the graph of X_t can be carried onto the graph of Y_t by a uniformly small perturbation of the ordinates, with the abscissa kept fixed. In contrast on $D[0,\infty)$, we would like allow also a uniformly small deformation of the time scale, so that, under a suitable metric the convergence $X^{(n)} \to X$ as $n \to \infty$ implies that the magnitudes and locations of the single jump of $X^{(n)}$ converge to those of X. Following this direction, the uniformly small deformation of the time scale will be represented by Λ, the set of all continuous functions $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ that are strictly increasing, with $\lambda_0 = 0$ and $\lambda_t \uparrow \infty$ as $t \to \infty$. So it is desirable that $d_{J_1}(X_n, X) \to 0$ as $n \to \infty$ if and only if there is a sequence $\{\lambda^{(n)}\} \subset \Lambda$ and the following conditions holds

$$\sup_s |\lambda_s^{(n)} - s| \to 0 \tag{A.1}$$

$$\sup_{s \le N} |X^{(n)} \circ \lambda_s^{(n)} - X_s| \quad \text{for all } n \in \bar{\mathbb{N}}. \tag{A.2}$$

In fact, we will requires that the time deformation λ that intervenes in (A.1) be near to the identity function in a sense more stringent, namely we need that the slope $(\lambda_t - \lambda_s)/(t-s)$ be nearly 1 or, what is the same thing and analytically more convenient that its logarithm be nearly 0. Hence for all $\lambda \in \Lambda$ we set

$$|||\lambda||| = \sup_{s<t} \left| \log \frac{\lambda_t - \lambda_s}{t-s} \right|.$$

The convergence $d_{J_1}(X_n, X) \to 0$ as $n \to \infty$ will be reduced to convergence of the restrictions of the functions involved to each finite interval $[0, N]$, in the sense of the J_1 metric on $D[0,N]$. Since projection maps are not automatically continuous for J_1 metric we need to avoid those values of N at which X has positive probability of jumping. To consider the restriction of $X^{(n)}$ to $[0,N]$, with the potential discontinuity at N smoothed out, for each $N \in \bar{\mathbb{N}}$ define the following function

$$k_N(t) = \begin{cases} 1 & \text{if } t \le N \\ N+1-t & \text{if } N < t < N+1 \\ 0 & \text{if } t \ge N+1, \end{cases}$$

and assume that $k_N X$ is the product of the real-valued function k_N with the $\mathbb{R}$-valued function X. Finally, for $X, Y \in D$ we set

$$d_{J_1}(X,Y) = \sum_{n=1}^{\infty} 2^{-n} (1 \wedge d_n(X,Y)), \tag{A.3}$$

where

$$d_N(X,Y) = \inf_{\lambda \in \Lambda} (|||\lambda||| + \|(k_N X) \circ \lambda - k_N Y\|_\infty), \quad N \in \bar{\mathbb{N}}. \tag{A.4}$$

The Polish space $D[0,\infty)$ is called the Skorokhod space (cf. also Skorokhod topology). Although the topology J_1 became the most famous on $D[0,\infty)$ is not the only reasonable,

in fact at the end of the 1980's it was found that in certain problems the other topologies introduced by Skorokhod in the space of càdlàg functions can be useful. In the remainder we will introduce the metrics that generates the J_2 and M_1, M_2 topologies.

Firstly we replace the set of function Λ in (A.4) by the larger set Λ' of all one-to-one maps of $\mathbb{R}_+$ onto $\mathbb{R}$ without requiring any continuity, so that

$$d_{J_2}^N(X,Y) = \inf_{\lambda\in\Lambda'}(|||\lambda||| + \|(k_N X)\circ\lambda - k_N Y\|_\infty), \quad N \in \bar{\mathbb{N}},$$

induce the J_2 topology on $D[0,\infty)$ as well d_N. Since $\Lambda \subset \Lambda'$, we obviously have

$$d_{J_2}(X,Y) \le d_{J_1}(X,Y), \quad \text{for all } X, Y \in D[0,\infty).$$

Besides, M_i-topologies on $D[0,T]$ are generated by metrics d_{M_i}, defined by means of completed graphs. For $X \in D[0,T]$ the completed graph of X is the set of point (z,t) where z belongs to the segment $[X(t-), X(t)]$, that is,

$$\Gamma_X := \{(z,t) \in \mathbb{R}\times[0,T] : z = aX_{t-} + (1-a)X_t \text{ for some } a \in [0,1]\}.$$

A (total) order on the graph Γ_X is established saying that $(z_1,t_1) \le (z_2,t_2)$ if either i)$t_1 < t_2$ or ii)$t_1 = t_2$ and $|X_{t_1-} - z_1| \le |X_{t_2-} - z_2|$. A parametric representation of the completed graph Γ_X (or of the function X) is a continuous function (u,r) mapping $[0,1]$ onto Γ_X with u being the spatial component and r being the time component. Let $\Pi(X)$ denote the set of nondecreasing, using the order above, parametric representation of X in $D[0,T]$.

For $X_1, X_2 \in D[0,T]$ we define the M_1 metric as

$$d_{M_1}(X_1,X_2) := \inf_{(u_j,r_j)\in\Pi(X_j)} \{\|u_1 - u_2\|_{[0,T]} \vee \|r_1 - r_2\|_{[0,T]}\}. \tag{A.5}$$

Notice that $\|u_1 - u_2\|_{[0,T]} \vee \|r_1 - r_2\|_{[0,T]}$ can also be written as $\|(u_1,r_1) - (u_2,r_2)\|_{[0,T]}$ where

$$\begin{aligned}\|(u_1,r_1) - (u_2,r_2)\|_{[0,T]} &:= \sup_{t\in[0,T]} \{|(u_1(t),r_1(t)) - (u_2(t),r_2(t))|\} \\ &= \sup_{t\in[0,T]} \{|u_1(t) - u_2(t)| \vee |r_1(t) - r_2(t)|\}.\end{aligned}$$

Then we have the following equivalent expression for the M_1-metric

$$d_{M_1}(X_1,X_2) = \inf_{(u_j,r_j)\in\Pi(x_j)} \left\{ \sup_{t\in[0,T]} \{|u_1(t) - u_2(t)| \vee |r_1(t) - r_2(t)|\} \right\}.$$

This metric induces the M_1 topology which is weaker that the J_1 topology. One of the advantages of the M_1 topology is that it allows for a jump in the limit function $X \in D$ to be approached by multiple jumps in the converging functions $X^{(n)} \in D$. To specify, a sequence $X_t^{(n)} \in D$ converges to $X_t \in D$ in the M_1- topology if

$$\lim_{n\to\infty} d_{M_1}(X^{(n)}, X) = 0.$$

In other words (see Theorem 12.5.1 of Whitt (2002)), we have the convergence in M_1 topology if there exist parametric representations (u,r) of the graph Γ_X and (u_n,r_n) of the graph $\Gamma_{X^{(n)}}$ such that

$$\lim_{n\to\infty} \|(u_n,r_n)-(u,r)\|_{[0,T]} = 0.$$

We are interested in the function space $D[0,\infty)$ with domain $[0,\infty)$ instead of the compact domain $[0,T]$. In that setting, let $r_t : D[0,\infty) \to D[0,t]$ be the restriction map with $r_t(X)(s) = X(s)$, $0 \le s \le t$. Suppose that $f : D[0,\infty) \to D[0,\infty)$ and $f : D[0,t] \to D[0,t]$ for $t > 0$ are functions with

$$f_t(r_t(X)) = r_t(f(X))$$

for all $X \in D[0,\infty)$ and all $t > 0$. We then call the functions f_t restrictions of the function f. From Teorema 12.9.1 of Whitt (2002), f is continuous in the topology of $D[0,\infty)$ if f has continuous restrictions f_t for all $t > 0$.

We now consider the extension of Lipschitz properties to subsets of $D[0,\infty)$. For this purpose, suppose that $d_{M_1,t}$ is the M_1-metric on $D[0,t]$ for $t > 0$. An associated metric $d_{M_1,\infty}$ on $D[0,\infty)$ can be defined by

$$d_{M_1,\infty}(X_1,X_2) = \int_0^\infty e^{-t}(d_{M_1,t}(r_t(X_1),r_t(X_2)) \wedge 1)dt.$$

The above integral is well defined (see Theorem 12.9.2) and finally we conclude with the following characterization of M_1 characterization of M_1 convergence in the domain $[0,\infty)$.

Theorem A.1 (Whitt (2002), Theorem 12.9.3)**.** *Suppose that $d_{M_1,\infty}$ and $d_{M_1,t}$, $t > 0$ are the M_1-metrics on $D[0,\infty)$ and $D[0,t]$, respectively. Then the following are equivalent for X and $X^{(n)}$, $n \ge 1$, in $D[0,\infty)$.*

i) $d_{M_1,\infty}(X^{(n)},X) \to 0$ *as* $n \to \infty$.

ii) $d_{M_1,t}(r_t(X^{(n)},r_t(X))) \to 0$ *as* $n \to \infty$ *for all* $t \notin Disc(X)$*, with*

$$Disc(X) := \{t \in [0,T] : X_{t-} \neq X_t\},$$

denoting the set of discontinuities of X;

iii) there exist parametric representations (u,r) and (u_n,r_n) of X and X_n mapping $[0,\infty)$ into the graphs such that

$$\|u_n - u\|_t \vee \|r_n - r\|_t \to 0 \quad as \quad n \to \infty,$$

for each $t > 0$.

Finally we have the topology M_2, usually this arises from a comparison of the completed graphs of paths by means of the Hausdorff distance, defined for K_1 and K_2 compact subsets of $\mathbb{R}$ as follows

$$|K_1 - K_2|_H := \sup_{x_1 \in K_1} |x_1 - K_2| \vee \sup_{x_2 - K_2} |x_2, K_1|,$$

where $|x - A|$ is the distance between the point x and the set A, that is

$$|x - A| \equiv |A - x| \equiv \inf_{y \in A} |x - y|$$

The M_2 metric on D is defined by

$$d_{M_2}(X_1, X_2) := |\Gamma_{X_1} - \Gamma_{X_2}|, \quad X_1, X_2 \in D[0,1],$$

where Γ_X is the completed graph of X before defined. A unified approach to the four Skorokhod topologies via graphs was provided in the thesis of Pomarede (1976). In that approach, the M_2 and J_2 topologies are generated by the Hausdorff metric applied to the completed and uncompleted graphs, respectively. Similarly, the M_1 and J_1 topologies are defined in terms of parametric representations of the completed and uncompleted graphs. Instead to present the characterization of topologies given by Pomarede to compare the topologies M, let us introduce the set $\Pi'(X)$ of parametric representation of X in $D[0,T]$ such that the time component function r be nondecreasing, now we replace $\Pi(X)$ by $\Pi'(X)$ in (A.5) to have an alternative equivalent characterization for the M_2 topology for which is evident that the M_1 topology is stronger than the $M2$ topology.

To summarize, the four non-uniform Skorokhod topologies are ordered by

$$J_1 > J_2 > M_2 \quad \text{and} \quad J_1 > M_1 > M_2$$

where $>$ means stronger than, with M_1 and J_2 not being comparable.

Consider now the following examples. Let $X_t = \mathbf{1}_{[1/2,1]}$ be the limit function. Then the sequence containing $X_t^{(n)} = \mathbf{1}_{[1/2+1/n,1]}$ converges J_1. Hovever, $X_t^{(n)} = \mathbf{1}_{[1/2-1/n,1/2)\cup[1/2+1/n,1]}$ converges J_2 but neither J_1 nor M_1. On the other hand, $X_t^{(n)} = \frac{1}{2}\mathbf{1}_{[1/2,1/2+1/n)]} + \mathbf{1}_{\cup[1/2+1/n,1]}$ converges M_1 but neither J_1 nor J_2. Finally $X_t^{(n)} = \frac{2}{3}\mathbf{1}_{[1/2-1/n,1/2)} + \frac{1}{3}\mathbf{1}_{[1/2,1/2+1/n)} + \mathbf{1}_{[1/2+1/n,1]}$ converges M_2 but neither J_2 nor M_1.

A.1. First passage time

Let $\tau_z(Y)$ the first passage time beyond z for $Y \in D$, a real-valued function closely related to the inverse function, so that

$$\tau_z(Y) = \inf\{t \geq 0 : Y_t > z\}.$$

The following continuity property is a key in our development.

Theorem A.2 (Whitt (2002), Theorem 13.6.4)**.** *Suppose that D_u is the subset of Y in $D[0,\infty)$ that are unbounded above and satisfy $Y(0) \geq 0$. If $\{Y^{(n)} : n \in \mathbb{N}\}$ is a sequence of functions on $D[0,\infty)$ such that in the sense of the Skorokhod topology M_2, $Y^{(n)}$ converges towards Y, an element of the following set*

$$D_{\tau_z} = \{Y \in D_u : Y(t) \neq z, \text{ where } t \in (\tau_z(Y) - \epsilon, \tau_z(Y)) \text{ for } \epsilon > 0 \text{ arbitrary}\}.$$

Then

$$G_z(Y^{(n)}) \xrightarrow[n\to\infty]{} G_z(Y) \quad \text{in } \mathbb{R}^4, \tag{A.6}$$

where

$$G_z(Y) = (\tau_z(Y), Y_{\tau_z(y)-}, Y_{\tau_z(y)}, Y_{\tau_z(y)} - Y_{\tau_z(y)-}).$$

Proof. The claimed convergence is a consequence of the continuity of the first passage time function mapping. This last assertion is given by Theorem 7.1 of Whitt (1980), we include the proof here for sake of completes. The supremum function, mapping $D[0,\infty)$ into itself according to

$$\mathcal{S}(X)_t = \sup_{0\leq s\leq t} X_s^{(n)}, \quad t \geq 0,$$

is continuos under the J_1 topology because

$$\sup_{0\leq t\leq m} \left| \sup_{0\leq s\leq m} X_s^n - \sup_{0\leq s\leq m} X(\lambda_s^n) \right| \leq \sup_{0\leq t\leq m} |X_t^n - X(\lambda_s^n)|,$$

for all m and n, hence the supremum is continuous in M_2. Besides, note that $(\mathcal{S}(X)(s), t(s))$ serves as a parametric representation for $\tau(\mathcal{S}(X))$ as well as $\mathcal{S}(X)$ when the roles of $\mathcal{S}(X)$ and t are switched because $\mathcal{S}(X)$ is non-decreasing. Hence $\tau(\mathcal{S}(X^n)) \xrightarrow[n\to\infty]{M_2} \tau(\mathcal{S}(X))$ if $\mathcal{S}(X^n) \xrightarrow[n\to\infty]{M_2} \mathcal{S}(X)$. Since $\mathcal{S}$ is M_2-continuous, $\tau(\mathcal{S}(X^n)) \xrightarrow[n\to\infty]{M_2} \tau(\mathcal{S}(X))$ if $X^n \xrightarrow[n\to\infty]{M_2} X$. This latter implies that τ is M_2-continuous because $\tau(\mathcal{S}) = \tau$. ∎

We know turn to the prove of lemma that we will use to establish Lemma 1.19

Lemma A.3. *Let $\{H^{(n)} : n \in \mathbb{N}\}$ be a sequence on $D[0,\infty)$ such that*

$$H^{(n)} \xrightarrow[n\to\infty]{J_1} I, \tag{A.7}$$

$$(Y^{(n)}, H^{(n)}) \xrightarrow[n\to\infty]{J_1} (Y, I), \tag{A.8}$$

where I is the identity on $D[0,\infty)$, i.e. $I_t = t$ for $t \geq 0$ and $Y \in D_{\tau_z}$. Then the following convergence holds

$$(G_z(Y^{(n)}), H^{(n)}(\tau_z(Y^{(n)}))) \xrightarrow[n\to\infty]{} (G_z(Y), I(\tau_z(Y))), \tag{A.9}$$

where G_z is the function defined in (A.2).

Proof. The convergence (A.6) implies that for every $\epsilon > 0$ there exist $\delta > 0$ and $N \in \mathbb{N}$ such that if $d_{J_1}((Y^{(n)}, H^{(n)}), (Y, I)) < \delta$ then

$$|G_z(Y^{(n)}) - G_z(Y)|_{\mathbb{R}^4} < \epsilon/2, \quad \text{for all } n \geq N.$$

Besides according to the definition of the metric d_{J_1} given in (A.3), we have from (A.7) that $d_N(H^{(n)}, I) \to 0$, if $\tau_z(Y^{(n)}) < N$. Then

$$d_{J_1}(H^{(n)}(\tau_z(Y^{(n)})), I(\tau_z(Y))) < \epsilon/2 \quad \text{for all } n \geq N,$$

from here we get

$$d_{J_1}(H(\tau_z(Y^{(n)})), I(\tau_z(Y)) \xrightarrow[n\to\infty]{} 0 \quad \text{for } \tau_z(Y^{(n)}) < n.$$

■

Appendix B

Proof of Lemma 1.13

A consequence of (1.19) is that the measure defined on $[0,\infty)$ by

$$\mu(x) \coloneqq \int_0^x z\bar{\pi}^+(z)dz, \quad x \geq 0. \tag{B.1}$$

is such that $x \mapsto \mu(x)$ is $RV_\infty^{2-\alpha}$. Then from the Tauberian-Abelian Theorem (see Theorem 1.7.1 in Bingham et al. (1987)), its Laplace transform $\mathcal{L}_\mu \in RV_0^{-(2-\alpha)}$ and

$$\mu(x) \sim \frac{1}{\Gamma(3-\alpha)}\mathcal{L}_\mu\left(1/x\right), \quad x \to \infty.$$

Observe that

$$\lambda^2\mathcal{L}_\mu(\lambda) = \mathbb{E}\left(1 - e^{-\lambda\xi^+} - \lambda\xi^+e^{-\lambda\xi^+}\right), \quad \lambda \to 0. \tag{B.2}$$

As consequence of the definition of the measure μ and the approximations above,

$$\bar{\pi}^+\left(1/\lambda\right) \sim c_\alpha\mathbb{E}\left(1 - e^{-\lambda\xi^+} - \lambda\xi^+e^{-\lambda\xi^+}\right), \quad \lambda \to 0, \tag{B.3}$$

where $c_\alpha = 1/\Gamma(3-\alpha)$. Hence for all $x > 0$,

$$\frac{1}{\mathbb{E}\left(1 - e^{-\lambda\xi^+} - \lambda\xi^+e^{-\lambda\xi^+}\right)}\bar{\pi}^+\left(x/\lambda\right) \sim \frac{\bar{\pi}^+\left(x/\lambda\right)}{c_\alpha\bar{\pi}^+\left(1/\lambda\right)} \xrightarrow[\lambda\to 0]{} c_\alpha x^{-\alpha}. \tag{B.4}$$

We set $r(n) = \left(\mathbb{E}\left[1 - e^{-\xi^+/n} - \xi^+e^{-\xi^+/n}/n\right]\right)^{-1}$ and define the measure $m_n(dy) = r(n)\bar{\pi}^+(ndy)$ on $(0,\infty)$. The convergence in (B.4) implies

$$m_n(x,\infty) \xrightarrow[n\to\infty]{} \int_x^\infty \frac{c_\alpha}{\alpha}\frac{dy}{y^{1+\alpha}}, \quad \text{for all } x > 0.$$

Therefore, for all $0 < x \leq y \leq \infty$

$$m_n(x,y] \xrightarrow[n\to\infty]{} \int_x^y \frac{c_\alpha}{\alpha}\frac{dz}{z^{1+\alpha}}.$$

This implies that the measure on $(0,\infty)$ defined by $m_n(dy) = r(n)\bar{\pi}^+(ndy)$ converges vaguely towards $c_\alpha \frac{dy}{y^{1+\alpha}}$. We also have

$$\int y^2 \mathbb{1}_{\{y\leq x\}} r(n)\pi^+(ndy) \xrightarrow[n\to\infty]{} c_\alpha \int y^2 \mathbb{1}_{\{y\leq x\}} \frac{dy}{y^{1+\alpha}}.$$

Using an argument of monote class to deduce the above convergence over $I \subset (0,\infty)$. Thus we obtain the convergence of the Laplace transform of the measure μ. This complete the proof because μ is regularly varying at infinity with indice $2-\alpha$ and its Laplace transform satisfies the identity (B.2).

Appendix C

Proof of Proposition 1.16

We prove the statement for clones, the mutants case is similar. First note that in the same way as in the proof above,

$$\mu^{cn}(x) = \int_0^x s\bar{\pi}^{cn}(s)ds, \quad x \geq 0, \tag{C.1}$$

is a measure on $[0,\infty)$ with Laplace transform $\mathcal{L}_{\mu^{cn}}$ such that

$$\lambda^2\mathcal{L}_{\mu^{cn}}(\lambda) = \mathbb{E}\left(1 - e^{-\lambda\xi^{(cn)}} - \lambda\xi^{(cn)}e^{-\lambda\xi^{(cn)}}\right), \quad \lambda \geq 0.$$

We now replace λ by a sequence $\{\lambda(n) : n \geq 0\}$ such that $\lambda(n) \to 0$ as $n \to \infty$,

$$\lambda(n)^2\mathcal{L}_{\mu^{cn}}(\lambda(n)) = 1 - \phi_n^c(\lambda(n)) + \lambda(n)(\phi_n^c)'(\lambda(n)).$$

From (1.36) we have an estimate of the term ϕ_n^c. In order to estimate $(\phi_n^c)'$ we use the fact that for every fixed n, conditionally to $\xi^{(+)} = k$ the distribution of $\xi^{(cn)}$ is Binomial with parameter $(k, 1-p(n))$. Then we apply the same techniques as in Lemma 1.14 to get the following estimate

$$(\phi_n^c)'(\lambda(n)) \sim (1-p(n))e^{-\lambda(n)}\phi^{+\prime}((1-p(n))(1-e^{-\lambda(n)})), \quad n \to \infty.$$

Putting both estimates together we infer as $n \to \infty$.

$$\begin{aligned}\lambda(n)^2\mathcal{L}_{\mu^{cn}}(\lambda(n)) \sim\ & 1 - \phi^+((1-p(n))(1-e^{-\lambda(n)})) \\ & + (1-p(n))(1-e^{-\lambda(n)})\phi_n^{+\prime}((1-p(n))(1-e^{-\lambda(n)})) \\ & - (1-p(n))(1-e^{-\lambda(n)} - \lambda(n)e^{-\lambda(n)})\phi_n^{+\prime}((1-p(n))(1-e^{-\lambda(n)})).\end{aligned}$$

Furthermore,

$$(1-p(n))(1-e^{-\lambda(n)} - \lambda(n)e^{-\lambda(n)})\phi_n^{+\prime}((1-p(n))(1-e^{-\lambda(n)})) \xrightarrow[n\to\infty]{} 0,$$

and from (B.2) we have

$$\lambda(n)^2\mathcal{L}_\mu(\lambda(n)) = 1 - \phi^+(\lambda(n)) + \lambda(n)\phi^{+\prime}(\lambda(n)), \quad n \to \infty,$$

where $\mathcal{L}_\mu$ is the Laplace transform of the measure μ defined in (B.1). From these last two displays we obtain

$$\lambda(n)^2\mathcal{L}_{\mu^{cn}}(\lambda(n)) \sim ((1-p(n))(1-e^{-\lambda(n)}))^2\mathcal{L}_\mu((1-p(n))(1-e^{-\lambda(n)})) + O(\lambda(n)^2), \quad n \to \infty.$$

Due to the estimate $\lambda(n) \sim 1 - e^{-\lambda(n)}$ as $n \to \infty$, the approximation of $\mathcal{L}_\mu$ given in (B.3) implies

$$c_\alpha\lambda(n)^2\mathcal{L}_{\mu^{cn}}(\lambda(n)) \sim \bar{\pi}^+\left(\frac{1}{\lambda(n)(1-p(n))}\right) + O(\lambda(n)^2), \quad n \to \infty. \tag{C.2}$$

Hence it is remains to prove

$$\lim_{n\to\infty} \frac{\bar{\pi}^{cn}(1/\lambda(n))}{(\lambda(n))^2\mathcal{L}_{\mu^{cn}}(\lambda(n))} = c_\alpha. \tag{C.3}$$

In this aim, define for every $y \geq 0$, the following measure

$$m^{cn}_{\lambda(n)}(0,y] := m^{cn}_{\lambda(n)}(y) = \frac{\mu^{cn}(y/\lambda(n))}{\mathcal{L}_{\mu^{cn}}(\lambda(n))}, \quad \forall\, y > 0.$$

Observe that

$$\int_{[0,\infty)} e^{-\theta s} d_y\left(\frac{\mu^{cn}(y/\lambda(n))}{\mathcal{L}_{\mu^{cn}}(\lambda(n))}\right) = \frac{1}{\mathcal{L}_{\mu^{cn}}(\lambda(n))}\int_{[0,\infty)} e^{-\theta\lambda(n)y}\mu^{cn}(dy) = \frac{\mathcal{L}_{\mu^{cn}}(\theta\lambda(n))}{\mathcal{L}_{\mu^{cn}}(\lambda(n))}, \quad \forall\theta > 0.$$

From the previous display and the estimate in (C.2) we get

$$\mathcal{L}_{m^{cn}_{\lambda(n)}}(\theta) \xrightarrow[n\to\infty]{} \theta^{-(2-\alpha)}, \quad \forall\theta > 0.$$

Writing now $\theta^{-(2-\alpha)}$ in terms of the gamma function we have

$$\theta^{-(2-\alpha)} = \frac{1}{\Gamma(2-\alpha)}\int_0^\infty s^{(2-\alpha)-1}e^{-\theta s}ds.$$

Since the convergence of the Laplace transform implies the weak convergence of measures (see Theorem 13.1.2 in Feller (1971)), we have

$$m^{cn}_{\lambda(n)}(y) \xrightarrow[n\to\infty]{} \frac{1}{\Gamma(2-\alpha)}\int_0^y s^{(2-\alpha)-1}ds = \frac{y^{2-\alpha}}{\Gamma(3-\alpha)}. \tag{C.4}$$

Moreover, by the definition of the measure $\mu^{(cn)}$ we can obtain the following inequality for any $y < 1$,

$$(1/\lambda(n))^2 y(1-y)\bar{\pi}^{cn}(y/\lambda(n)) \leq \mu^{cn}(1/\lambda(n)) - \mu^{cn}(y/\lambda(n)) \leq (1/\lambda(n))^2(1-y)\bar{\pi}^{cn}(1/\lambda(n)).$$

Due to (C.4) this implies for all $y < 1$.

$$c_\alpha\frac{1-y^{2-\alpha}}{1-y} \leq \liminf_{n\to\infty}\frac{\bar{\pi}^{cn}(1/\lambda(n))}{(\lambda(n))^2\mathcal{L}_{\mu^{cn}}(\lambda(n))} \leq \limsup_{n\to\infty}\frac{\bar{\pi}^{cn}(1/\lambda(n))}{(\lambda(n))^2\mathcal{L}_{\mu^{cn}}(\lambda(n))} \leq c_\alpha\frac{1-y^{2-\alpha}}{y(1-y)}.$$

To conclude we make $y \uparrow 1$.

Appendix D

Proof of Lemma 1.20

Before proving Lemma 1.20, we would like to present some basic aspect of functional convergence of stochastic process, further details can be found in Jacod and Shiryaev (1987). Is well know that the law of a Lévy process $\{X_t : t \geq 0\}$ on $\mathbb{R}^d$ is determined by that of random variable X_1, which is infinitely divisible random variable, and according to the Lévy-Khintchine formula has characteristic exponent

$$\Psi(\mathbf{u}) = \mathrm{i}\mathbf{u}\cdot\mathbf{b} - \frac{1}{2}\mathbf{u}\cdot c\mathbf{u}^T + \int \left(e^{\mathrm{i}\mathbf{u}\cdot\mathbf{x}} - 1 - \mathrm{i}\mathbf{u}\cdot\mathbf{h}(\mathbf{x})\right)\pi(d\mathbf{x})$$

where $\mathbf{b} \in \mathbb{R}^d$, c is a $d \times d$ symmetric nonnegative matrix, π is a positive measure on $\mathbb{R}^d$ with $\pi(\{\mathbf{0}\}) = 0$ and $\int (1 \wedge |\mathbf{x}|^2)\pi(d\mathbf{x}) < \infty$, $\mathbf{h}$ is a truncation function from $\mathbb{R}^d$ to $\mathbb{R}^d$, that is, bounded measurable satisfying

$$\mathbf{h}(\mathbf{x}) = o(|\mathbf{x}|), \quad |\mathbf{x}| \to 0.$$

Hence an infinitely divisible distribution, and therefore a Lévy process, is uniquely characterized by the triple $(\mathbf{b}, c, \pi)$. Another useful related quatity is a $d \times d$ symmetric nonnegative matrix, called the modified second characteristic, and defined as follows

$$\tilde{c}^{ij} = c^{ij} + \int h^i(\mathbf{x})h^j(\mathbf{x})\pi(d\mathbf{x}), \quad i, j = 1, 2, ..., d.$$

According to Theorem VII.2.9 of Jacod and Shiryaev (1987), if $\{\pi_n : n \in \mathbb{N}\}$ is a sequence of infinitely divisible distributions on $\mathbb{R}^d$. Then π_n converges weakly to π if and only if

$$\begin{aligned}
&\mathbf{b}_n \to \mathbf{b} \\
&\tilde{c}_n \to \tilde{c} \\
&\pi_n(\mathbf{g}) \to \pi(\mathbf{g}) \quad \text{for all } \mathbf{g} \in C_1(\mathbb{R}^d),
\end{aligned}$$

where $C_1(\mathbb{R}^d)$ is a convergence-determing class for the weak convergence induced by all continuous bounded non-negative functions $\mathbb{R}^d \to \mathbb{R}$, vanishing at the origin and with limit

at infinity. We will assume furthermore that $\mathbf{h}$ is a continuous function and that it is the same function for all the independent distributions considered here.

In a more general sense, a d-dimensional semimartingale W, has associated a characteristic triplet (B, C, ν) consisting in:

- $B = (B^i)_{i\leq d}$ a predictable process with components of finite variation over each interval $[0, t]$.
- $C = (C^{ij})_{i,j\leq d}$ a continuous process, namely

$$C^{ij} = \langle W^{i,c}, W^{j,c}\rangle,$$

  where W^c is the continuous martingale part of W.
- ν a predictable random measure on $\mathbb{R}_+ \times \mathbb{R}^d$.

A second modified characteristic $\tilde{C}$ is also defined,

$$\tilde{C}_t^{ij} = C_t^{ij} + (h^i h^j) * \nu_t - \sum_{s\leq t}\left(\int h^i(\mathbf{x})\nu(\{s\}\times d\mathbf{x})\right)\left(\int h^j(\mathbf{w})\nu(\{s\}\times d\mathbf{w})\right).$$

If W has no fixed times of discontinuity, in which case B is continuous, and $|h(x)|^2 * \nu_t < \infty$, it reduces to

$$\tilde{C}_t^{ij} = C_t^{ij} + (h^i h^j) * \nu_t.$$

According to Theorem VII.3.4 of Jacod and Shiryaev (1987), the necessary and sufficient conditions to assure the functional convergence of a sequence of semimartingales W^n towards W are given also in terms of their characteristics:

$$\begin{aligned}
&\sup_{s\leq t}|B_s^n - B_s| \to 0, \text{ for all } t \geq 0,\\
&\tilde{C}^n \to \tilde{C}, \text{ for all } t \in D,\\
&\mathbf{g} * \nu_t^n \to \mathbf{g} * \nu_t, \text{ for all } t \in D, \mathbf{g} \in C_1(\mathbb{R}^d),
\end{aligned} \tag{D.1}$$

where D is a dense subset of $\mathbb{R}_+$.

We now turn to prove the convergence claimed in Proposition 1.20. In order to apply the Theorem VII.3.4 of Jacod and Shiryaev (1987), first we will prove the convergence of the characteristics of the process

$$\tilde{\mathbf{S}}^n_{N_{r(n)t}} - (r(n)t/n, 0), \quad t \geq 0, \tag{D.2}$$

where

$$\tilde{\mathbf{S}}_k^n = \sum_{i=1}^{k}\left(\xi_i^{(cn)}/n, \xi_i^{(mn)}/r(n)p(n)\right), \quad k \in \mathbb{N},$$

and $\{N_t : t \geq 0\}$ is a Poisson process with parameter one, independent of the sequence

$$\xi^{(n)} = \left\{\left(\xi_k^{(cn)}, \xi_k^{(mn)}\right) : k \in \mathbb{Z}_+\right\}.$$

The following lemma establishes the previous statement. We will use this result as a device to study the characteristics of $\bar{\mathbf{S}}^n_{\lfloor r(n)t \rfloor}$, which are closely related to those of

$$\tilde{\mathbf{S}}^n_{N_{r(n)t}} - (r(n)t/n, 0).$$

.

Lemma D.1. *The process defined in (D.2) is a semimartingale with characteristics relatives to a continuous truncation function* $\mathbf{h}$ *given by*

$$\begin{aligned}
&\mathbf{b}_t^n = r(n)t\mathbb{E}\left[\mathbf{h}\left(b(n)\xi^{(n)}\right)\right] - (r(n)t/n, 0),\\
&c_t^{n,ij} = 0, \quad \tilde{c}_t^{n,ij} = r(n)t\mathbb{E}\left[h_i\left(b(n)\xi^{(n)}\right)h_j\left(b(n)\xi^{(n)}\right)\right], \quad i,j = 1,2,\\
&F_t^n(d\mathbf{x}) = r(n)t\pi(d\mathbf{x}),
\end{aligned} \tag{D.3}$$

where $b(n)\xi^{(n)} = (\xi^{(cn)}/n, \xi^{(mn)}/r(n)p(n))$, $\pi^{(n)}(d\mathbf{x}) = \mathbb{P}(\xi^{(cn)} \in dx_1, \xi^{(mn)} \in dx_2)$. *Moreover, in the regime determined by (1.19) and (1.21), we have the following weak convergence in the sense of Skorohod topology*

$$\left(\left\{\tilde{\mathbf{S}}^n_{N_{r(n)t}} - (r(n)t/n, 0) : t \geq 0\right\}, \mathbb{P}^{p(n)}_{a(n)}\right) \Longrightarrow \{(X_t, t) : t \geq 0\}, \tag{D.4}$$

where X_t *is a spectrally positive* α*-stable process with parameter* $\alpha \in (1,2)$. *In particular, we obtain the convergence of the characteristics in (D.3) towards those relatives to* $\{(X_t, t) : t \geq 0\}$ *and characteristic exponent* $c_\alpha|\lambda|^\alpha$, *that is*

$$\begin{aligned}
&\mathbf{b}_t = \left(t\left(\int_{(0,\infty)} \lambda(h(y) - y)c_\alpha y^{-(\alpha+1)}dy\right), t\right),\\
&c_t^{ij} = 0, \quad \tilde{c}_t^{ij} = \mathbb{E}\left[h_i\left(X_t\right)h_j\left(X_t\right)\right] \quad i,j = 1,2,\\
&F_t(d\mathbf{x}) = tc_\alpha x_1^{-(\alpha+1)}dx_1\delta_0(dx_2).
\end{aligned} \tag{D.5}$$

Proof of Lemma D.1. Note that for $\mathbf{u} = (\lambda, \theta) \in \mathbb{R}^2$,

$$\mathbb{E}\left(e^{\mathrm{i}\mathbf{u}\cdot\tilde{\mathbf{S}}^n_{N_{r(n)t}}}\right) = e^{r(n)t\left(\psi_n\left(\frac{\lambda}{n}, \frac{\theta}{r(n)p(n)}\right)-1\right)}, \quad t \geq 0. \tag{D.6}$$

Then the exponent in the righthand side of the previous equality can be written as follows

$$\begin{aligned}
\mathrm{i}t\int_{(0,\infty)}\int_{(0,\infty)} &\mathbf{u}\cdot\mathbf{h}(b(n)\mathbf{x})r(n)\pi^{(n)}(d\mathbf{x})\\
&+ t\int_{(0,\infty)}\int_{(0,\infty)}\left(e^{\mathrm{i}\mathbf{u}\cdot b(n)\mathbf{x}} - 1 - \mathrm{i}\mathbf{u}\cdot\mathbf{h}(b(n)\mathbf{x})\right)r(n)\pi^{(n)}(d\mathbf{x}),
\end{aligned} \tag{D.7}$$

where $b(n)\mathbf{x} = (x_1/n, x_2/r(n)p(n))$, $\pi^{(n)}(d\mathbf{x}) = \mathbb{P}(\xi^{(cn)} \in dx_1, \xi^{(mn)} \in dx_2)$. From here $\tilde{\mathbf{S}}_{N_{r(n)t}}$ is infinitely divisible, also we can deduce that the characteristics of the process $\{\tilde{\mathbf{S}}^n_{N_{r(n)t}} - (r(n)t/n, 0) : t \geq 0\}$ are given by (D.3). Thanks to Theorem II.3.11 of Jacod and Shiryaev (1987) this process is a Lévy process and also a semimartingale.

Besides, to get the convergence in (D.4) we shall prove the convergence of the characteristic functions. This fact is verified using Corollary 1.15, together with the fact that conditionally to $\xi^{(+)} = k$ the distribution of $\xi^{(cn)}$ is Binomial with parameter $(k, 1-p(n))$, as well as the assumption that ξ^+ has mean 1. Indeed, the expression (D.7) can be rewritten using the term

$$\mathrm{i}\left(\frac{\lambda}{n}(1-p(n)) + \frac{\theta}{r(n)p(n)}p(n)\right)r(n)t$$

Then by the continuity of $\mathbf{h}$ and the assumptions (1.11) and (1.20), the display (D.7) behaves as the following expression

$$\begin{aligned} t\int_{(0,\infty)} &\left(e^{-((1-p(n))(1-e^{\mathrm{i}\lambda/n})-p(n)(1-e^{\mathrm{i}\theta/r(n)p(n)}))y} - 1 - \mathrm{i}\lambda h\left(\frac{y}{n}\right)\right)r(n)\pi^+(dy) \\ &+ t\mathrm{i}\int_{(0,\infty)}\left(\lambda h\left(\frac{y}{n}\right) - \left(\frac{\lambda}{n}(1-p(n)) + \frac{\theta}{r(n)p(n)}p(n)\right)y\right)r(n)\pi^+(dy) \\ &+ \mathrm{i}\left(\frac{\lambda}{n}(1-p(n)) + \frac{\theta}{r(n)p(n)}p(n)\right)r(n)t. \end{aligned}$$

where h is the truncation function from $\mathbb{R}$ to $\mathbb{R}$ obtained as projection of $\mathbf{h}$ in the second coordinate. Making a change of variables $z = y/n$ we obtain

$$\begin{aligned} t\int_{(0,\infty)} &\left(e^{-((1-p(n))(1-e^{\mathrm{i}\lambda/n})-p(n)(1-e^{\mathrm{i}\theta/r(n)p(n)}))nz} - 1 - \mathrm{i}\lambda h(z)\right)r(n)\pi^+(ndz) \\ &+ t\mathrm{i}\int_{(0,\infty)}\left(\lambda h(z) - \left(\lambda(1-p(n)) + \frac{\theta}{r(n)p(n)}np(n)\right)z\right)r(n)\pi^+(ndz) \\ &+ \mathrm{i}\left(\frac{\lambda}{n}(1-p(n)) + \frac{\theta}{r(n)p(n)}p(n)\right)r(n)t. \end{aligned} \tag{D.8}$$

Finally we have the convergence

$$\mathbb{E}\left(e^{\mathrm{i}\mathbf{u}\cdot(\tilde{\mathbf{S}}^n_{N_{r(n)t}} - (r(n)t/n,0))}\right) \xrightarrow[n\to\infty]{} e^{t(\int_{(0,\infty)}(e^{\mathrm{i}\lambda y}-1-\mathrm{i}\lambda y)c_\alpha y^{-(\alpha+1)}dy)+\mathrm{i}t\theta}, \tag{D.9}$$

where c_α is a constant depending on α that appears in Lemma 1.13. Indeed, the result in Lemma 1.13 implies the following convergence

$$\begin{aligned} &\int_{(0,\infty)}\left(e^{-((1-p(n))(1-e^{\mathrm{i}\lambda/n})-p(n)(1-e^{\mathrm{i}\theta/r(n)p(n)}))ny} - 1 - \mathrm{i}\lambda h(y)\right)r(n)\pi^+(ndy) \\ &\qquad \xrightarrow[n\to\infty]{} t\int_{(0,\infty)}\left(e^{\mathrm{i}\lambda y} - 1 - \mathrm{i}\lambda h(y)\right)c_\alpha y^{-(\alpha+1)}dy, \end{aligned}$$

while the second adding in (D.8) converges towards

$$\mathrm{i}t\int_{(0,\infty)}\lambda(h(y)-y)c_\alpha y^{-(\alpha+1)}dy.$$

To finish, we observe that the assumption that $r(n)\in RV_\infty^\alpha$ with $\alpha\in(1,2)$ and $p(n)\sim cn^{-1}$ implies that

$$\frac{r(n)p(n)}{n}\xrightarrow[n\to\infty]{}0.$$

The final term in (D.9) is such that

$$\mathrm{i}\left(\frac{\lambda}{n}(1-p(n))+\frac{\theta}{r(n)p(n)}p(n)\right)r(n)t-\mathrm{i}\frac{\lambda}{n}r(n)t\xrightarrow[n\to\infty]{}\mathrm{i}\theta t,$$

for all $t\geq 0$. From (D.9) the characteristic of (X_t,t) are given by (D.5). As a consequence of Theorem VII.2.9 of Jacod and Shiryaev (1987) we have the convergence of the characteristics. Finally the characteristic exponent of X_t is $c_\alpha|\lambda|^\alpha$. ∎

We have now all the elements to prove Lemma 1.20.

Proof of Lemma 1.20. Thanks to Theorem 2.3.11 of Jacod and Shiryaev (1987), $\tilde{\mathbf{S}}^{(n)}_{\lfloor r(n)t\rfloor}$ is a semimartingale with characteristics relatives to $\mathbf{h}:\mathbb{R}^2\to\mathbb{R}^2$, given by

$$\begin{aligned}
&B_t^n=\lfloor r(n)t\rfloor\mathbb{E}\left[\mathbf{h}\left(b(n)\xi^{(n)}\right)\right]-(\lfloor r(n)t\rfloor/n,0),\\
&C_t^{n,ij}=0,\\
&\tilde{C}_t^{n,ij}=\lfloor r(n)t\rfloor\left(\mathbb{E}\left[h_i\left(b(n)\xi^{(n)}\right)h_j\left(b(n)\xi^{(n)}\right)\right]-\mathbb{E}\left[h_i\left(b(n)\xi^{(n)}\right)\right]\mathbb{E}\left[h_j(b(n)\xi^{(n)})\right]\right),\\
&\mathbf{g}*\nu_t^n=\lfloor r(n)t\rfloor\mathbb{E}\left(\mathbf{g}\left(b(n)\xi^{(n)}\right)\right),
\end{aligned}$$

where $i,j=1,2$ and $\mathbf{g}$ is a measurable and positive function. As usual $\lfloor\cdot\rfloor$ denotes the floor function. Now it only remains to verify the conditions (D.1) where the limit characteristics agree with these in (D.5). In this direction we recall that

$$\mathbf{b}_t^n=r(n)t\mathbb{E}\left[\mathbf{h}\left(b(n)\xi^{(n)}\right)\right]-(r(n)t/n,0),$$

and observe

$$|B_s^n-\mathbf{b}_s|\leq|\lfloor r(n)s\rfloor-r(n)s|\,\mathbb{E}\left[\mathbf{h}\left(b(n)\xi^{(n)}\right)\right]+|(r(n)s/n,0)-(\lfloor r(n)s\rfloor/n,0)|+|\mathbf{b}_s^n-\mathbf{b}_s|.$$

Then using the properties of the floor function we obtain

$$|B_s^n-\mathbf{b}_s|\leq\mathbb{E}\left[\mathbf{h}\left(b(n)\xi^{(n)}\right)\right]+|(s/n,0)|+|\mathbf{b}_s^n-\mathbf{b}_s|.$$

Thus, by the convergence of $\mathbf{b}_t^n$ established in the previous lemma, together with the fact that $r(n)\to\infty$, we get

$$\sup_{s\leq t}|B_s^n-\mathbf{b}_s|\leq\mathbb{E}\left[\mathbf{h}\left(b(n)\xi^{(n)}\right)\right]+|(t/n,0)|+|\mathbf{b}_t^n-\mathbf{b}_t|\xrightarrow[n\to\infty]{}0,$$

hence we have the first condition in (D.1). In order to determine the second one, let $b_t^{n,i}$ be the i-th coordinate of $\mathbf{b}_t^n$, $i = 1, 2$. Once more, applying the properties of the floor function we have

$$\left(1 - \frac{1}{r(n)t}\right)\tilde{c}_t^{n,ij} - \frac{1}{r(n)t}b_t^{n,i}b_t^{n,j} - \frac{r(n)t}{n}\mathbb{E}\left[h_2\left(b(n)\xi^{(n)}\right)\right] \leq \tilde{C}_t^{n,ij} \leq \tilde{c}_t^{n,ij} + \frac{1 - r(n)t}{(r(n)t)^2}b_t^{n,i}b_t^{n,j} + \frac{1 - r(n)t}{n}\mathbb{E}\left[h_2\left(b(n)\xi^{(n)}\right)\right].$$

Also

$$\left(1 - \frac{1}{r(n)t}\right)\tilde{c}_t^{n,22} - \frac{1}{r(n)t}(b_t^{n,2})^2 \leq \tilde{C}_t^{n,22} \leq \tilde{c}_t^{n,22} + \frac{1 - r(n)t}{(r(n)t)^2}(b_t^{n,2})^2,$$

and

$$\left(1 - \frac{1}{r(n)t}\right)\tilde{c}_t^{n,11} - \frac{1}{r(n)t}(b_t^{n,1})^2 \leq \tilde{C}_t^{n,11} \leq \tilde{c}_t^{n,11} + \frac{1 - r(n)t}{(r(n)t)^2}(b_t^{n,1})^2 + b_t^{n,1} + \frac{1}{n^2}[r(n) - 1].$$

As consequence of the convergence $\mathbf{b}_t^n \to \mathbf{b}_t$, $b_t^{n,i}$ converges for $i = 1, 2$. Then the above inequalities impliy $\tilde{C}_t^{n,ij} \to \tilde{c}_t^{ij}$, because $r(n) \in RV_\infty^\alpha$ and $\alpha \in (1, 2)$. It is easily prooved that also $\mathbf{g} * \nu_t^n \to \mathbf{g} * F_t$ for all g using that $F_t^n = r(n)t\pi(d\mathbf{x})$ converges to F_t, as we proved in the previous lemma, together with properties of the floor function. ■

Bibliography

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, 19(suppl 1):i7–i15.

Athreya, K. B. and Ney, P. E. (1972). *Branching processes*. Springer-Verlag, New York-Heidelberg. Die Grundlehren der mathematischen Wissenschaften, Band 196.

Barton, N. H., Etheridge, A. M., and Sturm, A. K. (2004). Coalescence in a random background. *Ann. Appl. Probab.*, 14(2):754–785.

Bertoin, J. (1996). *Lévy processes*, volume 121 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge.

Bertoin, J. (2006). *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Bertoin, J. (2010). A limit theorem for trees of alleles in branching processes with rare neutral mutations. *Stochastic Process. Appl.*, 120(5):678–697.

Bertoin, J. and Le Gall, J.-F. (2003). Stochastic flows associated to coalescent processes. *Probab. Theory Related Fields*, 126(2):261–288.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge.

Chauvin, B. (1986). Arbres et processus de Bellman-Harris. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(2):209–232.

Chung, K. L. and Walsh, J. B. (2005). *Markov processes, Brownian motion, and time symmetry*, volume 249 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, New York, second edition.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6):332–340.

Degnan, J. H., Rosenberg, N. A., and Stadler, T. (2012). The probability distribution of ranked gene trees on a species tree. *Math. Biosci.*, 235(1):45–55.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37.

Donnelly, P. and Kurtz, T. G. (1999). Genealogical processes for Fleming-Viot models with selection and recombination. *Ann. Appl. Probab.*, 9(4):1091–1148.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112; erratum, ibid. 3 (1972), 240; erratum, ibid. 3 (1972), 376.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual review of genetics*, 22(1):521–565.

Felsenstein, J. and Felenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland.

Foucart, C. (2011). Distinguished exchangeable coalescents and generalized Fleming-Viot processes with immigration. *Adv. in Appl. Probab.*, 43(2):348–374.

Foucart, C. (2012). Generalized Fleming-Viot processes with immigration via stochastic flows of partitions. *ALEA Lat. Am. J. Probab. Math. Stat.*, 9(2):451–472.

Greven, A., Limic, V., and Winter, A. (2005). Representation theorems for interacting Moran models, interacting Fisher-Wright diffusions and applications. *Electron. J. Probab.*, 10:no. 39, 1286–1356 (electronic).

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174.

Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene genealogies, variation and evolution.* Oxford University Press, Oxford. A primer in coalescent theory.

Heled, J. and Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8(1):289.

Herbots, H. M. (1997). The structured coalescent. In *Progress in population genetics and human evolution (Minneapolis, MN, 1994)*, volume 87 of *IMA Vol. Math. Appl.*, pages 231–255. Springer, New York.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Survey Evol. Biol.*, 7:1–44.

Jacod, J. and Shiryaev, A. N. (1987). *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Jiřina, M. (1958). Stochastic branching processes with continuous state space. *Czechoslovak Math. J.*, 8 (83):292–313.

Joffe, A. (1967). On the Galton-Watson branching process with mean less than one. *Ann. Math. Statist.*, 38:264–266.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132.

Kallenberg, O. (2002). *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition.

Kawazu, K. and Watanabe, S. (1971). Branching processes with immigration and related limit theorems. *Teor. Verojatnost. i Primenen.*, 16:34–51.

Kingman, J. F. C. (1978a). Random partitions in population genetics. *Proc. Roy. Soc. London Ser. A*, 361(1704):1–20.

Kingman, J. F. C. (1978b). The representation of partition structures. *J. London Math. Soc. (2)*, 18(2):374–380.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl.*, 13(3):235–248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Probab.*, (Special Vol. 19A):27–43. Essays in statistical science.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). Stem. *Bioinformatics*, 25:971–973. species tree estimation using maximum likelihood for gene trees under coalescence.

Lambert, A. (2007). Quasi-stationary distributions and the continuous-state branching process conditioned to be never extinct. *Electron. J. Probab.*, 12:no. 14, 420–446.

Lambert, A. (2008). Population dynamics and random genealogies. *Stoch. Models*, 24(suppl. 1):45–163.

Lamperti, J. and Ney, P. (1968). Conditioned branching processes and their limiting diffusions. *Teor. Verojatnost. i Primenen.*, 13:126–137.

Li, Z. (2011). *Measure-valued branching Markov processes*. Probability and its Applications (New York). Springer, Heidelberg.

Limic, V. and Sturm, A. (2006). The spatial Λ-coalescent. *Electron. J. Probab.*, 11:no. 15, 363–393 (electronic).

Linz, S., Radtke, A., and von Haeseler, A. (2007). A likelihood framework to measure horizontal gene transfer. *Molecular biology and evolution*, 24(6):1312–1319.

Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3):504–514.

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, 53(1):320–328.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic biology*, 46(3):523–536.

Maddison, W. P. and Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic biology*, 55(1):21–30.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution*, 25(7):1459–1471.

Möhle, M. (2000). Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Adv. in Appl. Probab.*, 32(4):983–993.

Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4):1547–1562.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.*, 29(1):59–75.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902.

Pomarede, J.-M. L. (1976). *A UNIFIED APPROACH VIA GRAPHS TO SKOROHOD'S TOPOLOGIES ON THE FUNCTION SPACE D*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Yale University.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical population biology*, 61(2):225–247.

Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4):1116–1125.

Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 5:Paper no. 12, 50 pp. (electronic).

Seneta, E. and Vere-Jones, D. (1966). On quasi-stationary distributions in discrete-time Markov chains with a denumerable infinity of states. *J. Appl. Probability*, 3:403–434.

Siri-Jégousse, A. (2009). *ÉTUDE DES GÉNÉALOGIES DANS DES MODÉLES DE GÉNÉTIQUE DES POPULATIONS*. Thesis (Ph.D.)–Université Paris Descartes.

Skorokhod, A. B. (1956). Convergence of random processes and limit theorems in probability theory. *Th. Probab. Appl.*, pages 261–290.

Sullivan, J. (2005). Maximum-likelihood methods for phylogeny estimation. *Methods in enzymology*, 395:757–779.

Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2014). The inference of gene trees with species trees. *Systematic biology*, page syu048.

Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460.

Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122(4):957–966.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical population biology*, 26(2):119–164.

Wakeley, J. (2008). Coalescent theory. roberts & co.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biology*, 7:256–276.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699.

Whitt, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res.*, 5(1):67–85.

Whitt, W. (2002). *Stochastic-process limits*. Springer Series in Operations Research. Springer-Verlag, New York. An introduction to stochastic-process limits and their application to queues.

Yaglom, A. M. (1947). Certain limit theorems of the theory of branching random processes. *Doklady Akad. Nauk SSSR (N.S.)*, 56:795–798.

Zähle, I., Cox, J. T., and Durrett, R. (2005). The stepping stone model. II. Genealogies and the infinite sites model. *Ann. Appl. Probab.*, 15(1B):671–699.