Accepted Manuscript

The split-and-drift random graph, a null model for speciation

François Bienvenu, Florence Débarre, Amaury Lambert

PII: S0304-4149(18)30300-4
DOI: https://doi.org/10.1016/j.spa.2018.06.009
Reference: SPA 3335

To appear in: *Stochastic Processes and their Applications*

Received date : 8 June 2017
Revised date : 17 March 2018
Accepted date : 29 June 2018

Please cite this article as: F. Bienvenu, F. Débarre, A. Lambert, The split-and-drift random graph, a null model for speciation, *Stochastic Processes and their Applications* (2018), https://doi.org/10.1016/j.spa.2018.06.009

The split-and-drift random graph, a null model for speciation

François Bienvenu[a,b,*], Florence Débarre[b], Amaury Lambert[a,b]

[a] *Center for Interdisciplinary Research in Biology (CIRB), CNRS UMR 7241, Collège de France, PSL Research University, Paris, France*
[b] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), CNRS UMR 8001, Sorbonne Université, Paris, France*

Abstract

We introduce a new random graph model motivated by biological questions relating to speciation. This random graph is defined as the stationary distribution of a Markov chain on the space of graphs on $\{1, \dots, n\}$. The dynamics of this Markov chain is governed by two types of events: vertex duplication, where at constant rate a pair of vertices is sampled uniformly and one of these vertices loses its incident edges and is rewired to the other vertex and its neighbors; and edge removal, where each edge disappears at constant rate. Besides the number of vertices n, the model has a single parameter r_n.

Using a coalescent approach, we obtain explicit formulas for the first moments of several graph invariants such as the number of edges or the number of complete subgraphs of order k. These are then used to identify five non-trivial regimes depending on the asymptotics of the parameter r_n. We derive an explicit expression for the degree distribution, and show that under appropriate rescaling it converges to classical distributions when the number of vertices goes to infinity. Finally, we give asymptotic bounds for the number of connected components, and show that in the sparse regime the number of edges is Poissonian.

Keywords: dynamical network, duplication-divergence, vertex duplication, genetic drift, species problem, coalescent

Contents

*Corresponding author: francois.bienvenu@normalesup.org

1. Introduction

In this paper, we introduce a random graph derived from a minimalistic model of speciation. This random graph bears superficial resemblance to classic models of protein interaction networks [1, 2, 3, 4] in that the events shaping the graph are the duplication of vertices and the loss of edges. However, our model is obtained as the steady state of a Markov process (rather than by repeatedly adding vertices), and has the crucial feature that the duplication of vertices is independent from the loss of edges. These differences result in a very different behavior of the model.

Before describing the model formally in Section 1.2, let us briefly explain the motivation behind its introduction.

1.1. Biological context

Although it is often presented as central to biology, there is no consensus about how the concept of species should be defined. A widely held view is that it should be based on the capacity of individuals to interbreed. This is the so-called "biological species concept", wherein a species is defined as a group of potentially interbreeding populations that cannot interbreed with populations outside the group.

This view, whose origins can be traced back to the beginning of the 20th century [5], was most famously promoted by Ernst Mayr [6] and has been most influential in biology [7]. However, it remains quite imprecise: indeed, groups of populations such that (1) all pairs of populations can interbreed and (2) no population can interbreed with a population outside the group are probably not common in nature – and, at any rate, do not correspond to what is considered a species in practice. Therefore, some leniency is required when applying conditions (1) and (2). But once we allow for this, there are several ways to formalize the biological species concept, as illustrated in Figure 1. Thus, it seems arbitrary to favor one over the others in the absence of a mechanism to explain why some kind of groups should be more relevant (e.g., arise more frequently) than others.

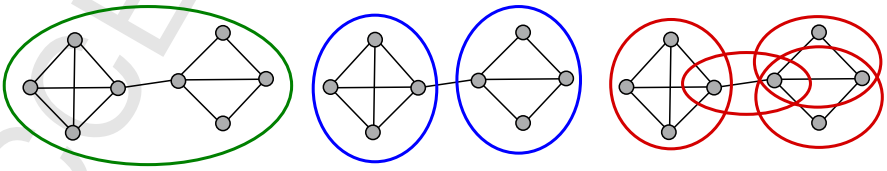

Figure 1: The vertices of the graph represent populations and its edges denote interbreeding potential (that is, individuals from two linked populations could interbreed, if given the chance). Even with such perfect information, it is not obvious how to delineate "groups of potentially interbreeding populations that cannot interbreed with populations outside the group": should these correspond to connected components (on the left, in green), maximal complete subgraphs (on the right, in red), or be based on some other clustering method (middle, in blue)?

Our aim is to build a minimal model of speciation that would make predictions about the structure and dynamics of the interbreeding network and allow one to recover species as an emergent property. To do so, we model speciation at the level of populations. Thus, we consider a set of n populations and we track the interbreeding ability for every pair of populations. All this information is encoded in a graph whose vertices correspond to populations and whose edges indicate potential interbreeding, i.e., two vertices are linked if and only if the corresponding populations can interbreed.

Speciation will result from the interplay between two mechanisms. First, populations can sometimes "split" into two initially identical populations which then behave as independent entities; this could happen as a result of the fragmentation of the habitat or of the colonization of a new patch. Second, because they behave as independent units, two initially identical populations will diverge (e.g., as a result of genetic drift) until they can no longer interbreed.

1.2. Formal description of the model

Start from any graph with vertex set $V = \{1, \dots, n\}$, and let it evolve according to the following rules

1. **Vertex duplication:** each vertex "duplicates" at rate 1; when a vertex duplicates, it chooses another vertex uniformly at random among the other vertices and replaces it with a copy of itself. The replacement of j by a copy of i means that j loses its incident edges and is then linked to i and to all of its neighbors, as depicted in Figure 2.

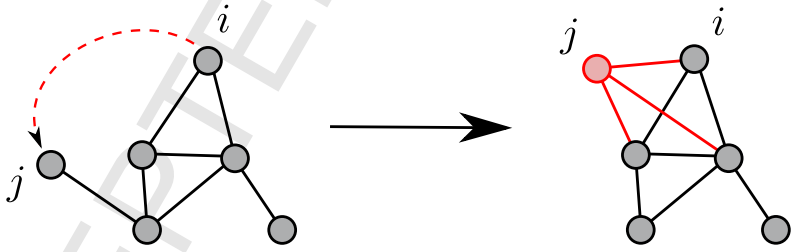

Figure 2: An illustration of vertex duplication. Here, i duplicates and replaces j. After the duplication, j is linked to i and to each of its neighbors.

2. **Edge removal:** each edge disappears at constant rate ρ.

This procedure defines a continuous-time Markov chain $(G_n(t))_{t\geqslant 0}$ on the finite state space of all graphs whose vertices are the integers $1, \dots, n$. It is easy to see that this Markov chain is irreducible. Indeed, to go from any graph $G^{(1)}$ to any graph $G^{(2)}$, one can consider the following sequence of events: first, a vertex is duplicated repeatedly in order to obtain the complete graph of order n (e.g., $\forall k \in \{2, \dots, n\}$, vertex k is replaced by a copy of vertex 1); then, all the edges that are not in $G^{(2)}$ are removed.

Because the Markov chain $(G_n(t))_{t\geqslant 0}$ is irreducible, it has a unique stationary probability distribution $\mu_{n,\rho}$. This probability distribution on the set of graphs of order n defines a random graph that is the object of study of this paper.

1.3. Notation

To study the asymptotic behavior of our model as $n \to +\infty$, we can let ρ, the ratio of the edge removal rate to the vertex duplication rate, be a function of n. As will become evident, it is more convenient to parametrize the model by

$$r_n \coloneqq \tfrac{n-1}{2}\,\rho_n\,.$$

Thus, we write G_{n,r_n} to refer to a random graph whose law is $\mu_{n,\frac{2r_n}{n-1}}$.

Although some of our results hold for any (n, r), in many cases we will be interested in asymptotic properties that are going to depend on the asymptotics of r_n. To quantify these, we will use the Bachmann–Landau notation, which for positive sequences r_n and $f(n)$ can be summarized as:

- $r_n \sim f(n)$ when $r_n/f(n) \to 1$.
- $r_n = o(f(n))$ when $r_n/f(n) \to 0$.
- $r_n = \Theta(f(n))$ when there exists positive constants α and β such that, asymptotically, $\alpha f(n) \leqslant r_n \leqslant \beta f(n)$.
- $r_n = \omega(f(n))$ when $r_n/f(n) \to +\infty$.

These notations also have stochastic counterparts, whose meaning will be recalled when we use them.

Finally, we occasionally use the expression *asymptotically almost surely* (abbreviated as a.a.s.) to that a property holds with probability that goes to 1 as n tends to infinity:

$$\mathcal{Q}_n \text{ a.a.s.} \iff \mathbb{P}(\mathcal{Q}_n) \xrightarrow[n\to+\infty]{} 1\,.$$

1.4. Statement of results

Table 1 lists the first moments of several graph invariants obtained in Section 3.1. These are then used to identify different regimes, depending on the asymptotic behavior of the parameter r_n, as stated in Theorem 3.10.

Theorem 3.10. *Let D_n be the degree of a fixed vertex of G_{n,r_n}. In the limit as $n \to +\infty$, depending on the asymptotics of r_n we have the following behaviors for G_{n,r_n}*

(i) Complete graph: *when $r_n = o(1/n)$, $\mathbb{P}(G_{n,r_n}$ is complete$)$ goes to 1, while when $r_n = \omega(1/n)$ it goes to 0; when $r_n = \Theta(1/n)$, this probability is bounded away from 0 and from 1.*

(ii) Dense regime: *when* $r_n = o(1)$, $\mathbb{P}(D_n = n-1) \to 1$.

(iii) Sparse regime: *when* $r_n = \omega(n)$, $\mathbb{P}(D_n = 0) \to 1$.

(iv) Empty graph: *when* $r_n = o(n^2)$, $\mathbb{P}(G_{n,r_n}$ *is empty) goes to* 0 *while when* $r_n = \omega(n^2)$ *it goes to* 1*; when* $r_n = \Theta(n^2)$*, this probability is bounded away from* 0 *and from* 1.

Variable	Expectation	Variance	Covariance
$\mathbb{1}_{\{i\leftrightarrow j\}}$	$\frac{1}{1+r}$	$\frac{r}{(1+r)^2}$	$\frac{r}{(1+r)^2(3+2r)}$ if vertex in common, $\frac{2\,r}{(1+r)^2(3+r)(3+2r)}$ otherwise.
$D_n^{(i)}$	$\frac{n-1}{1+r}$	$\frac{r(n-1)(1+2r+n)}{(1+r)^2\,(3+2r)}$	$\frac{r}{(1+r)^2}\left(1+\frac{3(n-2)}{3+2r}+\frac{2(n-2)(n-3)}{(3+r)(3+2r)}\right)$
$\lvert E_n\rvert$	$\frac{n(n-1)}{2(1+r)}$	$\frac{rn(n-1)(n^2+2r^2+2nr+n+5r+3)}{2\,(1+r)^2(3+r)\,(3+2r)}$	—
$X_{n,k}$	$\binom{n}{k}\left(\frac{1}{1+r}\right)^{k-1}$	unknown	—

Table 1: First and second moments of several graph invariants of $G_{n,r}$: $\mathbb{1}_{\{i\leftrightarrow j\}}$ is the variable indicating that $\{ij\}$ is an edge, $D_n^{(i)}$ the degree of vertex i, $|E_n|$ the number of edges and $X_{n,k}$ the number of complete subgraphs of order k. The covariance of the indicator variables of two edges depends on whether these edges share a common end, hence the two expressions. All expressions hold for every value of n and r.

In Section 4, we derive an explicit expression for the degree distribution, which holds for every value of n and r_n. We then show that, under appropriate rescaling, this degree converges to classical distributions.

Theorem 4.1. *Let* D_n *be the degree of a fixed vertex of* G_{n,r_n}*. Then, for each* $k \in \{0,\dots,n-1\}$*,*

$$\mathbb{P}(D_n = k) = \frac{2\,r_n\,(2\,r_n+1)}{(n+2\,r_n)(n-1+2\,r_n)}\,(k+1)\prod_{i=1}^{k}\frac{n-i}{n-i+2\,r_n-1}\,,$$

where the empty product is 1.

Theorem 4.2.

(i) *If* $r_n \to r > 0$*, then* $\frac{D_n}{n}$ *converges in distribution to a* $\mathrm{Beta}(2, 2\,r)$ *random variable.*

(ii) *If* r_n *is both* $\omega(1)$ *and* $o(n)$*, then* $\frac{D_n}{n/r_n}$ *converges in distribution to a size-biased exponential variable with parameter* 2*.*

(iii) *If* $2\,r_n/n \to \rho > 0$*, then* $D_n + 1$ *converges in distribution to a size-biased geometric variable with parameter* $\rho/(1+\rho)$*.*

Asymptotic bounds for the number of connected components are obtained in Section 5, where the following theorem is proved.

Theorem 5.1. *Let $\#\mathrm{CC}_n$ be the number of connected components of G_{n,r_n}. If r_n is both $\omega(1)$ and $o(n)$, then*

$$\frac{r_n}{2} + o_p(r_n) \leqslant \#\mathrm{CC}_n \leqslant 2\, r_n \log n + o_p(r_n \log n)$$

where, for a positive sequence (u_n), $o_p(u_n)$ denotes a given sequence of random variables (X_n) such that $X_n/u_n \to 0$ in probability.

Because the method used to obtain the upper bound in Theorem 5.1 is rather crude, we formulate the following conjecture, which is well supported by simulations.

Conjecture 5.4.

$$\exists \alpha, \beta > 0 \ \textit{s.t.}\ \mathbb{P}(\alpha r_n \leqslant \#\mathrm{CC}_n \leqslant \beta r_n) \xrightarrow[n\to\infty]{} 1.$$

Finally, in Section 6 we use the Stein–Chen method to show that the number of edges is Poissonian in the sparse regime, as shown by Theorem 6.1.

Theorem 6.1. *Let $|E_n|$ be the number of edges of G_{n,r_n}. If $r_n = \omega(n)$ then*

$$d_{\mathrm{TV}}\Big(|E_n|, \mathrm{Poisson}(\lambda_n)\Big) \xrightarrow[n\to+\infty]{} 0\,,$$

where d_{TV} stands for the total variation distance and $\lambda_n = \mathbb{E}(|E_n|) \sim \frac{n^2}{2r_n}$. If in addition $r_n = o(n^2)$, then $\lambda_n \to +\infty$ and as a result

$$\frac{|E_n| - \lambda_n}{\sqrt{\lambda_n}} \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}(0,1)\,,$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution.

These results are summarized in Figure 3.

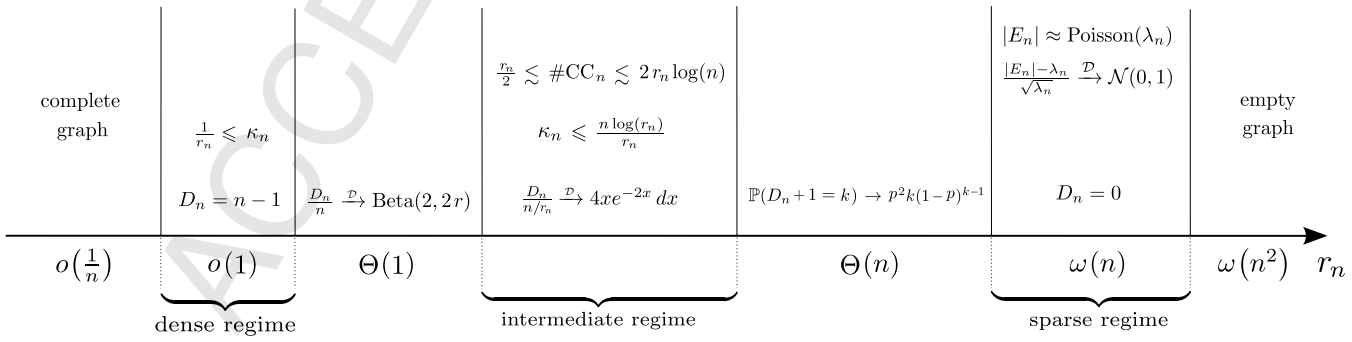

Figure 3: A graphical summary of the main results established in the paper; D_n is the degree of a fixed vertex, $|E_n|$ the number of edges, $\#\mathrm{CC}_n$ the number of connected components, and κ_n the clique number. All equalities and inequalities are to be understood "asymptotically almost surely" (i.e. hold with probability that goes to 1 as n tends to infinity).

2. Coalescent constructions of G_{n,r_n}

In this section, we detail coalescent constructions of G_{n,r_n} that will be used throughout the rest of the paper. Let us start by recalling some results about the Moran model.

2.1. The standard Moran process

The Moran model [8] is a classic model of population genetics. It consists in a set of n particles governed by the following dynamics: after an exponential waiting time with parameter $\binom{n}{2}$, a pair of particles is sampled uniformly at random. One of these particles is then removed (death) and replaced by a copy of the other (birth), and we iterate the procedure.

In this document, we will use the Poissonian representation of the Moran process detailed in the next definition.

Definition 2.1. The *driving measure of a standard Moran process on* V is a collection $\mathcal{M} = (M_{(ij)})_{(ij)\in V^2}$ of i.i.d. Poisson point processes with rate $1/2$ on $\mathbb{R}$.

We think of the elements of V as sites, each occupied by a single particle. In forward time, each atom $t \in M_{(ij)}$ indicates the replacement, at time t, of the particle in i by a copy of the particle in j.

For any given time $\alpha \in \mathbb{R}$, $\mathcal{M}$ defines a *genealogy of* V *on* $]-\infty, \alpha]$. Taking $\alpha = 0$ and working in backward time, i.e. writing $t \geqslant 0$ to refer to the absolute time $-t$, this genealogy is described by a collection of *ancestor functions* a_t, $t \in [0, +\infty[$, $a_t \colon V \to V$, defined as follows: $(a_t)_{t\geqslant 0}$ is the piecewise constant process such that

(i) a_0 is the identity on V.

(ii) If $t \in M_{(ij)}$ then

- For all k such that $a_{t-}(k) = i$, $a_t(k) = j$.
- For all k such that $a_{t-}(k) \neq i$, $a_t(k) = a_{t-}(k)$.

(iii) If for all $(ij) \in V^2$, $M_{(ij)} \cap [s, t] = \emptyset$, then $a_t = a_s$.

We refer to $a_t(i)$ as the ancestor of i at time t before the present – or, more simply, as the *ancestor of* i *at time* t.

The standard Moran process is closely related to the Kingman coalescent [9]. Indeed, let $\mathcal{R}_t$ denote the equivalence relation on V defined by

$$i \,\mathcal{R}_t\, j \iff a_t(i) = a_t(j)\,,$$

and let $K_t = V/\mathcal{R}_t$ be the partition of V induced by $\mathcal{R}_t$. Then, $(K_t)_{t\geqslant 0}$ is a Kingman coalescent on V. In particular, we will frequently use the next lemma.

Lemma 2.2. *Let $(a_t)_{t\geqslant 0}$ be the ancestor functions of a standard Moran process on V. For any $i \neq j$, let*

$$T_{\{ij\}} = \inf\{t \geqslant 0 : a_t(i) = a_t(j)\}$$

be the coalescence time of i and j and, for any $S \subset V$, let

$$T_S = \inf\{T_{\{ij\}} : i, j \in S, i \neq j\}.$$

Then, for all $t \geqslant 0$, conditional on $\{T_S > t\}$, $(T_S - t)$ is an exponential variable with parameter $\binom{|S|}{2}$.

For a more general introduction to Kingman's coalescent and Moran's model, one can refer to e.g. [10] or [11].

2.2. Backward construction

We now turn to the description of the coalescent framework on which our study relies. The crucial observation is that, for t large enough, every edge of $G_n(t)$ can ultimately be traced back to an initial edge that was inserted between a duplicating vertex and its copy. To find out whether two vertices i and j are linked in $G_n(t)$, we can trace back the ancestry of the potential link between them and see whether the corresponding initial edge and its subsequent copies survived up to time t. The first part of this procedure depends only on the vertex duplication process and, conditional on the sequence of ancestors of $\{ij\}$, the second one depends only on the edge removal process, making the whole procedure tractable. The next proposition formalizes these ideas.

Proposition 2.3. *Let $V = \{1, \dots, n\}$ and let $V^{(2)}$ be the set of unordered pairs of elements of V. Let $\mathcal{M}$ be the driving measure of a standard Moran process on V, and $(a_t)_{t\geqslant 0}$ the associated ancestor functions (that is, for each i in V, $a_t(i)$ is the ancestor of i at time t). Let $\mathcal{P} = (P_{\{ij\}})_{\{ij\}\in V^{(2)}}$ be a collection of i.i.d. Poisson point processes with rate r_n on $[0, +\infty[$ such that $\mathcal{M}$ and $\mathcal{P}$ are independent. For every pair $\{ij\} \in V^{(2)}$, define*

$$P^{\star}_{\{ij\}} = \left\{t \geqslant 0 : t \in P_{\{a_t(i)a_t(j)\}}\right\},$$

with the convention that, $\forall k \in V$, $P_{\{k\}} = \emptyset$. Finally, let $G = (V, E)$ be the graph defined by

$$E = \left\{\{ij\} \in V^{(2)} : P^{\star}_{\{ij\}} = \emptyset\right\}.$$

Then, $G \sim G_{n,r_n}$.

Throughout the rest of this document, we will write G_{n,r_n} for the graph obtained by the procedure of Proposition 2.3.

Proof of Proposition 2.3. First, consider the two-sided extension of $(G_n(t))_{t\geqslant 0}$, i.e. the corresponding stationary process on $\mathbb{R}$ (see, e.g., Section 7.1 of [12]), which by a slight abuse of notation we note $(G_n(t))_{t\in\mathbb{R}}$. Next, let $(\bar{G}_n(t))_{t\in\mathbb{R}}$ be the time-rescaled process defined by

$$\bar{G}_n(t) = G_n(t(n-1)/2)\,.$$

This time-rescaled process has the same stationary distribution as $(G_n(t))_{t\in\mathbb{R}}$ and so, in particular, $\bar{G}_n(0) \sim G_{n,r_n}$.

In the time-rescaled process, each vertex duplicates at rate $(n-1)/2$ and each edge disappears at rate $r_n = (n-1)\rho_n/2$. All these events being independent, we see that the vertex duplications correspond to the atoms of a standard Moran process on $V = \{1,\dots,n\}$, and the edge removals to the atoms of $\binom{n}{2}$ i.i.d. Poisson point processes with rate r_n on $\mathbb{R}$, that are also independent of the Moran process. Thus, there exists $(\bar{\mathcal{M}}, \bar{\mathcal{P}})$ with the same law as $(\mathcal{M}, \mathcal{P})$ from the proposition and such that, for $t \geqslant 0$,

- If $t \in \bar{M}_{(ij)}$, then j duplicates and replaces i in $\bar{G}_n(-t)$.
- If $t \in \bar{P}_{\{ij\}}$, then if there is an edge between i and j in $\bar{G}_n(-t)$, it is removed.

Since $(\bar{\mathcal{M}}, \bar{\mathcal{P}})$ has the same law as $(\mathcal{M}, \mathcal{P})$, if we show that

$$\{ij\} \in \bar{G}_n(0) \iff \bar{P}^\star_{\{ij\}} = \emptyset\,,$$

where $\bar{P}^\star_{\{ij\}} = \left\{t \geqslant 0 : t \in \bar{P}_{\{\bar{a}_t(i)\bar{a}_t(j)\}}\right\}$ is the same deterministic function of $(\bar{\mathcal{M}}, \bar{\mathcal{P}})$ as $P^\star_{\{ij\}}$ of $(\mathcal{M}, \mathcal{P})$, then we will have proved that $\bar{G}_n(0)$ has the same law as the graph G from the proposition.

Now to see why the edges of $\bar{G}_n(0)$ are exactly the pairs $\{ij\}$ such that $\bar{P}_{\{ij\}}$ is empty, note that, in the absence of edge-removal events, $\bar{G}_n(0)$ is the complete graph and the ancestor the edge $\{ij\}$ at time t is $\{a_t(i)\, a_t(j)\}$. Conversely, deleting the edge $\{k\ell\}$ from $\bar{G}_n(-t)$ will remove all of its subsequent copies from $\bar{G}_n(0)$, i.e. all edges $\{ij\}$ such that $\{a_t(i)\, a_t(j)\} = \{k\ell\}$. Thus, the edges of $\bar{G}_n(0)$ are exactly the edges that have no edge-removal events on their ancestral lineage – i.e, such that $\bar{P}^\star_{\{ij\}} = \emptyset$. □

Proposition 2.3 shows that G_{n,r_n} can be obtained as a deterministic function of the genealogy $(a_t)_{t\geqslant 0}$ of a Moran process and of independent Poisson point processes. Our next result shows that, in this construction, $(a_t)_{t\geqslant 0}$ can be replaced by a more coarse-grained process – namely, a Kingman coalescent (note that the Kingman coalescent contains less information because it only keeps track of blocks, not of which ancestor corresponds to which block at a given time t). This will be useful to give a forward construction of G_{n,r_n} in Section 2.3. The proof of this result is straightforward and can be found in Section A of the Appendix.

Proposition 2.4. *Let $(K_t)_{t\geqslant 0}$ be a Kingman coalescent on $V = \{1, \dots, n\}$, and let $\pi_t(i)$ denote the block containing i in the corresponding partition at time t. Let the associated genealogy of pairs be the set*

$$\mathcal{G} = \left\{ \Big(t,\, \{\pi_t(i)\, \pi_t(j)\}\Big) : \{ij\} \in V^{(2)},\, t \in \Big[0, T_{\{ij\}}\Big[\right\},$$

where $T_{\{ij\}} = \inf\{t \geqslant 0 : \pi_t(i) = \pi_t(j)\}$. Denote by

$$L_{\{ij\}} = \left\{ \Big(t,\, \{\pi_t(i)\, \pi_t(j)\}\Big) : t \in \Big[0, T_{\{ij\}}\Big[\right\}$$

the lineage of $\{ij\}$ in this genealogy. Finally, let $P^\bullet$ be a Poisson point process with constant intensity r_n on $\mathcal{G}$ and let $G = (V, E)$, where

$$E = \Big\{ \{ij\} \in V^{(2)} : P^\bullet \cap L_{\{ij\}} = \emptyset \Big\}.$$

Then, $G \sim G_{n,r_n}$.

We finish this section with a technical lemma that will be useful in the calculations of Section 3.1. Again, the proof of this result has no interest in itself and can be found in Section A of the Appendix.

Lemma 2.5. *Let S be a subset of $V^{(2)}$. Conditional on the measure $\mathcal{M}$, for any interval $I \subset [0, +\infty[$ such that*

(i) *For all $\{ij\} \in S$, $\forall t \in I$, $a_t(i) \neq a_t(j)$.*

(ii) *For all $\{k\ell\} \neq \{ij\}$ in S, $\forall t \in I$, $\{a_t(i)\, a_t(j)\} \neq \{a_t(k)\, a_t(\ell)\}$,*

$P^\star_{\{ij\}} \cap I$, $\{ij\} \in S$, are independent Poisson point processes with rate r_n on I.

Moreover, for any disjoint intervals I and J, $(P^\star_{\{ij\}} \cap I)_{\{ij\}\in S}$ is independent of $(P^\star_{\{ij\}} \cap J)_{\{ij\}\in S}$.

Before closing this section, let us sum up our results in words: if we think of $\{a_t(i)\, a_t(j)\}$ as being the ancestor of $\{ij\}$ at time t, then the genealogy of vertices induces a genealogy of pairs of vertices, as illustrated by Figure 4. Edge-removal events occur at constant rate r_n along the branches of this genealogy and the events affecting disjoint sections of branches are independent, so that we can think of $P^\star_{\{ij\}}$, $\{ij\} \in V^{(2)}$, as a single Poisson point process $P^\star$ on the lineages of pairs of vertices. A pair of vertices is an edge of G_{n,r_n} if and only if there is no atom of $P^\star$ on its lineage.

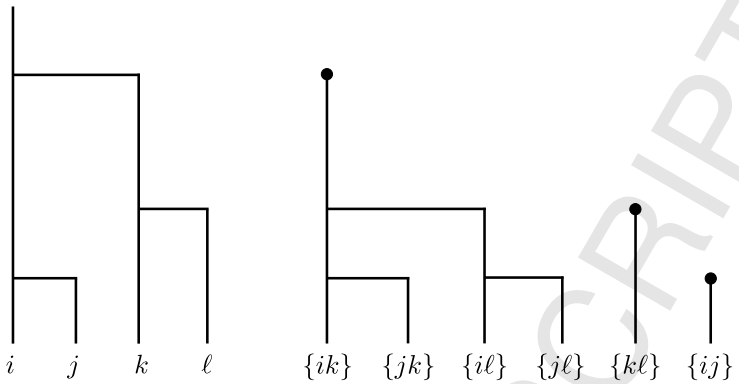

Figure 4: On the left, a genealogy on $\{i, j, k, \ell\}$ and on the right the corresponding genealogy of the pairs. Edge removal events occur at constant rate along the lineages of pairs of vertices, and a pair of vertices is an edge of G_{n,r_n} if and only if there is no atom on its lineage.

2.3. Forward construction

We now give a forward version of the coalescent construction presented in the previous section. Here, unlike in the previous section, the graph $G_{n,r}$ is built by adding vertices one at a time. This construction will be useful in proofs and provides a computationally efficient way to sample $G_{n,r}$.

Consider the Markov process $(G_r^\dagger(t))_{t\geqslant 0}$ defined by

(i) $G_r^\dagger(0) = (\{1,2\}, \{\{1,2\}\})$ is the complete graph of order 2.

(ii) Conditional on $V_t = \{1, \dots, n\}$, where V_t is the set of vertices of $G_r^\dagger(t)$: at rate $\binom{n}{2}$, a vertex is sampled uniformly in V_t and duplicated without replacement – that is, we copy the vertex and all incident edges, and label the new vertex $n+1$, resulting in a graph with vertex set $\{1, \dots, n+1\}$.

(iii) During the whole process, each edge disappears at constant rate r.

Next, for every integer $n \geqslant 2$, let $G_r^\star(n) = G_r^\dagger(t_n-)$, where

$$t_n = \sup\Big\{t \geqslant 0 : G_r^\dagger(t) \text{ has } n \text{ vertices}\Big\}.$$

Finally, let $\Phi_n(G_r^\star(n))$ denote the graph obtained by shuffling the labels of the vertices of $G_r^\star(n)$ uniformly at random, i.e. let Φ_n be picked uniformly at random among all the permutations of $\{1, \dots, n\}$ and, by a slight abuse of notation, let

$$\Phi_n(G_r^\star(n)) = \Big(\{1, \dots, n\}, \Big\{\{\Phi_n(i)\, \Phi_n(j)\} : \{ij\} \in G_r^\star(n)\Big\}\Big)$$

Proposition 2.6. *For any $r > 0$, for any integer $n \geqslant 2$,*

$$\Phi_n(G_r^\star(n)) \sim G_{n,r}\,.$$

Going from a backward construction such as Proposition 2.4 to a forward construction such as Proposition 2.6 is common in coalescent theory. The proofs, though straightforward, are somewhat tedious. They can be found in Section A of the Appendix, and we save the rest of this section to comment on the forward construction.

Proposition 2.6 shows that, for any given sequence (r_n), for any $n \geqslant 2$, $\Phi_n(G^\star_{r_n}(n)) \sim G_{n,r_n}$. Note however that this is *not* a compatible construction of a sequence $(G_{n,r_n})_{n\geqslant 2}$. In particular, all elements of a sequence $(\Phi_n(G^\star_r(n)))_{n\geqslant 2}$ are associated to the same value of r, while each term of a sequence $(G_{n,r_n})_{n\geqslant 2}$ corresponds to a different value of r_n.

Finally, it is necessary to relabel the vertices of $G^\star_r(n)$ in Proposition 2.6, as failing to do so would condition on $\{k, k-1\}$ being the $(n-k+1)$-th pair of vertices to coalesce in the genealogy of $G_{n,r}$ (in particular, the edges of $G^\star_r(n)$ are not exchangeable: "old" edges such as $\{1,2\}$ are least likely to be present than more recent ones such as $\{n-1, n\}$). However, since $G^\star_r(n)$ and $\Phi_n(G^\star_r(n))$ are isomorphic, when studying properties that are invariant under graph isomorphism (such as the number of connected components in Section 5 or the positive association of the edges in Section 6), we can work directly on $G^\star_r(n)$.

3. First and second moment methods

In this section, we apply Proposition 2.3 and Lemma 2.5 to obtain the expressions presented in Table 1. These are then used to identify different regimes for G_{n,r_n}, depending on the asymptotic behavior of the parameter r_n.

In order to be able to use Lemma 2.5, we will always reason conditionally on the genealogy of the vertices (i.e. on the vertex duplication process $\mathcal{M}$) and then integrate against its law.

3.1. First moments of graph invariants

3.1.1. Degree and number of edges

Proposition 3.1. *For any fixed vertices i and j, $i \neq j$, the probability that i and j are linked in G_{n,r_n} is*

$$\mathbb{P}(i \leftrightarrow j) = \frac{1}{1+r_n}\,.$$

Corollary 3.2. *Let D_n be the degree of a fixed vertex of G_{n,r_n}, and $|E_n|$ be the number of edges of G_{n,r_n}. Then,*

$$\mathbb{E}(D_n) = \frac{n-1}{1+r_n} \quad \textit{and} \quad \mathbb{E}(|E_n|) = \binom{n}{2}\frac{1}{1+r_n}\,.$$

Proof. By Proposition 2.3,

$$\{i \leftrightarrow j\} \iff P^\star_{\{ij\}} \cap [0, T_{\{ij\}}[= \emptyset\,.$$

Reasoning conditionally on $T_{\{ij\}}$ and applying Lemma 2.5 to $S = \{\{ij\}\}$ and $I = [0, T_{\{ij\}}[$, we see that $P^\star_{\{ij\}}$ is a Poisson point process with rate r_n on I. Since $T_{\{ij\}} \sim \mathrm{Exp}(1)$,

$$\mathbb{P}(i \leftrightarrow j) = \mathbb{P}\left(e_1 > T_{\{ij\}}\right),$$

where $e_1 = \inf P^\star_{\{ij\}}$ is an exponential variable with rate r_n that is independent of $T_{\{ij\}}$.

The corollary follows directly from the fact that the degree of a vertex v can be written as

$$D_n^{(v)} = \sum_{i \neq v} \mathbb{1}_{\{i \leftrightarrow v\}}$$

and that the number of edges of G_{n,r_n} is

$$|E_n| = \sum_{\{ij\} \in V^{(2)}} \mathbb{1}_{\{i \leftrightarrow j\}}. \qquad \square$$

Proposition 3.3. *Let i, j and k be three distinct vertices of G_{n,r_n}. We have*

$$\mathrm{Cov}\left(\mathbb{1}_{\{i \leftrightarrow j\}}, \mathbb{1}_{\{i \leftrightarrow k\}}\right) = \frac{r_n}{(3 + 2r_n)(1 + r_n)^2}$$

Corollary 3.4. *Let D_n be the degree of a fixed vertex of G_{n,r_n}. We have*

$$\mathrm{Var}(D_n) = \frac{r_n(n-1)(1 + 2\,r_n + n)}{(1 + r_n)^2\,(3 + 2\,r_n)}$$

Proof. For all $t \geqslant 0$, let $S_t = \{a_t(i), a_t(j), a_t(k)\}$. Let $\tau_1 = \inf\{t \geqslant 0 : |S_t| = 2\}$ and $\tau_2 = \inf\{t \geqslant \tau_1 : |S_t| = 1\}$. Recall from Lemma 2.2 that τ_1 and $\tau_2 - \tau_1$ are independent exponential variables with parameter 3 and 1, respectively. Finally, let $\{u, v\} = S_{\tau_1}$.

By Proposition 2.3, $\{ij\}$ and $\{ik\}$ are edges of G_{n,r_n} if and only if $P^\star_{\{ij\}} \cap [0, T_{\{ij\}}[$ and $P^\star_{\{ik\}} \cap [0, T_{\{ik\}}[$ are empty, which can also be written

$$\left(P^\star_{\{ij\}} \cap [0, \tau_1[\right) \cup \left(P^\star_{\{ik\}} \cap [0, \tau_1[\right) \cup \left(P^\star_{\{uv\}} \cap [\tau_1, \tau_2[\right) = \emptyset$$

Conditionally on τ_1 and τ_2, by Lemma 2.5, $(P^\star_{\{ij\}} \cap [0, \tau_1[) \cup (P^\star_{\{ik\}} \cap [0, \tau_1[)$ is independent of $P^\star_{\{uv\}} \cap [\tau_1, \tau_2[$, $P^\star_{\{ij\}}$ and $P^\star_{\{ik\}}$ are independent Poisson point processes with rate r_n on $[0, \tau_1[$, and $P^\star_{\{uv\}}$ is a Poisson point process with rate r_n on $[\tau_1, \tau_2[$. Therefore,

$$\mathbb{P}(i \leftrightarrow j, i \leftrightarrow k) = \mathbb{P}(e_1 > \tau_1)\,\mathbb{P}(e_2 > \tau_2 - \tau_1),$$

where $e_1 = \inf(P^\star_{\{ij\}} \cup P^\star_{\{ik\}}) \sim \mathrm{Exp}(2\,r_n)$ is independent of τ_1 and $e_2 = \inf(P^\star_{\{uv\}} \cap [\tau_1, +\infty[) \sim \mathrm{Exp}(r_n)$ is independent of $\tau_2 - \tau_1$. As a result,

$$\mathbb{P}(i \leftrightarrow j, i \leftrightarrow k) = \frac{3}{3 + 2\,r_n} \times \frac{1}{1 + r_n}.$$

A short calculation shows that

$$\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow j\}}, \mathbb{1}_{\{i\leftrightarrow k\}}\Big) = \frac{r_n}{(3+2r_n)(1+r_n)^2}\,,$$

proving the proposition.

As before, the corollary follows from writing the degree of v as $D_n^{(v)} = \sum_{i\neq v} \mathbb{1}_{\{i\leftrightarrow v\}}$, which gives

$$\mathrm{Var}\Big(D_n^{(v)}\Big) = (n-1)\,\mathrm{Var}\Big(\mathbb{1}_{\{i\leftrightarrow v\}}\Big) + (n-1)(n-2)\,\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow v\}}, \mathbb{1}_{\{j\leftrightarrow v\}}\Big)\,.$$

Substituting $\mathrm{Var}\Big(\mathbb{1}_{\{i\leftrightarrow v\}}\Big) = r_n/(1+r_n)^2$ and $\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow v\}}, \mathbb{1}_{\{j\leftrightarrow v\}}\Big)$ yields the desired expression. □

Proposition 3.5. *Let i, j, k and ℓ be four distinct vertices of G_{n,r_n}. We have*

$$\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow j\}}, \mathbb{1}_{\{k\leftrightarrow \ell\}}\Big) = \frac{2\,r_n}{(1+r_n)^2(3+r_n)(3+2\,r_n)}$$

Corollary 3.6. *Let $D_n^{(i)}$ and $D_n^{(j)}$ be the respective degrees of two fixed vertices i and j, and let $|E_n|$ be the number of edges of G_{n,r_n}. We have*

$$\mathrm{Cov}\Big(D_n^{(i)}, D_n^{(j)}\Big) = \frac{r_n}{(1+r_n)^2}\left(1 + \frac{3(n-2)}{3+2\,r_n} + \frac{2(n-2)(n-3)}{(3+r_n)(3+2\,r_n)}\right)$$

and

$$\mathrm{Var}(|E_n|) = \frac{r_n\,n\,(n-1)(n^2+2\,r_n^2+2\,n\,r_n+n+5\,r_n+3)}{2\,(1+r_n)^2\,(3+r_n)\,(3+2\,r_n)}$$

The proof of Proposition 3.5 and its corollary are conceptually identical to the proofs of Propositions 3.1 and 3.3 and their corollaries, but the calculations are more tedious and so they have been relegated to Section B of the Appendix.

3.1.2. Complete subgraphs

From a biological perspective, complete subgraphs are interesting because they are related to how fine the partition of the set of populations into species can be. Indeed, the vertices of a complete subgraph – and especially of a large one – should be considered as part of the same species. A complementary point of view will be brought by connected components in Section 5.

In this section we establish the following results.

Proposition 3.7. *Let $X_{n,k}$ be the number of complete subgraphs of order k in G_{n,r_n}. Then,*

$$\mathbb{E}(X_{n,k}) = \binom{n}{k}\left(\frac{1}{1+r_n}\right)^{k-1}.$$

Corollary 3.8. *Let κ_n be the clique number of G_{n,r_n}, i.e. the maximal number of vertices in a complete subgraph of G_{n,r_n}. If (k_n) is such that*

$$\binom{n}{k_n}\Big(\frac{1}{1+r_n}\Big)^{k_n-1} \xrightarrow[n\to\infty]{} 0\,,$$

then k_n is asymptotically almost surely an upper bound on κ_n, i.e. $\mathbb{P}(\kappa_n \leqslant k_n) \to 1$ as $n \to +\infty$. In particular, when $r_n \to +\infty$,

(i) *If $r_n = o(n)$, then $\kappa_n \leqslant \log(r_n)n/r_n$ a.a.s.*

(ii) *If $r_n = O(n/\log(n))$, $\kappa_n = O_p(n/r_n)$, i.e.*

$$\forall \varepsilon > 0,\ \exists M > 0,\ \exists N \text{ s.t. } \forall n \geqslant N,\ \mathbb{P}(\kappa_n > Mn/r_n) < \varepsilon\,.$$

Proof of Proposition 3.7. The number of complete subgraphs of order k of G_{n,r_n} is

$$X_{n,k} = \sum_{S\in V^{(k)}} \mathbb{1}_{\{G_{n,r_n}[S] \text{ is complete}\}}$$

where the elements of $V^{(k)}$ are the k-subsets of $V = \{1,\dots,n\}$, and $G_{n,r_n}[S]$ is the subgraph of G_{n,r_n} induced by S. By exchangeability,

$$\mathbb{E}(X_{n,k}) = \binom{n}{k}\mathbb{P}\Big(G_{n,r_n}[S] \text{ is complete}\Big)\,,$$

where S is any fixed set of k vertices. Using the notation of Proposition 2.3,

$$G_{n,r_n}[S] \text{ is complete} \iff \forall\{ij\} \in S,\ P^\star_{\{ij\}} = \emptyset\,.$$

For all $t \geqslant 0$, let $A_t = \{a_t(i) : i \in S\}$ be the set of ancestors of S at t. Let $\tau_0 = 0$ and for each $\ell = 1,\dots,k-1$ let τ_ℓ be the time of the ℓ-th coalescence between two lineages of S, i.e.

$$\tau_\ell = \inf\Big\{t > \tau_{\ell-1} : |A_t| = |A_{\tau_{\ell-1}}| - 1\Big\}$$

Finally, let $\widetilde{A}_\ell = A_{\tau_\ell}$ and $I_\ell = [\tau_\ell, \tau_{\ell+1}[$. With this notation,

$$\Big\{\forall\{ij\} \in S, P^\star_{\{ij\}} = \emptyset\Big\} = \bigcap_{\ell=0}^{k-2} B_\ell\,,$$

where

$$B_\ell = \bigcap_{\{ij\}\in\widetilde{A}_\ell^{(2)}} \{P^\star_{\{ij\}} \cap I_\ell = \emptyset\}$$

and $\widetilde{A}_\ell^{(2)}$ denotes the (unordered) pairs of $\widetilde{A}_\ell$. Since for $\ell \neq m$, $I_\ell \cap I_m = \emptyset$, Lemma 2.5 shows that conditionally on $I_0,\dots,I_{k-1}$, the events $B_0,\dots,B_{k-2}$ are independent. By construction, for all $\{ij\} \neq \{uv\}$ in $\widetilde{A}_\ell^{(2)}$,

$$\forall t \in I_\ell,\ \{a_t(i), a_t(j)\} \neq \{a_t(u), a_t(v)\} \neq \emptyset$$

and so it follows from Lemma 2.5 that, conditional on I_ℓ, $(P^\star_{\{ij\}} \cap I_\ell)$, $\{ij\} \in \widetilde{A}^{(2)}_\ell$, are i.i.d. Poisson point processes with rate r_n on I_ℓ. Therefore,

$$\mathbb{P}(B_\ell) = \mathbb{P}\Big(\min\Big\{e^{(\ell)}_{\{ij\}} : \{ij\} \in \widetilde{A}^{(2)}_\ell\Big\} > |I_\ell|\Big)\,,$$

where $e^{(\ell)}_{\{ij\}}$, $\{ij\} \in \widetilde{A}^{(2)}_\ell$, are $\binom{k-\ell}{2}$ i.i.d. exponential variables with parameter r_n that are also independent of $|I_\ell|$. Since $|I_\ell| \sim \mathrm{Exp}\Big(\binom{k-\ell}{2}\Big)$,

$$\mathbb{P}(B_\ell) = \frac{1}{1+r_n}$$

and Proposition 3.7 follows. □

Proof of Corollary 3.8. The first part of the corollary is a direct consequence of Proposition 3.7. First, note that

$$X_{n,k_n} = 0 \iff \kappa_n < k_n$$

that a complete subgraph of order k contains complete subgraphs of order ℓ for all $\ell < k$. As a result, any k_n such that $\mathbb{P}(X_{n,k_n} = 0) \to 1$ is asymptotically almost surely an upper bound on the clique number κ_n. Now, observe that since X_{n,k_n} is a non-negative integer, $X_{n,k_n} \geqslant \mathbb{1}_{\{X_{n,k_n} \neq 0\}}$ and therefore

$$\mathbb{E}(X_{n,k_n}) \geqslant \mathbb{P}(X_{n,k_n} \neq 0)\,.$$

Finally, $X_{n,k}$ being integer-valued, $\mathbb{P}(X_{n,k_n} \neq 0) \to 0$ implies $\mathbb{P}(X_{n,k_n} = 0) \to 1$.

To prove the second part of the corollary, using Stirling's formula we find that whenever r_n and k_n are $o(n)$ and go to $+\infty$ as $n \to +\infty$,

$$\binom{n}{k_n}\Big(\frac{1}{1+r_n}\Big)^{k_n-1} \sim \frac{C}{\sqrt{k_n}}\frac{n^n}{k_n^{k_n}(n-k_n)^{n-k_n}}\Big(\frac{1}{1+r_n}\Big)^{k_n-1},$$

where $C = \sqrt{2\pi}$. The right-hand side goes to zero if and only if its logarithm goes to $-\infty$, i.e. if and only if

$$A_n \coloneqq k_n \log\left(\frac{n-k_n}{k_n(1+r_n)}\right) - n\log\left(1-\frac{k_n}{n}\right) + \log\left(\frac{1+r_n}{\sqrt{k_n}}\right)$$

goes to $-\infty$. Now let $k_n = ng_n/r_n$, where $g_n \to +\infty$ and is $o(r_n)$, so that $k_n = o(n)$. Then,

$$k_n \log\left(\frac{n-k_n}{k_n(1+r_n)}\right) \sim -k_n \log(g_n)$$

and

$$-n\log\left(1-\frac{k_n}{n}\right) \sim k_n\,.$$

Moreover, as long as it does not go to zero,

$$\log\left(\frac{1+r_n}{\sqrt{k_n}}\right) \sim \frac{3}{2}\log(r_n) - \frac{1}{2}\log(ng_n)\,.$$

Putting the pieces together, we find that A_n is asymptotically equivalent to

$$-\frac{ng_n}{r_n}\log(g_n) + \frac{3}{2}\log(r_n) - \frac{1}{2}\log(ng_n)\,.$$

Taking $g_n = \log(r_n)$, this expression goes to $-\infty$ as $n \to +\infty$, yielding (i). If $r_n = O(n/\log(n))$, then it goes to $-\infty$ for any $g_n \to +\infty$, which proves (ii). Indeed, if there exists $\varepsilon > 0$ such that

$$\forall M > 0,\ \forall N,\ \exists n \geqslant N \text{ s.t. } \mathbb{P}(\kappa_n > Mn/r_n) \geqslant \varepsilon\,,$$

then considering successively $M = 1, 2, \ldots$, we can find $n_1 < n_2 < \cdots$ such that

$$\forall k \in \mathbb{N},\ \mathbb{P}(\kappa_{n_k} > kn_k/r_{n_k}) \geqslant \varepsilon\,.$$

Defining (g_n) by

$$\forall n \in \{n_k, \ldots, n_{k+1} - 1\},\ g_n = k\,,$$

we obtain a sequence (g_n) that goes to infinity and yet is such that for all N there exists $n := \min\{n_k : n_k \geqslant N\}$ such that $\mathbb{P}(\kappa_n > g_n n/r_n) \geqslant \varepsilon$. □

A natural pendant to Proposition 3.7 and Corollary 3.8 would be to use the variance of $X_{n,k}$ to find a lower bound on the clique number. Indeed, it follows from Chebychev's inequality that

$$\mathbb{P}(X_{n,k} = 0) \leqslant \frac{\mathrm{Var}(X_{n,k})}{\mathbb{E}(X_{n,k})^2}\,.$$

However, computing $\mathrm{Var}(X_{n,k})$ requires being able to compute the probability that two subsets of k vertices S and S' both induce a complete subgraph, which we have not managed to do. Using the probability that $G_{n,r_n}[S]$ is complete as an upper bound for this quantity, we have the very crude inequality

$$\mathrm{Var}(X_{n,k}) \leqslant \binom{n}{k}^2 p\,(1-p)\,,$$

where $p = 1/(1+r_n)^{k-1}$. This shows that when $r_n \to 0$ and $k_n = o(1/r_n)$, $\mathbb{P}(X_{n,k_n} = 0)$ tends to zero, proving that κ_n is at least $\Theta(1/r_n)$.

Finally, because we expect our model to form dense connected components, whose number we conjecture to be on the order of r_n in the intermediate regime (see Theorem 5.1 and Conjecture 5.4), and since the degree of a typical vertex is approximately n/r_n in that regime, it seems reasonable to conjecture

Conjecture 3.9. *In the intermediate regime, i.e. when $r_n \to +\infty$ and $r_n = o(n)$,*

$$\exists \alpha, \beta > 0 \text{ s.t. } \mathbb{P}(\alpha n/r_n \leqslant \kappa_n \leqslant \beta n/r_n) \xrightarrow[n\to+\infty]{} 1.$$

3.2. Identification of different regimes

We now use the results of the previous section to identify different regimes for the behavior of G_{n,r_n}. The proof of our next theorem relies in part on results proved later in the paper (namely, Theorems 4.1 and 6.1), but no subsequent result depends on it, avoiding cyclic dependencies. While this section could have been placed at the end of the paper, it makes more sense to present it here because it relies mostly on Section 3.1 and because it helps structure the rest of the paper.

Theorem 3.10. *Let D_n be the degree of a fixed vertex of G_{n,r_n}. In the limit as $n \to +\infty$, depending on the asymptotics of r_n we have the following behaviors for G_{n,r_n}*

(i) Transition for the complete graph: *when $r_n = o(1/n)$, $\mathbb{P}(G_{n,r_n}$ is complete$)$ goes to 1, while when $r_n = \omega(1/n)$ it goes to 0; when $r_n = \Theta(1/n)$, this probability is bounded away from 0 and from 1.*

(ii) Dense regime: *when $r_n = o(1)$, $\mathbb{P}(D_n = n-1) \to 1$.*

(iii) Sparse regime: *when $r_n = \omega(n)$, $\mathbb{P}(D_n = 0) \to 1$.*

(iv) Transition for the empty graph: *when $r_n = o(n^2)$, $\mathbb{P}(G_{n,r_n}$ is empty$)$ goes to 0 while when $r_n = \omega(n^2)$ it goes to 1; when $r_n = \Theta(n^2)$, this probability is bounded away from 0 and from 1.*

Proof. (i) is a direct consequence of Proposition 3.7 which, applied to $k = n$, yields

$$\mathbb{P}(G_{n,r_n} \text{ is complete}) = \left(\frac{1}{1+r_n}\right)^{n-1}.$$

(ii) is intuitive since $\mathbb{E}(D_n) = (n-1)/(1+r_n)$; but because it takes $r_n = o(1/n^2)$ for $\mathrm{Var}(D_n)$ to go to zero, a second moment method is not sufficient to prove it. However, using Theorem 4.1, we see that $\mathbb{P}(D_n = n-1)$ can be written as

$$\mathbb{P}(D_n = n-1) = \frac{\Gamma(2+2\,r_n)\Gamma(n+1)}{\Gamma(n+1+2\,r_n)},$$

where Γ is the gamma function. The results follows by letting r_n go to zero and using the continuity of Γ.

(iii) follows from the same argument as in the proof of Corollary 3.8, by which, D_n being a non-negative integer, $\mathbb{P}(D_n \neq 0) \leqslant \mathbb{E}(D_n) = \frac{n-1}{1+r_n}$.

In (iv), the fact that G_{n,r_n} is empty when $r_n = \omega(n^2)$ is yet another application of this argument, but this time using the expected number of edges, $\mathbb{E}(|E_n|) = \frac{n(n-1)}{2(1+r_n)}$, in conjunction with the fact that G_{n,r_n} is empty if and only if $|E_n| = 0$; to see why the graph cannot be empty when $r_n = o(n^2)$, consider the edge that was created between the duplicated vertex and its copy in the most recent duplication. Clearly, if this edge has not disappeared

yet then G_{n,r_n} cannot be empty. But the probability that this edge has disappeared is just

$$\frac{r_n}{\binom{n}{2} + r_n},$$

which goes to zero when $r_n = o(n^2)$. Finally, the fact that $\mathbb{P}(G_{n,r_n}$ is empty) is bounded away from 0 and from 1 when $r_n = \Theta(n^2)$ is a consequence of Theorem 6.1, which shows that the number of edges is Poissonian when $r_n = \omega(n)$. As a result, $\mathbb{P}(|E_n| = 0) \sim e^{-\mathbb{E}(|E_n|)}$. □

Remark 3.11. Note that when $r_n = o(1)$, $\mathrm{Var}(D_n) \sim r_n n^2/3$ can go to infinity even though $D_n = n - 1$ with probability that goes to 1. Similarly, when $r_n = o(1/n)$, $\mathrm{Var}(|E_n|) \sim r_n n^4/18$ and $|E_n| = \binom{n}{2}$ a.a.s. Notably, $\overline{D_n} = (n-1) - D_n$ converges to 0 in probability while $\mathrm{Var}\left(\overline{D_n}\right)$ goes to infinity.

4. The degree distribution

The degree distribution is one of the most widely studied graph invariants in network science. Our model makes it possible to obtain an exact expression for its probability distribution:

Theorem 4.1 (degree distribution)**.** *Let D_n be the degree of a fixed vertex of G_{n,r_n}. Then, for each $k \in \{0, \dots, n-1\}$,*

$$\mathbb{P}(D_n = k) = \frac{2\,r_n\,(2\,r_n + 1)}{(n + 2\,r_n)(n - 1 + 2\,r_n)}\,(k+1)\prod_{i=1}^{k}\frac{n-i}{n-i+2\,r_n-1},$$

where the empty product is 1.

The expression above holds for any positive sequence (r_n) and any n; but as $n \to +\infty$ it becomes much simpler and, under appropriate rescaling, the degree converges to classical distributions:

Theorem 4.2 (convergence of the rescaled degree)**.**

(i) *If $r_n \to r > 0$, then $\frac{D_n}{n}$ converges in distribution to a* $\mathrm{Beta}(2, 2\,r)$ *random variable.*

(ii) *If r_n is both $\omega(1)$ and $o(n)$, then $\frac{D_n}{n/r_n}$ converges in distribution to a size-biased exponential variable with parameter* 2*.*

(iii) *If $2\,r_n/n \to \rho > 0$, then $D_n + 1$ converges in distribution to a size-biased geometric variable with parameter $\rho/(1+\rho)$.*

In this section we prove Theorem 4.1 by coupling the degree to the number of individuals descended from a founder in a branching process with immigration. Theorem 4.2 is then easily deduced by a standard study that has been relegated to Section C of the Appendix.

4.1. Ideas of the proof of Theorem 4.1

Before jumping to the formal proof of Theorem 4.1, we give a verbal account of the main ideas of the proof.

In order to find the degree of a fixed vertex v, we have to consider all pairs $\{iv\}$ and look at their ancestry to assess the absence/presence of atoms in the corresponding Poisson point processes. To do so, we can restrict our attention to the genealogy of the vertices, and consider that edge-removal events occur along the lineages of this genealogy: a point that falls on the lineage of vertex i at time t means that $t \in P^\star_{\{iv\}}$. In this setting, edge-removal events occur at constant rate r_n on every lineage different from that of v.

Next, the closed neighborhood of v (i.e. the set of vertices that are linked to v, plus v itself) can be obtained through the following procedure: we trace the genealogy of vertices, backwards in time; if we encounter an edge-removal event on lineage i at time t, then we mark all vertices that descend from this lineage, i.e. all vertices whose ancestor at time t is i; only the lineages of unmarked vertices are considered after t. We stop when there is only one lineage left in the genealogy. The unmarked vertices are then exactly the neighbors of v (plus v itself). The procedure is illustrated in Figure 5.

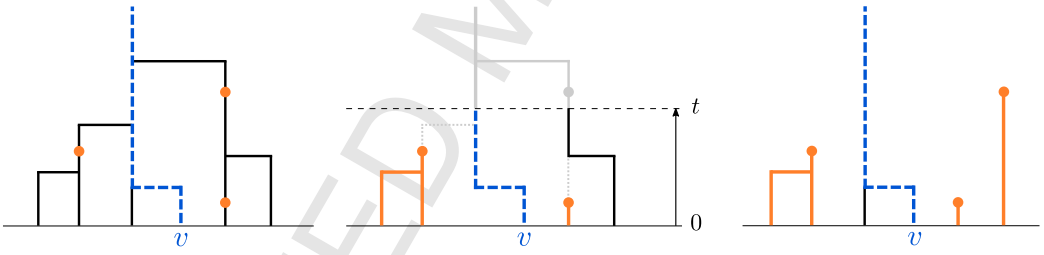

Figure 5: Illustration of the procedure used to find the neighborhood of v. On the left, the genealogy of the vertices. The dashed blue line represents the lineage of the focal vertex v, and a dot on lineage k corresponds to a point in $P_{\{ka_t(v)\}}$. In the middle, we uncover the genealogy and edge-removal events in backward time, as described in the main text. On the right, the forest that we get when the procedure is complete. The non-colored (black) branches are exactly the neighbors of v.

This vertex marking process is not convenient to describe in backward time because we typically mark several vertices simultaneously. By contrast, the forest that results from the completed process seems much easier to describe in forward time. Indeed, the arrival of a new lineage corresponds either to the addition of a new unmarked vertex or to the addition of a marked one, depending on whether the new lineage belongs to the same tree as v or not.

Moreover, in forward time, the process is reminiscent of a branching process with immigration: new lineages are either grafted to existing ones (branching) or sprout spontaneously (immigration). Let us try to find what the branching and immigration rates should be. In backward time, when there are $k+1$ lineages then a coalescence occurs at rate $\binom{k+1}{2}$, while an edge-removal event occurs at rate $k\,r_n$. Reversing time, these events occur

at the same rates. As a result, when going from k to $k+1$ lineages, the probability that the next event is a branching is $(k+1)/(k+1+2\,r_n)$.

Next, we have to find the probability that each lineage has to branch, given that the next event is a branching. Here, a spinal decomposition [13, 14] suggests that every lineage branches at rate 1, except for the lineage of v, which branches at rate 2. To see why, observe that this is coherent with the fact that, in backward time, when going from $k+1$ to k lineages there are k pairs out of $\binom{k+1}{2}$ that involve the lineage of v, so that the probability that the lineage of v is involved in the next coalescence is $2/(k+1)$.

If this heuristic is correct, then in forward time it is easy to track the number of branches of the tree of v versus the number of branches of other trees: when there are p branches in the tree of v and q branches in the other trees, the probability that the next branch is added to the tree of v is just $(p+1)/(p+1+q+2\,r_n)$. Moreover, when the total number of branches reaches n, the number of branches in the tree of v is also the number of unmarked vertices at the end of the vertex marking procedure, which is itself $D_n^{(v)}+1$, the degree of v plus one.

4.2. Formal proof of Theorem 4.1

The ideas and outline of the proof parallels the account given in the previous section: first, given a realization of the vertex-duplication process $\mathcal{M}$ and of the edge-removal process $\mathcal{P}$, we describe a deterministic procedure that gives the closed neighborhood of any vertex v,

$$N_G[v] = \Big\{ i \in V : \{iv\} \in E \Big\} \cup \Big\{ v \Big\},$$

where $G=(V,E)$ is the graph associated to $\mathcal{M}$ and $\mathcal{P}$; then, we identify the law of the process $(F_t)_{t\geqslant 0}$ corresponding to this procedure, and recognize it as the law of a branching process with immigration.

Definition 4.3. A *rooted forest with marked vertices* is a triple $F=(V^\circ, V^\bullet, \vec{E})$ such that

(i) $V^\circ \cap V^\bullet = \emptyset$.

(ii) Letting $V = V^\circ \cup V^\bullet$, $(V,\vec{E})$ is an acyclic digraph with the property that $\forall i \in V$, $\deg^+(i) \in \{0,1\}$, where $\deg^+(i)$ is the out-degree of vertex i.

The marked vertices are the elements of $V^\bullet$; the roots of F are the vertices with out-degree 0 (that is, edges are oriented towards the root), whose set we denote by $R(F)$; finally, the trees of F are its connected components (in the weak sense, i.e. considering the underlying undirected graph), and we write $T_F(i)$ for the tree containing i in F.

4.2.1. The vertex-marking process

We now define the backward-time process $(F_t)_{t\geqslant 0}$ that corresponds to the procedure described informally in Section 4.1. Recall the notation of Proposition 2.3. For a given realization of $\mathcal{M}$ and $\mathcal{P}$, and for any fixed vertex v, let $(F_t)_{t\geqslant 0}$ be the piecewise constant process defined deterministically by

- $F_0 = (V, \emptyset, \emptyset)$.
- If $t \in M_{(ij)}$, then $\forall k, \ell \in R(F_{t-}) \cap V^\circ_{t-}$ such that $(a_{t-}(k), a_{t-}(\ell)) = (i, j)$,
$$\vec{E}_t = \vec{E}_{t-} \cup \Big\{(k, \ell)\Big\}.$$
- If $t \in P_{\{ia_t(v)\}}$, then letting $d_t(i) = \{j \in V : a_t(j) = i\}$ be the set of descendants of i born after time t,
$$\begin{cases} V^\circ_t = V^\circ_{t-} \setminus d_t(i) \\ V^\bullet_t = V^\bullet_{t-} \cup d_t(i)\,. \end{cases}$$

What makes $(F_t)_{t\geqslant 0}$ interesting is that

$$N_G[v] = V^\circ_\infty\,.$$

Indeed, by construction,

$$i \in V^\circ_t \iff \bigcup_{s\in[0,t]} \left(\bigcup_{j : i\in d_s(j)} P_{\{ja_s(v)\}} \right) = \emptyset\,,$$

and since for every s the unique j such that $i \in d_s(j)$ is $a_s(i)$, we have

$$V^\circ_t = \Big\{ i \in V : P^\star_{\{iv\}} \cap [0, t] = \emptyset \Big\}\,.$$

The Poissonian construction given above shows that $(F_t, a_t)_{t\geqslant 0}$ is a Markov process. Now, observe that conditional on a_t

(i) $M_{(ij)} \cap]t, +\infty[\;\sim\; M_{(a_t(i)a_t(j))} \cap]t, +\infty[$ and is independent of $(F_s, a_s)_{s\leqslant t}$

(ii) $P_{\{ia_t(v)\}} \cap]t, +\infty[\;\sim\; P_{\{a_t(i)a_t(v)\}} \cap]t, +\infty[$ and is independent of $(F_s, a_s)_{s\leqslant t}$

(iii) $j \in d_t(i) \iff i \in R(F_t)$ and $j \in T_{F_t}(i)$

As a consequence, $(F_t)_{t\geqslant 0}$ is also a Markov process, whose law is characterized by

- $F_0 = (V, \emptyset, \emptyset)$.
- F_t goes from $(V^\circ_t, V^\bullet_t, \vec{E}_t)$ to
  - $\Big(V^\circ_t,\, V^\bullet_t,\, \vec{E}_t \cup \{(i, j)\}\Big)$ at rate $1/2$, for all i, j in $R(F_t)$
  - $\Big(V^\circ_t \setminus T_{F_t}(i),\, V^\bullet_t \cup T_{F_t}(i),\, \vec{E}_t\Big)$ at rate r_n, for all i in $R(F_t)$.

Let $(\widetilde{F}_k)_{k\in\{1,\dots,n\}}$ be the chain embedded in $(F_t)_{t\geqslant 0}$, i.e. defined by

$$\widetilde{F}_k = F_{t_k}, \text{ where } t_k = \inf\Big\{t \geqslant 0 : |R(F_t)| = n-k+1\Big\}.$$

The rooted forests with marked vertices that correspond to realizations of $\widetilde{F}_n$ are exactly the $f_n = (V^\circ, V^\bullet, \vec{E})$ that have n vertices and are such that $V^\circ = T_{f_n}(v)$. Moreover, for each of these there exists a unique trajectory $(f_1, \dots, f_n)$ of $(\widetilde{F}_1, \dots, \widetilde{F}_n)$ such that $\widetilde{F}_n = f_n$ and it follows from the transition rates of $(F_t)_{t\geqslant 0}$ that

$$\begin{aligned}\mathbb{P}\Big(\widetilde{F}_n = f_n\Big) &= \frac{(1/2)^{n-|R(f_n)|}\, r_n^{|R(f_n)|-1}}{\prod_{k=2}^{n}\big(k(k-1)/2 + (k-1)r_n\big)} \\ &= \frac{1}{(n-1)!} \times \frac{(2\,r_n)^{|R(f_n)|-1}}{\prod_{k=2}^{n}(k+2\,r_n)} \qquad (1)\end{aligned}$$

Finally, note that $\widetilde{V}_n^\circ = V_\infty^\circ$ is the closed neighborhood of v in our graph.

4.2.2. The branching process

The process with which we will couple the vertex-marking process described in the previous section is a simple random function of the trajectories of a branching process with immigration $(Z_t)_{t\geqslant 0}$. In this branching process, immigration occurs at rate $2\,r_n$ and each particle gives birth to a new particle at rate 1 – except for one particle, which carries a special item that enables it to give birth at rate 2; when this lineage reproduces, it keeps the item with probability 1/2, and passes it to its offspring with probability 1/2.

Formally, we consider the Markov process on the set of rooted forests with marked vertices (augmented with an indication of the carrier of the item), defined by $Z_0 = (\{1\}, \emptyset, \emptyset, 1)$ and by the following transition rates:

$(Z_t)_{t\geqslant 0}$ goes from $(W^\circ,\, W^\bullet,\, \vec{E},\, c)$ to

- $\Big(W^\circ \cup \{N\},\, W^\bullet,\, \vec{E} \cup \{(N,i)\},\, c\Big)$ at rate 1, for all $i \in W^\circ$
- $\Big(W^\circ,\, W^\bullet \cup \{N\},\, \vec{E} \cup \{(N,i)\},\, c\Big)$ at rate 1, for all $i \in W^\bullet$
- $\Big(W^\circ \cup \{N\},\, W^\bullet,\, \vec{E} \cup \{(N,c)\},\, N\Big)$ at rate 1
- $\Big(W^\circ,\, W^\bullet \cup \{N\},\, \vec{E},\, c\Big)$ at rate $2\,r_n$

where $N = |W^\circ \cup W^\bullet| + 1$ is the label of the new particle. The fourth coordinate of $(Z_t)_{t\geqslant 0}$ tracks the carrier of the item.

As previously, the Markov chain $(\widetilde{Z}_k)_{k\in\mathbb{N}^*}$ embedded in $(Z_t)_{t\geqslant 0}$ is defined by

$$\widetilde{Z}_k = Z_{t_k}, \text{ where } t_k = \inf\Big\{t \geqslant 0 : |W_t^\circ \cup W_t^\bullet| = k\Big\}.$$

The realizations of $\widetilde{Z}_n$ are exactly the $(W_n^\circ, W_n^\bullet, \vec{E}_n, c_n)$ such that $f_n = (W_n^\circ, W_n^\bullet, \vec{E}_n)$ is a rooted forest with marked vertices on $\{1, \dots, n\}$ and $W_n^\circ = T_{f_n}(1) = T_{f_n}(c_n)$. For these, it follows from the transition rates of $(Z_t)_{t\geqslant 0}$ that

$$\mathbb{P}\Big(\widetilde{Z}_n = (W_n^\circ, W_n^\bullet, \vec{E}_n, c_n)\Big) = \frac{(2\,r_n)^{|R(f_n)|-1}}{\prod_{k=1}^{n-1}(k+1+2\,r_n)}\,. \tag{2}$$

Finally, note that $(X_k)_{k\in\mathbb{N}^*} = \big(|\widetilde{W}_k^\circ|, |\widetilde{W}_k^\bullet|\big)_{k\in\mathbb{N}^*}$, which counts the number of descendants of the first particle and the number of descendants of immigrants, is a Markov chain whose law is characterized by $X_1 = (1,0)$ and X_k goes from (p,q) to

- $(p+1, q)$ with probability $\frac{p+1}{p+1+q+2r_n}$
- $(p, q+1)$ with probability $\frac{q+2r}{p+1+q+2r_n}$.

4.2.3. Relabeling and end of proof

The last step before finishing the proof of Theorem 4.1 is to shuffle the vertices of the forest associated to $\widetilde{Z}_n$ appropriately. For any fixed n, v and c in $\{1, \dots, n\}$, let $\Phi_{(c,v)}$ be uniformly and independently of anything else picked among the permutations of $\{1, \dots, n\}$ that map c to v; define $\Phi_v(\widetilde{Z}_n)$ by

$$\Phi_v\Big(\widetilde{W}_n^\circ, \widetilde{W}_n^\bullet, \widetilde{E}_n, \tilde{c}_n\Big) = \Big(\Phi_{(\tilde{c}_n,v)}(\widetilde{W}_n^\circ),\ \Phi_{(\tilde{c}_n,v)}(\widetilde{W}_n^\bullet),\ \Phi_{(\tilde{c}_n,v)}(\widetilde{E}_n)\Big)$$

where $\Phi_{(\tilde{c}_n,v)}(\widetilde{E}_n)$ is to be understood as $\Big\{\Big(\Phi_{(\tilde{c}_n,v)}(i), \Phi_{(\tilde{c}_n,v)}(j)\Big) : (i,j) \in \widetilde{E}_n\Big\}$.

With all these elements, the proof of Theorem 4.1 goes as follows. First, from equations (1) and (2) and the definition of Φ_v, we see that for all rooted forest with marked vertices f_n,

$$\mathbb{P}\Big(\widetilde{F}_n = f_n\Big) = \mathbb{P}\Big(\Phi_v(\widetilde{Z}_n) = f_n\Big)\,.$$

In particular, $\widetilde{V}_n^\circ$, the set of unmarked vertices in the vertex-marking process, and $\Phi_{(\tilde{c}_n,v)}(\widetilde{W}_n^\circ)$, the relabeled set of descendants of the first particle in the branching process, have the same law. Now, on the one hand we have

$$\Big|\widetilde{V}_n^\circ\Big| = \Big|N_G[v]\Big| = D_n^{(v)} + 1\,,$$

and on the other hand we have

$$\Big|\Phi_{(\tilde{c}_n,v)}(\widetilde{W}_n^\circ)\Big| = \Big|\widetilde{W}_n^\circ\Big|\,.$$

Since $\Big|\widetilde{W}_n^\circ\Big|$ is the first coordinate of the Markov chain $(X_k)_{k\in\mathbb{N}^*}$ introduced in the previous section, it follows directly from the transition probabilities of $(X_k)_{k\in\mathbb{N}^*}$ that

$$\mathbb{P}\Big(X_n = (k+1, n-k-1)\Big) = \binom{n-1}{k} \frac{\prod_{p=1}^{k}(p+1)\ \prod_{q=0}^{n-k-2}(q+2\,r_n)}{\prod_{(p+q)=1}^{n-1}\Big((p+q)+1+2\,r_n\Big)}\,,$$

from which the expression of Theorem 4.1 can be deduced through elementary calculations.

5. Connected components in the intermediate regime

From a biological perspective, connected components are good candidates to define species, and have frequently been used to that end. Moreover, among the possible definitions of species, they play a special role because they indicate how coarse the partition of the set of populations into species can be; indeed, it would not make sense biologically for distinct connected components to be part of the same species. As a result, connected components are in a sense the "loosest" possible definition of species. This complements the perspective brought by complete subgraphs, which inform us on how fine the species partition can be (see Section 3.1.2). For a discussion of the definition of species in a context where traits and ancestral relationships between individuals are known, see [15].

The aim of this section is to prove the following theorem.

Theorem 5.1. *Let $\#\mathrm{CC}_n$ be the number of connected components of G_{n,r_n}. If r_n is both $\omega(1)$ and $o(n)$, then*

$$\frac{r_n}{2} + o_p(r_n) \leqslant \#\mathrm{CC}_n \leqslant 2\, r_n \log n + o_p(r_n \log n)$$

where, for a positive sequence (u_n), $o_p(u_n)$ denotes a given sequence of random variables (X_n) such that $X_n/u_n \to 0$ in probability.

5.1. Lower bound on the number of connected components

The proof of the lower bound on the number of connected components uses the forward construction introduced in Section 2.2 and the associated notation. It relies on the simple observation that, letting $\#\mathrm{CC}(G)$ denote the number of connected components of a graph G, $\#\mathrm{CC}(G^\star_{r_n}(k))$ is a nondecreasing function of k. Indeed, in the sequence of events defining $(G^\star_{r_n}(k))_{k\geqslant 2}$, vertex duplications do not change the number of connected components – because a new vertex is always linked to an existing vertex (its 'mother') and her neighbors – and edge removals can only increase it. Thus, if $m_n \leqslant n$ and ℓ_n are such that $\mathbb{P}\left(\#\mathrm{CC}(G^\star_{r_n}(m_n) \geqslant \ell_n\right) \to 1$ as $n \to \infty$, then ℓ_n is asymptotically almost surely a lower bound on the number of connected components of $G^\star_{r_n}(n)$ — and therefore of G_{n,r_n}.

To find such a pair (m_n, ℓ_n), note that, for every graph G of order m,

$$\#\mathrm{CC}(G) \geqslant m - \#\mathrm{edges}(G)\,.$$

Moreover, since for any fixed n, $G^\star_{r_n}(m_n)$ has the same law as G_{m_n,r_n}, the exact expressions for the expectation and the variance of $|E^\star_{m_n}|$, the number

of edges of $G^\star_{r_n}(m_n)$, are given in Table 1. We see that, if r_n and m_n are both $\omega(1)$ and $o(n)$,

$$\mathbb{E}\Big(|E^\star_{m_n}|\Big) \sim \frac{m_n^2}{2\,r_n} \quad \text{and} \quad \mathrm{Var}\Big(|E^\star_{m_n}|\Big) \sim \frac{m_n^2}{4\,r_n^3}\Big(m_n^2 + 2\,r_n^2\Big)\,.$$

By Chebychev's inequality,

$$\mathbb{P}\Big(\,\Big||E^\star_{m_n}| - \mathbb{E}\Big(|E^\star_{m_n}|\Big)\Big| \geqslant m_n^{1-\varepsilon}\Big) \;\leqslant\; \frac{\mathrm{Var}\Big(|E^\star_{m_n}|\Big)}{m_n^{2-2\varepsilon}}\,.$$

When $m_n = \Theta(r_n)$, since $r_n = \omega(1)$ the right-hand side of this inequality goes to 0 as $n \to +\infty$, for all $\varepsilon < 1/2$. It follows that

$$|E^\star_{m_n}| = \mathbb{E}\Big(|E^\star_{m_n}|\Big) + o_p(r_n)\,.$$

Taking $m_n \coloneqq \lfloor \alpha\, r_n \rfloor$, we find that

$$\#\mathrm{CC}(G^\star_{r_n}(m_n)) \geqslant m_n - |E^\star_{m_n}| = \alpha\left(1 - \frac{\alpha}{2}\right) r_n + o_p(r_n)\,.$$

The right-hand side is maximal for $\alpha = 1$ and is then $r_n/2 + o_p(r_n)$.

5.2. Upper bound on the number of connected components

Our strategy to get an upper bound on the number of connected components is to find a spanning subgraph whose number of connected components we can estimate. A natural idea is to look for a spanning forest, because forests have the property that their number of connected components is their number of vertices minus their number of edges.

Definition 5.2. A pair of vertices $\{ij\}$ is said to be a *founder* if it has no ancestor other than itself, i.e., letting $T_{\{ij\}} = \sup\{t \geqslant 0 : a_t(i) \neq a_t(j)\}$ be the coalescence time of i and j, $\{ij\}$ is a founder if and only if $\forall t < T_{\{ij\}}$, $\{a_t(i)\, a_t(j)\} = \{ij\}$.

Let $\mathcal{F}$ be the set of founders of $G_{n,r_n} = (V, E)$, and let $T_n = (V, \mathcal{F})$. Note that $\#\mathcal{F} = n - 1$ and that T_n is a tree. Therefore, letting $F_n = (V, \mathcal{F} \cap E)$ be the spanning forest of G_{n,r_n} induced by T_n, we have

$$\#\mathrm{CC}_n \leqslant n - \#\mathrm{edges}(F_n)\,.$$

Let us estimate the number of edges of F_n. Recall Proposition 2.3. By construction, $\forall \{ij\} \in \mathcal{F}$, $P^\star_{\{ij\}} = P_{\{ij\}} \cap [0, T_{\{ij\}}]$. It follows that

$$\#\mathrm{edges}(F_n) = \sum_{\{ij\}\in\mathcal{F}} \mathbb{1}_{\left\{P_{\{ij\}}\cap[0,T_{\{ij\}}]=\emptyset\right\}}$$

and, as a consequence,

$$\#\mathrm{CC}_n \leqslant 1 + \sum_{\{ij\}\in\mathcal{F}} \mathbb{1}_{\left\{P_{\{ij\}}\cap[0,T_{\{ij\}}]\neq\emptyset\right\}}\,.$$

Now, $\mathbb{1}_{\{P_{\{ij\}}\cap[0,T_{\{ij\}}]\neq\emptyset\}} \leqslant \#(P_{\{ij\}} \cap [0, T_{\{ij\}}])$, and since $(P_{\{ij\}})_{\{ij\}\in\mathcal{F}}$ are i.i.d. Poisson point processes with intensity r_n that are also independent of $(T_{\{ij\}})_{\{ij\}\in\mathcal{F}}$,

$$\sum_{\{ij\}\in\mathcal{F}} \#(P_{\{ij\}} \cap [0, T_{\{ij\}}]) \leqslant \#(P \cap [0, L_n])\,,$$

where P is a Poisson point process on $]0, +\infty]$ with intensity r_n and $L_n = T_{\mathrm{MRCA}} + \sum_{\{ij\}\in\mathcal{F}} T_{\{ij\}}$ is the total branch length of the genealogy of the vertices. Putting the pieces together,

$$\#\mathrm{CC}_n \leqslant 1 + \#(P \cap [0, L_n])\,.$$

Conditional on L_n, $\#(P\cap[0, L_n])$ is a Poisson random variable with parameter $r_n L_n$. Moreover, it is known [16] that

$$\mathbb{E}(L_n) = 2\sum_{i=1}^{n-1} \frac{1}{i} \quad \text{and} \quad \mathrm{Var}(L_n) = 4\sum_{i=1}^{n-1} \frac{1}{i^2}$$

As a result,

$$\mathbb{E}(\#(P \cap [0, L_n])) = r_n \mathbb{E}(L_n) \sim 2\, r_n \log n$$

and

$$\mathrm{Var}(\#(P \cap [0, L_n])) = r_n \mathbb{E}(L_n) + \mathrm{Var}(r_n L_n) \sim 2\, r_n \log n + \alpha\, r_n^2\,,$$

with $\alpha = 2\pi^2/3$. Using Chebychev's inequality, we find that for all $\varepsilon > 0$,

$$\mathbb{P}\Big(|\#(P \cap [0, L_n]) - 2\, r_n \log n| \geqslant \varepsilon\, r_n \log n\Big) = O\bigg(\frac{2}{\varepsilon^2\, r_n \log(n)} + \frac{\alpha}{\varepsilon^2 \log(n)^2}\bigg).$$

The right-hand side goes to 0 as $n \to +\infty$, which shows that $\#(P\cap[0, L_n]) - 2\, r_n \log n = o_p(r_n \log n)$ and finishes the proof.

Remark 5.3. Using $\#(P\cap[0, L_n])$ as an upper bound for $\sum_{\{ij\}\in\mathcal{F}} \mathbb{1}_{\{P_{\{ij\}}\cap[0,T_{\{ij\}}]\neq\emptyset\}}$ turns out not to be a great source of imprecision, because most of the total branch length of a Kingman coalescent comes from very short branches. As a result, when $r_n = o(n)$, only a negligible proportion of the $P_{\{ij\}}\cap[0, T_{\{ij\}}]$'s, $\{ij\} \in \mathcal{F}$, have more than one point.

By contrast, using $n - \#\mathrm{edges}(F_n)$ as an upper bound on $\#\mathrm{CC}_n$ is very crude. This leads us to formulate the following conjecture:

Conjecture 5.4.

$$\exists \alpha, \beta > 0 \ \textit{s.t.}\ \mathbb{P}(\alpha r_n \leqslant \#\mathrm{CC}_n \leqslant \beta r_n) \xrightarrow[n\to\infty]{} 1.$$

6. Number of edges in the sparse regime

From the expressions obtained in section 3.1.1 and recapitulated in Table 1, we see that when $r_n = \omega(n)$,

$$\mathbb{E}(|E_n|) \sim \mathrm{Var}(|E_n|) \sim \frac{n^2}{2\,r_n}\,.$$

This suggests that the number of edges is Poissonian in the sparse regime, and this is what the next theorem states.

Theorem 6.1. *Let $|E_n|$ be the number of edges of G_{n,r_n}. If $r_n = \omega(n)$ then*

$$d_{\mathrm{TV}}\Big(|E_n|, \mathrm{Poisson}(\lambda_n)\Big) \xrightarrow[n\to+\infty]{} 0\,,$$

where d_{TV} stands for the total variation distance and $\lambda_n = \mathbb{E}(|E_n|) \sim \frac{n^2}{2r_n}$. If in addition $r_n = o(n^2)$, then $\lambda_n \to +\infty$ and as a result

$$\frac{|E_n| - \lambda_n}{\sqrt{\lambda_n}} \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}(0,1)\,,$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution.

The proof of Theorem 6.1 is a standard application of the Stein–Chen method [17, 18]. A reference on the topic is [19], and another excellent survey is given in [20]. Let us state briefly the results that we will need.

Definition A. The Bernoulli variables $X_1, \dots, X_N$ are said to be *positively related* if for each $i = 1, \dots, N$ there exists $(X_1^{(i)}, \dots, X_N^{(i)})$, built on the same space as $(X_1, \dots, X_N)$, such that

(i) $\big(X_1^{(i)}, \dots, X_N^{(i)}\big) \sim (X_1, \dots, X_N) \mid X_i = 1.$

(ii) For all $j = 1, \dots N$, $X_j^{(i)} \geqslant X_j$.

Note that there are other equivalent definitions of positive relation (see e.g. Lemma 4.27 in [20]). Finally, we will need the following classic theorem, which appears, e.g., as Theorem 4.20 in [20].

Theorem A. *Let $X_1, \dots, X_N$ be positively related Bernoulli variables with $\mathbb{P}(X_i = 1) = p_i$. Let $W = \sum_{i=1}^N X_i$ and $\lambda = \mathbb{E}(W)$. Then,*

$$d_{\mathrm{TV}}(W, \mathrm{Poisson}(\lambda)) \leqslant \min\{1, \lambda^{-1}\} \left(\mathrm{Var}(W) - \lambda + 2\sum_{i=1}^N p_i^2\right).$$

6.1. Proof of the positive relation between the edges

It is intuitive that the variables indicating the presence of edges in our graph are positively related, because the only way through which these variables depend on each other is through the fact that the edges share ancestors. Our proof is nevertheless technical.

6.1.1. Preliminary lemmas

In this section we isolate the proof of two useful results that are not tied to the particular setting of our model.

Lemma 6.2. *Let* $\mathbf{X} = (X_1, \dots, X_N)$ *be a vector of Bernoulli variables. The distribution of* $\mathbf{X}$ *is uniquely characterized by the quantities*

$$\mathbb{E}\left(\prod_{i\in I} X_i\right), \quad I \subset \{1, \dots, N\},\, I \neq \emptyset$$

Proof. For all $I \subset \{1, \dots, N\}$, $I \neq \emptyset$, let

$$p_I = \mathbb{E}\left(\prod_{i\in I} X_i\right) \quad \text{and} \quad q_I = \mathbb{E}\left(\prod_{i\in I} X_i \prod_{j\in I^{\mathrm{c}}} (1 - X_j)\right)$$

where the empty product is understood to be 1.

Clearly, the distribution of $\mathbf{X}$ is fully specified by (q_I). Now observe that, by the inclusion-exclusion principle,

$$q_I = \sum_{J\supset I} (-1)^{|J|-|I|}\, p_J\,,$$

which terminates the proof. □

Lemma 6.3. *Let* $X_1, \dots, X_N$ *be independent random nondecreasing functions from* $[0, +\infty[$ *to* $\{0, 1\}$ *such that*

$$\forall i \in \{1, \dots, N\}, \quad \inf\{t \geqslant 0 : X_i(t) = 1\} < +\infty \text{ almost surely.}$$

Let T *be a non-negative random variable that is independent of* $(X_1, \dots, X_N)$. *Then,* $X_1(T), \dots, X_N(T)$ *are positively related.*

Proof. Pick $i \in \{1, \dots, N\}$. Now, let $\tau_i = \inf\{t \geqslant 0 : X_i(t) = 1\}$. Assume without loss of generality that X_i is left-continuous, so that $\{X_i(T) = 1\} = \{T > \tau_i\}$. Next, note that,

$$\forall x, t \geqslant 0, \quad \mathbb{P}(T > x, T > t) \geqslant \mathbb{P}(T > x)\mathbb{P}(T > t)\,.$$

Integrating in t against the law of τ_i, we find that

$$\forall x \geqslant 0, \quad \mathbb{P}(T > x \,|\, T > \tau_i) \geqslant \mathbb{P}(T > x)\,.$$

This shows that T is stochastically dominated by $T^{(i)}$, where $T^{(i)}$ has the law of T conditioned on $\{T > \tau_i\}$. As a result, there exists S, built on the same space as $X_1, \dots, X_N$ and independent of $(X_j)_{j\neq i}$, such that $S \sim T^{(i)}$ and $S \geqslant T$. For all $j \neq i$, let $X_j^{(i)} = X_j(S)$. Since X_j is nondecreasing, $X_j^{(i)} \geqslant X_j(T)$, and since $(X_j)_{j\neq i} \perp\!\!\!\perp (T, \tau_i)$, $(X_j^{(i)})_{j\neq i} \sim ((X_j(T))_{j\neq i} \mid X_i(T) = 1)$. This shows that $X_1(T), \dots, X_N(T)$ are positively related. □

Remark 6.4. Lemma 6.3 and its proof are easily adapted to the case where $X_1, \dots, X_N$ are nonincreasing and such that $\inf\{t \geqslant 0 : X_i(t) = 0\} < +\infty$ almost surely.

6.1.2. Stein–Chen coupling

Proposition 6.5. *For any $n \geqslant 2$ and $r > 0$, the random variables $\mathbb{1}_{\{i \leftrightarrow j\}}$ for $\{ij\} \in V^{(2)}$, which indicate the presence of edges in $G_{n,r}$, are positively related.*

Proof. We use the forward construction described in Section 2.3 and proceed by induction. To keep the notation light, throughout the rest of the proof we index the pairs of vertices of $G_r^\star(n) = (\{1, \dots, n\}, E_n^\star)$ by the integers from 1 to $N = \binom{n}{2}$ and, for $i \in \{1, \dots, N\}$, we let $X_i = \mathbb{1}_{\{i \in E_n^\star\}}$. We also make consistent use of bold letters to denote vectors, i.e., given any family of random variables $Z_1, \dots, Z_p$, we write $\mathbf{Z}$ for $(Z_1, \dots, Z_p)$.

For $n = 2$, the family X_i for $i \in \{1, \dots, n\}$ consists of a single variable X_1, so it is trivially positively related.

Now assume that $X_1, \dots, X_N$ are positively related in $G_r^\star(n)$, i.e.

$$\begin{aligned} &\forall i \leqslant N, \exists \mathbf{Y}^{(i)} = \left(Y_1^{(i)}, \dots, Y_N^{(i)}\right) \text{ such that} \\ &\quad \text{(i)} \quad \mathbf{Y}^{(i)} \sim (\mathbf{X} \mid X_i = 1) \\ &\quad \text{(ii)} \quad \forall k \leqslant N,\ Y_k^{(i)} \geqslant X_k \end{aligned} \tag{3}$$

Remember that $G_r^\star(n+1)$ is obtained by (1) adding a vertex to $G_r^\star(n)$ (which, without loss of generality, we label $n+1$) and linking it to a uniformly chosen vertex u_n of $G_r^\star(n)$ as well as to the neighbors of u_n; and (2) waiting for an exponential time T with parameter $\binom{n}{2}$ while removing each edge at constant rate r.

Formally, $\forall k \leqslant N + n$, define the "mother" of k, $M_k \in \{1, \dots, N\} \cup \{\emptyset\}$, by

- If $k \leqslant N$ (i.e., if k is the label of $\{u, v\}$, with $1 \leqslant u < v \leqslant n$), then $M_k = k$.
- If $k > N$ is the label of $\{v, n+1\}$, with $1 \leqslant v \leqslant n$, then $M_k = \ell$, where ℓ is the label of $\{u_n, v\}$.
- If $k > N$ is the label of $\{u_n, n+1\}$, then $M_k = \emptyset$.

Letting $X'_k = \mathbb{1}_{\left\{k \in E_{n+1}^\star\right\}}$, we then have

$$X'_k = \begin{cases} A_k & \text{if } M_k = \emptyset \\ X_{M_k} A_k & \text{otherwise} \end{cases}$$

with $A_k = \mathbb{1}_{\{e_k > T\}}$, where we recall that $T \sim \mathrm{Exp}(N)$ and, $e_1, \dots, e_{N+n}$ are i.i.d. exponential variables with parameter r that are also independent of everything else.

Note that the random functions $\widetilde{A}_k \colon t \mapsto \mathbb{1}_{\{e_k > t\}}$, $k \in \{1, \dots N + n\}$ are nonincreasing and such that $\inf\{t \geqslant 0 : \widetilde{A}_k(t) = 0\} < +\infty$ almost surely. By

Lemma 6.3 (see also Remark 6.4), it follows that $A_1, \dots, A_{N+n}$ are positively related.

We now pick any $i \leqslant \binom{n+1}{2} = N+n$ and build a vector $\mathbf{Y}'^{(i)}$ that has the same law as $(\mathbf{X}' \mid X_i = 1)$ and satisfies $\mathbf{Y}'^{(i)} \geqslant \mathbf{X}'$.

Assume that $M_i \neq \emptyset$. In that case,

1. By the induction hypothesis, there exists $\mathbf{Y}^{(M_i)}$ that satisfies (3).
2. Since by $A_1, \dots, A_{N+n}$ are positively related, $\exists \mathbf{B}^{(i)} \sim (\mathbf{A} \mid A_i = 1)$ such that $\mathbf{B}^{(i)} \geqslant \mathbf{A}$.

Note that $\mathbf{A}$, $\mathbf{B}^{(i)}$, $\mathbf{X}$ and $\mathbf{Y}^{(M_i)}$ are all built on the same space. Therefore, omitting the (M_i) and (i) superscripts to keep the notation light, we can set $Y'_i = 1$ and, for $k \neq i$,

$$Y'_k = \begin{cases} B_k & \text{if } M_k = \emptyset \\ Y_{M_k} B_k & \text{otherwise.} \end{cases}$$

With this definition, $\forall J \subset \{1, \dots, N+n\}$,

$$\mathbb{E}\left(\prod_{j\in J} Y'_j\right) = \mathbb{E}\left(\prod_{j\in \widetilde{J}} Y_j\right) \mathbb{E}\left(\prod_{j\in J} B_j\right),$$

where $\widetilde{J} = \{M_j : j \in J, M_j \neq \emptyset\}$. By hypothesis,

$$\mathbb{E}\left(\prod_{j\in \widetilde{J}} Y_j\right) = \mathbb{E}\left(\prod_{j\in \widetilde{J}} X_j \,\middle|\, X_{M_i} = 1\right) = \mathbb{E}\Big(X_{M_i} \prod_{j\in \widetilde{J}} X_j\Big) \,/\, \mathbb{E}\big(X_{M_i}\big)$$

Similarly,

$$\mathbb{E}\left(\prod_{j\in J} B_j\right) = \mathbb{E}\left(A_i \prod_{j\in J} A_j\right) \,/\, \mathbb{E}\big(A_i\big).$$

As a result,

$$\begin{aligned} \mathbb{E}\left(\prod_{j\in J} Y'_j\right) &= \frac{\mathbb{E}\big(X_{M_i} \prod_{j\in \widetilde{J}} X_j\big)\mathbb{E}\big(A_i \prod_{j\in J} A_j\big)}{\mathbb{E}(X_{M_i})\mathbb{E}(A_i)} \\ &= \frac{\mathbb{E}\big(X_{M_i} A_i \prod_{j\in J} X'_j\big)}{\mathbb{E}(X_{M_i} A_i)} \\ &= \mathbb{E}\left(\prod_{j\in J} X'_j \,\middle|\, X'_i = 1\right) \end{aligned}$$

By Lemma 6.2, this shows that $\mathbf{Y}' \sim (\mathbf{X}' \mid X'_i = 1)$.

If $M_i = \emptyset$, we can no longer choose $\mathbf{Y}^{(M_i)}$. However, in that case, X'_i depends only on A_i. Therefore, we set $Y'_i = 1$ and, for $k \neq i$,

$$Y'_k = X_{M_k} B_k$$

Remembering that $X'_i = A_i$, we then check that

$$\mathbb{E}\left(\prod_{j\in J} Y'_j\right) = \frac{\mathbb{E}\big(\prod_{j\in\widetilde{J}} X_j\big)\mathbb{E}\big(A_i \prod_{j\in J} A_j\big)}{\mathbb{E}(A_i)} = \mathbb{E}\left(\prod_{j\in J} X'_j \,\middle|\, X'_i = 1\right).$$

Finally, it is clear that, with both constructions of $\mathbf{Y}^{(i)}$, $\mathbf{Y}'^{(i)}_k \geqslant \mathbf{X}'_k$. □

6.2. Proof of Theorem 6.1

Applying Theorem A to $|E_n| = \sum_{\{ij\}} \mathbb{1}_{\{i\leftrightarrow j\}}$ and using the expressions in Table 1, we get

$$d_{\mathrm{TV}}\Big(|E_n|, \mathrm{Poisson}(\lambda_n)\Big) \leqslant \min\Big\{1, \lambda_n^{-1}\Big\}\, C_n\,,$$

with $\lambda_n = \frac{n(n-1)}{2(r_n+1)}$ and

$$C_n = \frac{n(n-1)(n^2 r_n + 2n r_n^2 + n r_n - 2r_n^2 + 3r_n + 9)}{2\,(2\,r_n + 3)(r_n + 3)(r_n + 1)^2}\,.$$

When $r_n = \omega(n)$,

$$C_n = \Theta\left(\frac{n^4}{r_n^3} + \frac{n^3}{r_n^2}\right).$$

Now, if $r_n > \frac{n(n-1)}{2} - 1$, so that $\min\{1, \lambda_n^{-1}\} = 1$, we see that $C_n = \Theta(n^3/r_n^2)$. If by contrast $r_n \leqslant \frac{n(n-1)}{2} - 1$ then $\lambda_n^{-1} C_n = \Theta(n/r_n)$. In both cases, $\min\{1, \lambda_n^{-1}\}\, C_n$ goes to zero as $n \to +\infty$, proving the first part of Theorem 6.1.

The convergence of $\frac{|E_n| - \lambda_n}{\sqrt{\lambda_n}}$ to the standard normal distribution is a classic consequence of the conjunction of $d_{\mathrm{TV}}(|E_n|, \mathrm{Poisson}(\lambda_n)) \to 0$ with $\lambda_n \to +\infty$. See, e.g., [19], page 17, where this is recovered as a consequence of inequality (1.39).

Acknowledgements

François Bienvenu thanks Jean-Jil Duchamps for helpful discussions. All authors thank the *Center for Interdisciplinary Research in Biology* (CIRB) for funding. Florence Débarre thanks the Agence Nationale de la Recherche for funding (grant ANR-14-ACHN- 0003-01).

References

[1] F. Chung, L. Lu, T. G. Dewey, D. J. Galas, Duplication models for biological networks, J. Comput. Biol. 10 (5) (2003) 677–687.

[2] I. Ispolatov, P. Krapivsky, A. Yuryev, Duplication-divergence model of protein interaction network, Phys. Rev. E 71 (6) (2005) 061911.

[3] R. Solé, R. Pastor-Satorras, E. Smith, T. B. Kepler, A model of large-scale proteome evolution, Adv. Complex Syst. 5 (1) (2002) 43–54.

[4] A. Vázquez, A. Flammini, A. Maritan, A. Vespignani, Modeling of protein interaction networks, ComPlexUs 1 (2003) 38–44.

[5] E. B. Poulton, What is a species?, Proc. Entomol. Soc. Lond. 1903 (1904) lxxvii–cxvi.

[6] E. Mayr, Systematics and the Origin of Species from the Viewpoint of a Zoologist, Columbia University Press, 1942.

[7] J. A. Coyne, H. A. Orr, Speciation, Sinauer Associates, 2004.

[8] P. A. P. Moran, Random processes in genetics, Math. Proc. Camb. Philos. Soc. 54 (1) (1958) 60–71.

[9] J. F. C. Kingman, The coalescent, Stoch. Process. Appl. 13 (3) (1982) 235–248.

[10] R. Durrett, Probability models for DNA sequence evolution, 2nd Edition, Springer-Verlag New York, 2008.

[11] A. Etheridge, Some mathematical models from population genetics. École d'été de probabilités de Saint-Flour XXXIX-2009, Vol. 2012, Springer Science & Business Media, 2011.

[12] R. Durrett, Probability: theory and examples, 4th Edition, Cambridge University Press, 2010.

[13] B. Chauvin, A. Rouault, A. Wakolbinger, Growing conditioned trees, Stoch. Process. Appl. 39 (1) (1991) 117–130.

[14] R. Lyons, R. Pemantle, Y. Peres, Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes, Ann. Probab. 23 (3) (1995) 1125–1138.

[15] M. Manceau, A. Lambert, The species problem from the modeler's point of view, bioRxiv 075580 doi:10.1101/075580.

[16] S. Tavaré, Line-of-descent and genealogical processes, and their applications in population genetics models, Theor. Popul. Biol. 26 (2) (1984) 119–164.

[17] C. M. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, in: L. M. Le Cam, J. Neyman, E. L. Scott (Eds.), Proc. Sixth Berkeley Symp. Math. Stat. Probab. Vol. 2, University of California Press, 1972, pp. 583–602.

[18] L. H. Y. Chen, Poisson approximation for dependent trials, Ann. Probab. 3 (3) (1975) 534–545.

[19] A. D. Barbour, L. Holst, S. Janson, Poisson approximation, Oxford Studies in Probability, Clarendon Press, 1992.

[20] N. Ross, Fundamentals of Stein's method, Probab. Surv. 8 (2011) 201–293.

[21] A. Lambert, Probabilistic models for the (sub)tree(s) of life, Braz. J. Probab. Stat. 31 (3) (2017) 415–475.

A. Proofs of Propositions 2.4 and 2.6 and of Lemma 2.5

A.1. Proof of Propositions 2.4 and 2.6

Proposition 2.4. *Let $(K_t)_{t\geqslant 0}$ be a Kingman coalescent on $V=\{1,\dots,n\}$, and let $\pi_t(i)$ denote the block containing i in the corresponding partition at time t. Let the associated genealogy of pairs be the set*

$$\mathcal{G}=\left\{\Big(t,\,\{\pi_t(i)\,\pi_t(j)\}\Big):\{ij\}\in V^{(2)},\,t\in\Big[0,T_{\{ij\}}\Big[\right\},$$

where $T_{\{ij\}}=\inf\{t\geqslant 0:\pi_t(i)=\pi_t(j)\}$. Denote by

$$L_{\{ij\}}=\left\{\Big(t,\,\{\pi_t(i)\,\pi_t(j)\}\Big):t\in\Big[0,T_{\{ij\}}\Big[\right\}$$

the lineage of $\{ij\}$ in this genealogy. Finally, let $P^\bullet$ be a Poisson point process with constant intensity r_n on $\mathcal{G}$ and let $G=(V,E)$, where

$$E=\left\{\{ij\}\in V^{(2)}:P^\bullet\cap L_{\{ij\}}=\emptyset\right\}.$$

Then, $G\sim G_{n,r_n}$.

Proof. Let $(a_t)_{t\geqslant 0}$ and $\mathcal{P}^\star$ be as in Proposition 2.3, and let

$$\mathcal{G}^\star=\left\{\Big(t,\,\{a_t(i)\,a_t(j)\}\Big):\{ij\}\in V^{(2)},\,t\in\Big[0,T^\star_{\{ij\}}\Big[\right\},$$

where $T^\star_{\{ij\}}=\inf\{t\geqslant 0:a_t(i)=a_t(j)\}$. Being essentially a finite union of intervals, $\mathcal{G}^\star$ can be endowed with the Lebesgue measure.

As already suggested, conditional on $(a_t)_{t\geqslant 0}$, $\mathcal{P}^\star$ can be seen as a Poisson point process $P^\star$ with constant intensity r_n on $\mathcal{G}^\star$. More specifically,

$$P^\star=\left\{\Big(t,\,\{a_t(i)\,a_t(j)\}\Big):\{ij\}\in V^{(2)},\,t\in P^\star_{\{ij\}}\right\}.$$

With this formalism, writing

$$L^\star_{\{ij\}}=\left\{\Big(t,\,\{a_t(i)\,a_t(j)\}\Big):t\in\Big[0,T^\star_{\{ij\}}\Big[\right\}$$

for the lineage of $\{ij\}$ in this genealogy, we see that $P^\star_{\{ij\}}$ is isomorphic to $P^\star\cap L^\star_{\{ij\}}$. In particular,

$$P^\star_{\{ij\}}=\emptyset\iff P^\star\cap L^\star_{\{ij\}}=\emptyset$$

Now let $(\bar\pi_t)_{t\geqslant 0}$ be defined by

$$\forall i\in V,\quad \bar\pi_t(i)=\{j\in V:a_t(j)=a_t(i)\}.$$

Then, $\psi\colon(t,\,\{a_t(i)\,a_t(j)\})\mapsto(t,\,\{\bar\pi_t(i)\,\bar\pi_t(j)\})$ is a measure-preserving bijection from $\mathcal{G}^\star$ to $\psi(\mathcal{G}^\star)$. Therefore, $\psi(P^\star)$ is a Poisson point process with constant intensity r_n on $\psi(\mathcal{G}^\star)$. Since $(\bar\pi_t)_{t\geqslant 0}$ has the same law as $(\pi_t)_{t\geqslant 0}$ from the proposition, we conclude that

$$\Big(\psi(\mathcal{G}^\star),\,\psi(P^\star)\Big)\sim(\mathcal{G},P^\bullet)$$

which terminates the proof. □

Proposition 2.6. *For any $r > 0$, for any integer $n \geqslant 2$,*

$$\Phi_n(G_r^\star(n)) \sim G_{n,r}\,.$$

Proof. First, let us give a Poissonian construction of $(G_r^\dagger(t))_{t\geqslant 0}$. The edge-removal events can be recovered from a collection $\mathcal{P}^\dagger = \left(P^\dagger_{\{ij\}}\right)_{\{ij\}\in V^{(2)}}$ of i.i.d. Poisson point processes with rate r on $\mathbb{R}$ such that, if $t \in P_{\{ij\}}$ and there is an edge between i and j in $G_r^\dagger(t-)$, it is removed at time t. The duplication events induce a genealogy on the vertices of $G_r^\star(n)$ that is independent of $\mathcal{P}^\dagger$. Using a backward-time notation, let $a_t^\dagger(i)$ denote the ancestor of i at time $(t_n - t)$, i.e. t time-units before we reach $G_r^\star(n)$. Observe that, by construction of $G_r^\star(n)$,

$$\{ij\} \in G_r^\star(n) \iff \left\{t \geqslant 0 : t \in P^\dagger_{\{a_t^\dagger(i)\, a_t^\dagger(j)\}}\right\} = \emptyset\,.$$

Taking the relabeling of vertices into account, the genealogy on the vertices of $G_r^\star(n)$ translates into a genealogy on the vertices of $\Phi_n(G_r^\star(n))$, where the ancestor $\bar{a}_t$ function is given by $\bar{a}_t = \Phi_n \circ a_t^\dagger \circ \Phi_n^{-1}$. To keep only the relevant information about this genealogy, define

$$\bar{\pi}_t(i) = \{j \in V : \bar{a}_t(j) = \bar{a}_t(i)\}$$

and let

$$\bar{\mathcal{G}} = \left\{\left(t,\, \{\bar{\pi}_t(i)\,\bar{\pi}_t(j)\}\right) : \{ij\} \in V^{(2)},\, t \in \left[0, T_{\{ij\}}\right[\right\},$$

where $T_{\{ij\}} = \inf\{t \geqslant 0 : \bar{\pi}_t(i) = \bar{\pi}_t(j)\}$. As before, let us denote by

$$\bar{L}_{\{ij\}} = \left\{\left(t,\, \{\bar{\pi}_t(i)\,\bar{\pi}_t(j)\}\right) : t \in \left[0, T_{\{ij\}}\right[\right\}$$

the lineage of $\{ij\}$ in this genealogy. Finally, define

$$\bar{P} = \left\{\left(t,\, \{\bar{\pi}_t(i)\,\bar{\pi}_t(j)\}\right) : \{ij\} \in V^{(2)},\, t \in P^\dagger_{\{a_t^\dagger(\Phi_n^{-1}(i))\, a_t^\dagger(\Phi_n^{-1}(j))\}}\right\}.$$

Then, conditional on $\bar{\mathcal{G}}$, $\bar{P}$ is a Poisson point process with constant intensity r_n on $\bar{\mathcal{G}}$. Moreover,

$$\begin{aligned}\{ij\} \in \Phi_n(G_r^\star(n)) &\iff \{\Phi_n^{-1}(i)\,\Phi_n^{-1}(j)\} \in G_r^\star(n)\\ &\iff \left\{t \geqslant 0 : t \in P^\dagger_{\{a_t^\dagger(\Phi_n^{-1}(i))\, a_t^\dagger(\Phi_n^{-1}(j))\}}\right\} = \emptyset\\ &\iff \bar{P} \cap \bar{L}_{\{ij\}} = \emptyset\,.\end{aligned}$$

Therefore, by Proposition 2.4, to conclude the proof it is sufficient to show that $(\bar{\pi}_t)_{t\geqslant 0}$ has the same law as the corresponding process for a Kingman coalescent. By construction, the time to go from k to $k-1$ blocks in $(\bar{\pi}_t)_{t\geqslant 0}$ is an exponential variable with parameter $\binom{k}{2}$ and thus it only remains to

prove that the tree encoded by $(\bar{\pi}_t)_{t\geqslant 0}$ has the same topology as the Kingman coalescent tree. This follows directly from the standard fact that the shape of a Yule tree with n tips labeled uniformly at random with the integers from 1 to n is the same as that of the shape of a Kingman n-coalescent tree – namely, the uniform law on the set of ranked tree shapes with n tips labeled by $\{1,\dots,n\}$ (see e.g. [21]).

Alternatively, we can finish the proof as follows: working in backward time, for $i=1,\dots,n-1$, consider the i-th coalescence and let U_i denote the mother in the corresponding duplication in the construction of $G_r^\star(n)$. Note that $U_i \sim \text{Uniform}(\{1,\dots,n-i\})$, and that the coalescing blocks are then the block that contains $\Phi_n(U_i)$ and the block that contains $\Phi_n(n-i+1)$. Let us record the information about the i first coalescences in the variable Λ_i defined by $\Lambda_0=\emptyset$ and, for $i\geqslant 1$,

$$\Lambda_i = \Big(\Phi_n(n-k+1),\, \Phi_n(U_k)\Big)_{k=1,\dots,i}.$$

Thus, we have to show that, conditional on Λ_{i-1}, the block containing $\Phi_n(n-i+1)$ and the block containing $\Phi_n(U_i)$ are uniformly chosen. We proceed by induction. For $i=1$, this is trivial. Now, for $i>1$, observe that, conditional on Λ_{i-1}, the restriction of Φ_n to

$$I_i = \{1,\dots,n\}\setminus\{\Phi_n(n),\dots,\Phi_n(n-i)\}$$

is a uniform permutation on I_i. As a result, $\{\Phi_n(n-i+1),\,\Phi_n(U_i)\}$ is a uniformly chosen pair of elements of I_i (note that the fact that U_i is uniformly distributed on $\{1,\dots,n-i\}$ is not necessary for this, but is needed to ensure that the restriction of Φ_n to I_{i+1} remains uniform when conditioning on Λ_i in the next step of the induction). Since each block contains exactly one element of I_i, this terminates the proof. □

A.2. Proof of Lemma 2.5

Lemma 2.5. *Let S be a subset of $V^{(2)}$. Conditional on the measure $\mathcal{M}$, for any interval $I\subset[0,+\infty[$ such that*

(i) *For all $\{ij\}\in S$, $\forall t\in I$, $a_t(i)\neq a_t(j)$.*

(ii) *For all $\{k\ell\}\neq\{ij\}$ in S, $\forall t\in I$, $\{a_t(i),a_t(j)\}\neq\{a_t(k),a_t(\ell)\}$,*

$P^\star_{\{ij\}}\cap I$, $\{ij\}\in S$, are independent Poisson point processes with rate r_n on I.

Moreover, for any disjoint intervals I and J, $(P^\star_{\{ij\}}\cap I)_{\{ij\}\in S}$ is independent of $(P^\star_{\{ij\}}\cap J)_{\{ij\}\in S}$.

Proof. For all $t\geqslant 0$, define S_t by

$$S_t = \{\{a_t(i),a_t(j)\} : \{ij\}\in S\}.$$

Set $t_0=\inf I$ and let $t_1,\dots,t_{m-1}$ be the jump times of $(S_t)_{t\geqslant 0}$ on I, i.e.

$$t_p = \inf\Big\{t>t_{p-1} : S_t\neq S_{t_{p-1}}\Big\},\quad p=1,\dots,m-1.$$

Finally, set $t_m = \sup I$ and, for $p = 0, \dots, m-1$, let $I_p = [t_p, t_{p+1}[$ and $\tilde{a}_p = a_{t_p}$, so that $(\tilde{a}_p)_{p\in\{0,\dots,m\}}$ is the embedded chain of $(a_t)_{t\in I}$. With this notation, for all $\{ij\} \in S$,

$$P^\star_{\{ij\}} \cap I = \bigcup_{p=0}^{m-1} \left(P_{\{\tilde{a}_p(i),\, \tilde{a}_p(j)\}} \cap I_p \right),$$

where for $p \neq q$, $I_p \cap I_q = \emptyset$, and $P_{\{uv\}}$, $\{uv\} \in V^{(2)}$, are i.i.d. Poisson point processes on $[0, +\infty[$ with rate r_n. By assumption, for all $p = 0, \dots, m-1$, for all $\{ij\} \neq \{k\ell\}$ in S, $\tilde{a}_p(i) \neq \tilde{a}_p(j)$, $\tilde{a}_p(k) \neq \tilde{a}_p(\ell)$ and $\{\tilde{a}_p(i), \tilde{a}_p(j)\} \neq \{\tilde{a}_p(k), \tilde{a}_p(\ell)\}$. This shows that $(P_{\{\tilde{a}_p(i),\tilde{a}_p(j)\}} \cap I_p)$, $\{ij\} \in S$ and $p = 0, \dots, m-1$, are i.i.d. Poisson point processes with rate r_n on the corresponding intervals, proving the first part of the lemma.

The second assertion is proved similarly. Adapting the previous notation to work with two disjoint intervals I and J, i.e. letting $(\tilde{a}^I_p)_{p\in\{0,\dots,m_I\}}$ be the embedded chain of $(a_t)_{t\in I}$ and $(\tilde{a}^J_p)_{p\in\{0,\dots,m_J\}}$ that of $(a_t)_{t\in J}$, for all $\{ij\} \in S$ we write

$$P^\star_{\{ij\}} \cap I = \bigcup_{p=0}^{m_I-1} \left(P_{\{\tilde{a}^I_p(i),\, \tilde{a}^I_p(j)\}} \cap I_p \right),$$

and

$$P^\star_{\{ij\}} \cap J = \bigcup_{p=0}^{m_J-1} \left(P_{\{\tilde{a}^J_p(i),\, \tilde{a}^J_p(j)\}} \cap J_p \right).$$

We conclude the proof by noting that the families

$$\left(P_{\{\tilde{a}^I_p(i),\, \tilde{a}^I_p(j)\}} \cap I_p \right)_{\{ij\}\in S,\, p\in\{0,\dots,m_I\}}$$

and

$$\left(P_{\{\tilde{a}^J_p(i),\, \tilde{a}^J_p(j)\}} \cap J_p \right)_{\{ij\}\in S,\, p\in\{0,\dots,m_J\}}$$

are independent, because the elements of these families are either deterministic (if, e.g, $\tilde{a}^I_p(i) = \tilde{a}^I_p(j)$, in which case $P_{\{\tilde{a}^I_p(i),\, \tilde{a}^I_p(j)\}} = \emptyset$) or Poisson point processes on intervals that are disjoint from each of the intervals involved in the definition of the other family. □

B. Proofs of Proposition 3.5 and Corollary 3.6

Proposition 3.5. *Let i, j, k and ℓ be four distinct vertices of G_{n,r_n}. We have*

$$\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow j\}}, \mathbb{1}_{\{k\leftrightarrow \ell\}}\Big) = \frac{2\,r_n}{(1+r_n)^2(3+r_n)(3+2\,r_n)}$$

Corollary 3.6. *Let $D_n^{(i)}$ and $D_n^{(j)}$ be the respective degrees of two fixed vertices i and j, and let $|E_n|$ be the number of edges of G_{n,r_n}. We have*

$$\mathrm{Cov}\Big(D_n^{(i)}, D_n^{(j)}\Big) = \frac{r_n}{(1+r_n)^2}\left(1 + \frac{3(n-2)}{3+2\,r_n} + \frac{2(n-2)(n-3)}{(3+r_n)(3+2\,r_n)}\right)$$

and

$$\mathrm{Var}(|E_n|) = \frac{r_n\, n\,(n-1)(n^2+2\,r_n^2+2\,n\,r_n+n+5\,r_n+3)}{2\,(1+r_n)^2\,(3+r_n)\,(3+2\,r_n)}$$

B.1. Proof of Proposition 3.5

The proof of Proposition 3.5 parallels that of Proposition 3.3, but this time the topology of the genealogy of the pairs of vertices has to be taken into account. Indeed, define

$$S_t = \{a_t(i), a_t(j), a_t(k), a_t(\ell)\}$$

and let $\tau_1 < \tau_2 < \tau_3$ be the times of coalescence in the genealogy of $\{i,j,k,\ell\}$, i.e.

$$\tau_p = \inf\{t \geqslant 0 : |S_t| = 4-p\}, \quad p = 1,2,3\,.$$

Write $I_1 = [0,\tau_1[$, $I_2 = [\tau_1,\tau_2[$ and $I_3 = [\tau_2,\tau_3[$. Finally, for $m = 1,2$, let

$$A^{(m)}_{\{uv\}} = \{a_{\tau_m-}(u) \neq a_{\tau_m-}(v)\} \cap \{a_{\tau_m}(u) = a_{\tau_m}(v)\}$$

be the event that the m-th coalescence in the genealogy of $\{i,j,k,\ell\}$ involved the lineages of u and v (note that the third coalescence is uniquely determined by the first and the second, so we do not need $A^{(3)}_{\{uv\}}$).

On $A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{k\ell\}}$, $\{i \leftrightarrow j, k \leftrightarrow \ell\}$ is equivalent to

$$(P^\star_{\{ij\}} \cap I_1) \cup (P^\star_{\{k\ell\}} \cap I_1) \cup (P^\star_{\{k\ell\}} \cap I_2) = \emptyset$$

so that, conditionally on I_1 and I_2, by Lemma 2.5,

$$\begin{aligned}\mathbb{P}\Big(i \leftrightarrow j, k \leftrightarrow \ell \;\Big|\; A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{k\ell\}}\Big) &= \mathbb{P}\Big((P^\star_{\{ij\}} \cup P^\star_{\{k\ell\}}) \cap I_1 = \emptyset\Big) \times \mathbb{P}\Big(P^\star_{\{k\ell\}} \cap I_2 = \emptyset\Big) \\ &= \frac{6}{6+2\,r_n} \times \frac{3}{3+r_n}\,.\end{aligned}$$

By contrast, on $A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{ik\}}$, $\{i \leftrightarrow j, k \leftrightarrow \ell\}$ is

$$(P^\star_{\{ij\}} \cap I_1) \cup (P^\star_{\{k\ell\}} \cap I_1) \cup (P^\star_{\{k\ell\}} \cap I_2) \cup (P^\star_{\{k\ell\}} \cap I_3) = \emptyset$$

and thus

$$\mathbb{P}\Big(i \leftrightarrow j, k \leftrightarrow \ell \,\Big|\, A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{ik\}}\Big) = \frac{6}{6+2\,r_n} \times \frac{3}{3+r_n} \times \frac{1}{1+r_n}\,.$$

Given a realization of the topology of the genealogy of the form $A^{(1)}_{\{u_1v_1\}} \cap A^{(2)}_{\{u_2v_2\}}$, we can always express $\{i \leftrightarrow j, k \leftrightarrow \ell\}$ as a union of intersections of $P^\star_{\{ij\}}$ and $P^\star_{\{k\ell\}}$ with I_1, I_2 and I_3. In total, there are $\binom{4}{2} \times \binom{3}{2} = 18$ possible events $A^{(1)}_{\{u_1v_1\}} \cap A^{(2)}_{\{u_2v_2\}}$, each having probability $1/18$. This enables us to compute $\mathbb{P}(i \leftrightarrow j, k \leftrightarrow \ell)$, but in fact the calculations can be simplified by exploiting symmetries, such as the fact that $\{ij\}$ and $\{k\ell\}$ are interchangeable. In the end, it suffices to consider four cases, as depicted in Figure B.6.

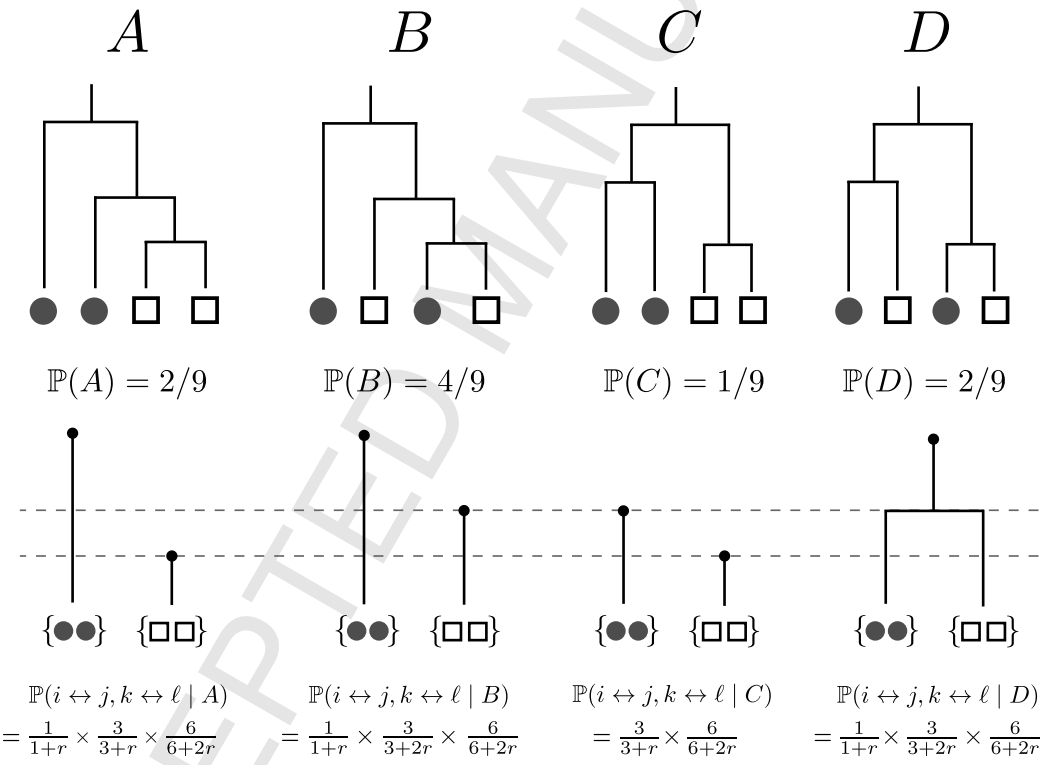

Figure B.6: The four cases that we consider to compute $\mathbb{P}(i \leftrightarrow j, k \leftrightarrow \ell)$. Top, the "aggregated" genealogies of vertices and their probability. Each of these correspond to several genealogies on $\{i, j, k, \ell\}$, which are obtained by labeling symbols in such a way that a pair of matching symbols has to correspond to either $\{ij\}$ or $\{k\ell\}$. For instance, $C = (A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{k\ell\}}) \cup (A^{(1)}_{\{k\ell\}} \cap A^{(2)}_{\{ij\}})$ and therefore $\mathbb{P}(C) = 2/18$. Similarly, $A = (A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{ik\}}) \cup (A^{(1)}_{\{ij\}} \cap A^{(2)}_{\{i\ell\}}) \cup (A^{(1)}_{\{k\ell\}} \cap A^{(2)}_{\{ik\}}) \cup (A^{(1)}_{\{k\ell\}} \cap A^{(2)}_{\{jk\}})$ and $\mathbb{P}(A) = 4/18$, etc. Bottom, the associated genealogy of the pairs and the corresponding conditional probability of $\{i \leftrightarrow j, k \leftrightarrow \ell\} \Leftrightarrow \{\square \leftrightarrow \square, \bullet \leftrightarrow \bullet\}$.

Putting the pieces together, we find that

$$\begin{aligned}\mathbb{P}(i \leftrightarrow k, j \leftrightarrow \ell) = &\frac{6}{9} \times \frac{1}{1+r_n} \times \frac{3}{3+2r} \times \frac{6}{6+2\,r_n} \\ &+ \frac{2}{9} \times \frac{1}{1+r_n} \times \frac{3}{3+r_n} \times \frac{6}{6+2\,r_n} \\ &+ \frac{1}{9} \times \frac{3}{3+r_n} \times \frac{6}{6+2\,r_n} \\ &= \frac{9+2\,r_n}{(1+r_n)(3+r_n)(3+2\,r_n)}\,.\end{aligned}$$

and Proposition 3.5 follows, since

$$\mathbb{P}(i \leftrightarrow j)\,\mathbb{P}(k \leftrightarrow \ell) = \left(\frac{1}{1+r_n}\right)^2.$$

B.2. Proof of Corollary 3.6

Corollary 3.6 is proved by standard calculations. First,

$$\begin{aligned}\mathrm{Cov}\Big(D_n^{(i)}, D_n^{(j)}\Big) &= \mathrm{Cov}\left(\sum_{k\neq i} \mathbb{1}_{\{i\leftrightarrow k\}}, \sum_{\ell\neq j} \mathbb{1}_{\{j\leftrightarrow \ell\}}\right) \\ &= \mathrm{Var}\Big(\mathbb{1}_{\{i\leftrightarrow j\}}\Big) \\ &\quad + 3(n-2)\,\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow k\}}, \mathbb{1}_{\{j\leftrightarrow k\}}\Big) \\ &\quad + (n-2)(n-3)\,\mathrm{Cov}\Big(\mathbb{1}_{\{i\leftrightarrow k\}}, \mathbb{1}_{\{j\leftrightarrow \ell\}}\Big)\end{aligned}$$

Remembering from Proposition 3.1 that $\mathrm{Var}(\mathbb{1}_{\{i\leftrightarrow j\}}) = r_n/(1+r_n)^2$ and from Proposition 3.3 that $\mathrm{Cov}(\mathbb{1}_{\{i\leftrightarrow k\}}, \mathbb{1}_{\{j\leftrightarrow k\}}) = \frac{r_n}{(1+r_n)^2(3+2\,r_n}$, and using Proposition 3.5, we find that

$$\mathrm{Cov}\Big(D_n^{(i)}, D_n^{(j)}\Big) = \frac{r_n}{(1+r_n)^2}\left(1 + \frac{3(n-2)}{3+2\,r_n} + \frac{2(n-2)(n-3)}{(3+r_n)(3+2\,r_n)}\right).$$

Finally, to compute $\mathrm{Var}(|E_n|)$, we could do a similar calculation. However, it is easier to note that

$$|E_n| = \frac{1}{2}\sum_{i=1}^{n} D_n^{(i)}\,.$$

As a result,

$$\begin{aligned}\mathrm{Var}(|E_n|) &= \frac{1}{4}\Big(n\,\mathrm{Var}\Big(D_n^{(i)}\Big) + n(n-1)\,\mathrm{Cov}\Big(D_n^{(i)}, D_n^{(j)}\Big)\Big) \\ &= \frac{r_n\,n\,(n-1)(n^2+2\,r_n^2+2\,n\,r_n+n+5\,r_n+3)}{2(1+r_n)^2(3+r_n)(3+2\,r_n)}\,.\end{aligned}$$

C. Proof of Theorem 4.2

In this section, we prove Theorem 4.2.

Theorem 4.2 (convergence of the rescaled degree)**.**

(i) *If* $r_n \to r > 0$*, then* $\frac{D_n}{n}$ *converges in distribution to a* $\mathrm{Beta}(2, 2\,r)$ *random variable.*

(ii) *If* r_n *is both* $\omega(1)$ *and* $o(n)$*, then* $\frac{D_n}{n/r_n}$ *converges in distribution to a size-biased exponential variable with parameter* 2*.*

(iii) *If* $2\,r_n/n \to \rho > 0$*, then* $D_n + 1$ *converges in distribution to a size-biased geometric variable with parameter* $\rho/(1+\rho)$*.*

The proof of (iii) is immediate: indeed, by Theorem 4.1,

$$\mathbb{P}(D_n + 1 = k) = \frac{2\,r_n\,(2\,r_n + 1)}{(n + 2\,r_n)(n - 1 + 2\,r_n)}\,k \prod_{i=1}^{k-1} \frac{n-i}{n-i+2\,r_n-1}\,.$$

If $2\,r_n/n \to \rho$, then for any fixed k this goes to $k\left(\frac{\rho}{1+\rho}\right)^2\left(\frac{1}{1+\rho}\right)^{k-1}$ as $n \to +\infty$.

C.1. Outline of the proof

To prove (i) and (ii), we show the pointwise convergence of the cumulative distribution function F_n of the rescaled degree. To do so, in both cases,

1. We show that, for any $\varepsilon > 0$, for n large enough,
$$\forall y \geqslant 0, \quad \int_0^y f_n(x)\,dx \leqslant F_n(y) \leqslant \int_0^{y+\varepsilon} f_n(x)\,dx$$
for some function f_n to be introduced later.
2. We identify the limit of f_n as a classical probability density f, and use dominated convergence to conclude that
$$\forall y \geqslant 0, \quad \int_0^y f_n(x)\,dx \to \int_0^y f(x)\,dx\,.$$

In order to factorize as much of the reasoning as possible, we introduce the rescaling factor N_n:

- When $r_n \to r$, i.e. when we want to prove (i), $N_n = n$.
- When r_n is both $\omega(1)$ and $o(n)$, i.e. when we want to prove (ii), $N_n = n/r_n$.

Thus, in both cases the rescaled degree is D_n/N_n and its cumulative distribution function is

$$F_n(y) = \sum_{k=0}^{\lfloor N_n y \rfloor} \mathbb{P}(D_n = k)\,.$$

C.2. Step 1

For all $x > 0$, let

$$f_n(x) = N_n \mathbb{P}(D_n = \lfloor N_n x \rfloor),$$

so that

$$\forall k \in \mathbb{N}, \quad \mathbb{P}(D_n = k) = \int_{k/N_n}^{(k+1)/N_n} f_n(x)\, dx\,.$$

If follows that

$$F_n(y) = \int_0^{(\lfloor N_n y \rfloor + 1)/N_n} f_n(x)\, dx\,.$$

Finally, since $y \leqslant \frac{\lfloor N_n y \rfloor + 1}{N_n} \leqslant y + \frac{1}{N_n}$ and f_n is non-negative, for any $\varepsilon > 0$, for n large enough,

$$\forall y \geqslant 0, \quad \int_0^y f_n(x)\, dx \leqslant F_n(y) \leqslant \int_0^{y+\varepsilon} f_n(x)\, dx\,,$$

and the rank after which these inequalities hold is uniform in y, because the convergence of $(\lfloor N_n y \rfloor + 1)/N_n$ to y is.

C.3. Step 2

To identify the limit of f_n, we reexpress it in terms of the gamma function. Using that $\Gamma(z) = z\Gamma(z)$, by induction,

$$\prod_{i=1}^{k}(n-i) = \frac{\Gamma(n)}{\Gamma(n-k)} \quad \text{and} \quad \prod_{i=1}^{k}(n-i+2\,r_n-1) = \frac{\Gamma(n+2\,r_n-1)}{\Gamma(n-k+2\,r_n-1)}\,.$$

Therefore, $f_n(x)$ can also be written

$$f_n(x) = \frac{N_n\, 2\,r_n\,(2\,r_n+1)}{(n+2\,r_n)(n-1+2\,r_n)} \Big(\lfloor N_n x \rfloor + 1\Big) \times P_n(x)\,, \tag{C.1}$$

where

$$P_n(x) = \frac{\Gamma(n)\,\Gamma(n-\lfloor N_n x \rfloor + 2\,r_n - 1)}{\Gamma(n-\lfloor N_n x \rfloor)\,\Gamma(n+2\,r_n-1)}\,. \tag{C.2}$$

We now turn to the specificities of the proofs of (i) and (ii).

C.3.1. Proof of (i)

In this subsection, $r_n \to r > 0$ and $N_n = n$.

Limit of f_n. Recall that

$$\forall \alpha \in \mathbb{R}, \quad \frac{\Gamma(n+\alpha)}{\Gamma(n)} \sim n^\alpha .$$

Using this in (C.2), we see that, for all $x \in [0,1[$,

$$P_n(x) \to (1-x)^{2r-1} .$$

Therefore, for all $x \in [0,1[$,

$$f_n(x) \to 2r(2r+1)\, x\, (1-x)^{2r-1} .$$

Noting that $2r(2r+1) = 1/B(2,2r)$, where B denotes the beta function, we can write $f = \lim_n f_n$ as

$$f\colon x \mapsto \frac{x(1-x)^{2r-1}}{B(2,2r)}\, \mathbb{1}_{[0,1[}(x)$$

and we recognize the probability density function of the Beta$(2,2r)$ distribution.

Domination of (f_n). First note that, for all $x \in [0,1[$,

$$\frac{1}{n-1+2\,r_n} \prod_{i=1}^{\lfloor nx \rfloor} \frac{n-i}{n-i+2\,r_n-1} \quad = \quad \frac{1}{n-\lfloor nx \rfloor + 2\,r_n - 1} \prod_{i=1}^{\lfloor nx \rfloor} \frac{n-i}{n-i+2\,r_n},$$

where the empty product is understood to be 1. Since $2\,r_n > 0$, this enables us to write that, for all $x \in [0,1[$,

$$f_n(x) = \underbrace{\frac{n\,2r\,(2r+1)}{n+2r}}_{\leqslant (2r+1)^2} \times \frac{\lfloor nx \rfloor + 1}{n-1+2r} \times \underbrace{\frac{1}{n-\lfloor nx \rfloor + 2r - 1}}_{\leqslant \frac{1}{2r}} \times \underbrace{\prod_{i=1}^{\lfloor nx \rfloor} \frac{n-i}{n-i+2r}}_{\leqslant 1} .$$

where, to avoid cluttering the expression, the n index of r_n has been dropped. Since

$$\frac{\lfloor nx \rfloor + 1}{n-1+2\,r_n} \leqslant \frac{(n-1)x + x + 1}{n-1} \leqslant x + \frac{2}{n-1} \xrightarrow[n\to+\infty]{\text{uniformly}} x ,$$

there exists c such that, for all $x \in [0,1[$ and n large enough,

$$f_n(x) \leqslant c\, x$$

Since f_n is zero outside of $[0,1[$, this shows that (f_n) is dominated by $g\colon x \mapsto c\, x\, \mathbb{1}_{[0,1[}(x)$.

C.3.2. Proof of (ii)

In this subsection, r_n is both $\omega(1)$ and $o(n)$, and $N_n = n/r_n$. For brevity, we will write k_n for $\lfloor nx/r_n \rfloor$. It should be noted that

- k_n is both $\omega(1)$ and $o(n)$.
- $k_n r_n / n \to x$ uniformly in x on $[0, +\infty[$.

Limit of f_n. In this paragraph, we will need Stirling's formula for the asymptotics of Γ:

$$\Gamma(t+1) \sim \sqrt{2\pi t}\,\frac{t^t}{e^t}\,.$$

Using this in Equation (C.2),

$$\begin{aligned} P_n(x) &= \frac{\Gamma(n)\,\Gamma(n - \lfloor N_n x \rfloor + 2\,r_n - 1)}{\Gamma(n - \lfloor N_n x \rfloor)\,\Gamma(n + 2\,r_n - 1)} \\ &\sim \underbrace{\sqrt{\frac{(n-1)(n-2-k_n+2\,r_n)}{(n-1-k_n)(n-2+2\,r_n)}}}_{\sim 1} \times \underbrace{\frac{e^{n-1-k_n}\,e^{n-2+2r_n}}{e^{n-1}\,e^{n-2-k_n+2r_n}}}_{=1} \times Q_n \end{aligned}$$

where

$$Q_n = \frac{(n-1)^{n-1}\,(n-2-k_n+2\,r_n)^{n-2-k_n+2r_n}}{(n-1-k_n)^{n-1-k_n}\,(n-2+2\,r_n)^{n-2+2r_n}}\,.$$

Let us show that $Q_n \to e^{-2x}$:

$$\begin{aligned} \log Q_n = \;& (n-1)\log(n-1) \\ &+ (n-a+b)\log(n-a+b) \\ &- (n-a)\log(n-a) \\ &- (n-1+b)\log(n-1+b) \end{aligned}$$

where, to avoid cluttering the text, we have written a for $k_n + 1$ and b for $2\,r_n - 1$. Factorizing, we get

$$\log Q_n = n\log\left(\frac{(n-1)(n-a+b)}{(n-a)(n-1+b)}\right) - a\log\left(\frac{n-a+b}{n-a}\right) + b\log\left(\frac{n-a+b}{n-1+b}\right) - \log\left(\frac{n-1}{n-1+b}\right)\,.$$

Now,

$$\frac{(n-1)(n-a+b)}{(n-a)(n-1+b)} = 1 + \underbrace{\frac{(a-1)b}{n^2 - n + nb - na + a - ab}}_{\sim \frac{2k_n r_n}{n^2} = o(1)}$$

so that

$$n\log\left(\frac{(n-1)(n-a+b)}{(n-a)(n-1+b)}\right) \sim \frac{2k_n r_n}{n} \to 2x$$

Similarly,

$$-a\log\left(\frac{n-a+b}{n-a}\right) = -a\log\left(1+\frac{b}{n-a}\right) \sim -\frac{ab}{n} \to -2x$$

$$b\log\left(\frac{n-a+b}{n-1+b}\right) = b\log\left(1+\frac{1-a}{n-1+b}\right) \sim -\frac{ab}{n} \to -2x$$

and, finally, $-\log\left(\frac{n-1}{n-1+b}\right) \to 0$. Putting the pieces together,

$$\log Q_n \to -2x\,.$$

Having done that, we note that

$$\frac{2\,n(2\,r_n+1)}{(n+2\,r_n)(n-1+2\,r_n)}(k_n+1) \;\to\; 4\,x\,.$$

Plugging these results in Equation (C.1), we see that

$$\forall x \in \mathbb{R}, \quad f_n(x) \;\to\; 4\,x\,e^{-2x}\,\mathbb{1}_{[0,+\infty[}(x)$$

and we recognize the probability density function of a size-biased exponential distribution with parameter 2.

Domination of (f_n). Recall that, since $N_n = n/r_n$, for all $x \in [0,1[$,

$$f_n(x) = \frac{2\,n\,(2\,r_n+1)}{(n+2\,r_n)(n-1+2\,r_n)}\,(k_n+1)\,\prod_{i=1}^{k_n}\frac{n-i}{n-i+2\,r_n-1}\,.$$

Next, note that, for all i,

$$\frac{n-i}{n-i+2\,r_n-1} = 1-\frac{2\,r_n-1}{n-i+2\,r_n-1} \leqslant \exp\left(-\frac{2\,r_n-1}{n-i+2\,r_n-1}\right)$$

so that

$$\prod_{i=1}^{k_n}\frac{n-i}{n-i+2\,r_n-1} \leqslant \exp\left(-\sum_{i=1}^{k_n}\frac{2\,r_n-1}{n-i+2\,r_n-1}\right)\,,$$

with

$$\sum_{i=1}^{k_n}\frac{2\,r_n-1}{n-i+2\,r_n-1} \geqslant k_n\,\frac{2\,r_n-1}{n-1+2\,r_n-1}\,.$$

Because $r_n = \omega(1)$, for all $\varepsilon > 0$, $2\,r_n - 1 \geqslant (1-\varepsilon)2\,r_n$ for n large enough. Similarly, since $r_n = o(n)$, $\frac{1}{n+2\,r_n} \geqslant \frac{1}{(1+\varepsilon)n}$. As a result, there exists $c > 0$ such that

$$k_n\,\frac{2\,r_n-1}{n-1+2\,r_n-1} \geqslant c\,k_n\,\frac{2\,r_n}{n} \xrightarrow{\text{uniformly}} 2cx\,.$$

We conclude that

$$\forall x \geqslant 0, \quad \prod_{i=1}^{k_n} \frac{n-i}{n-i+2\,r_n-1} \leqslant \exp(-2cx)$$

for n large enough. Finally,

$$2 \times \underbrace{\frac{n}{n+2\,r_n}}_{\leqslant 1} \times \underbrace{\frac{(2\,r_n+1)(k_n+1)}{(n-1+2\,r_n)}}_{\to 2x, \text{ uniformly}} \leqslant 4cx$$

and so (f_n) is dominated by $g\colon x \mapsto 4c\,x\,e^{-2cx}\,\mathbb{1}_{[0,+\infty[}(x)$.