Random graphs in evolution

François Bienvenu

École Doctorale de Sciences Mathématiques de Paris Centre

Laboratoire de Probabilités, Statistiques et Modélisation
Sorbonne Université

Thèse de doctorat

Discipline : mathématiques

présentée par François Bienvenu

Random Graphs in Evolution

Sous la direction d'Amaury Lambert

Après avis des rapporteurs :

M. Simon Harris (University of Auckland)
Mme Régine Marchand (Université de Lorraine)

Soutenue le 13 septembre 2019 devant le jury composé de :

M. Nicolas Broutin	Sorbonne Université	Examinateur
M. Amaury Lambert	Sorbonne Université	Directeur de thèse
Mme Régine Marchand	Université de Lorraine	Rapporteuse
Mme. Céline Scornavacca	Chargée de recherche CNRS	Examinatrice
M. Viet Chi Tran	Université de Lille	Examinateur

Contents

Contents **3**

1 Introduction **5**
1.1 A very brief history of random graphs 6
1.2 A need for tractable random graphs in evolution 7
1.3 Outline of the thesis 8
Literature cited in the introduction 9

2 The split-and-drift random graph **11**
2.1 Introduction 13
2.2 Coalescent constructions of G_{n,r_n} 17
2.3 First and second moment methods 22
2.4 The degree distribution 28
2.5 Connected components in the intermediate regime 33
2.6 Number of edges in the sparse regime 36
Chapter references 40
2.A Proofs of Propositions 2.2.4 and 2.2.6 and of Lemma 2.2.5 42
2.B Proofs of Proposition 2.3.5 and Corollary 2.3.6 45
2.C Proof of Theorem 2.4.2 48

3 The Moran forest **53**
3.1 Introduction 55
3.2 Sampling of the stationary distribution 57
3.3 Number of trees 60
3.4 Degrees 63
3.5 Tree sizes 70
Chapter references 78
3.A Proof of point (ii) of Proposition 3.4.4 80
3.B Technical lemmas used in the proof of Theorem 3.5.8 82

4 Ranked tree-child networks **86**
4.1 Introduction 88
4.2 Counting and generating RTCNs 94
4.3 RTCNs and ranked trees 100
4.4 Cherries and reticulated cherries 104
4.5 Random paths between the roots and the leaves 109
4.6 Number of lineages in the ancestry of a leaf 114
Chapter references 121

4.A Permutations and subexcedant sequences . . . 123
4.B Lemmas used in Section 4.4 . . . 125
4.C Variance of χ_ℓ . . . 126

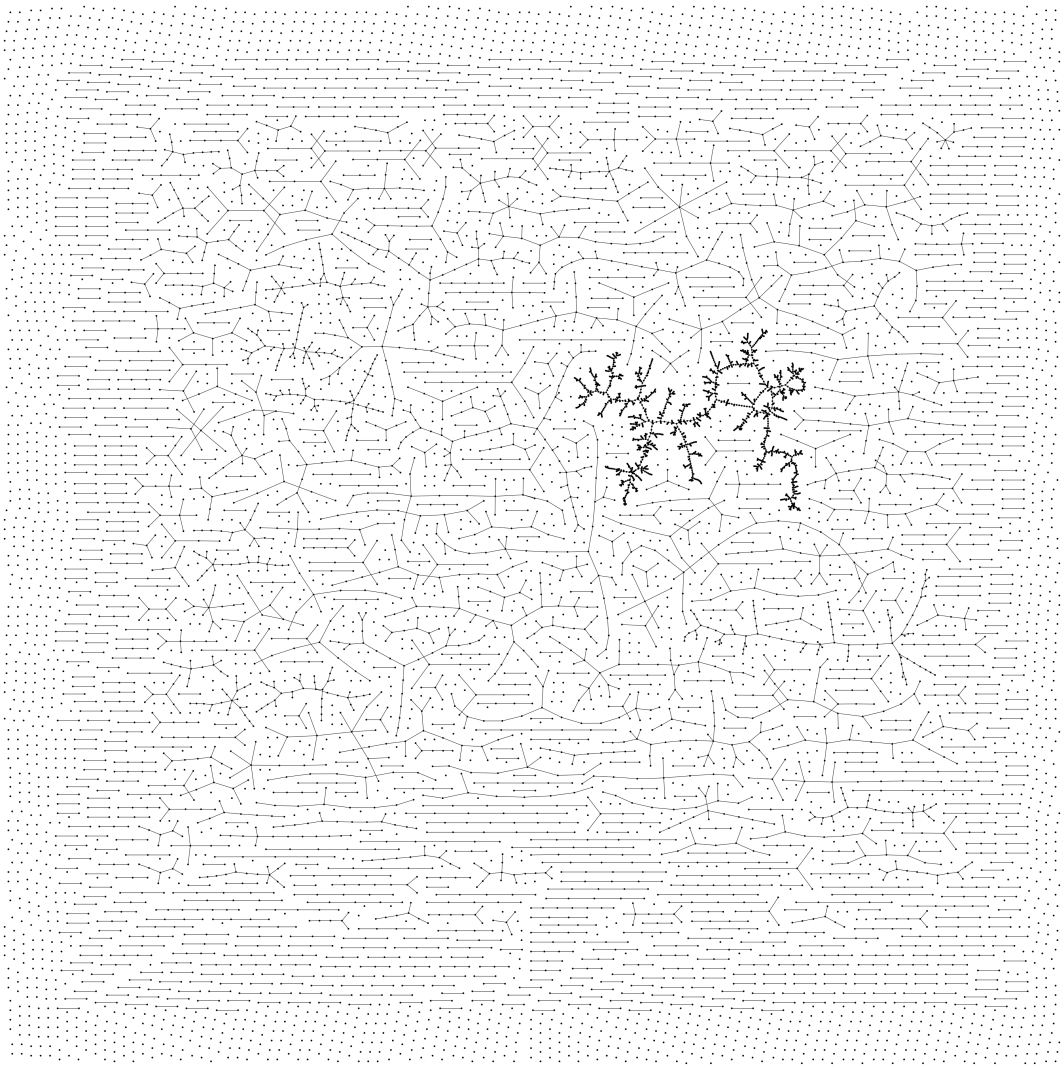
5 Oriented percolation in randomly oriented graphs 127
5.1 Introduction . . . 129
5.2 Positive association of the percolation cluster . . . 130
5.3 Percolation from the leaves of a binary tree . . . 133
Chapter references . . . 144

6 The equivocal "mean age of parents at birth" 145
6.1 Introduction . . . 147
6.2 Derivation and interpretation of the expressions of μ_1 and τ . . . 148
6.3 Examples . . . 153
6.4 Discussion . . . 157
Chapter references . . . 158
6.A Basic facts about Poisson point processes . . . 160
6.B Expression of τ for discrete age structures . . . 161
6.C Proof of $\mathbb{E}(\tilde{T}) \to \mathbb{E}(T^2)/\mathbb{E}(T)$ as $m \to 0$ . . . 163
6.D Proof of $\tau \leqslant \mu_1$ . . . 163
6.E Computing μ_1 and τ for Compadre/Comadre . . . 166
6.F Projection matrices for *A. mexicanum* . . . 167

Bibliography 169

Chapter 1

Introduction



1.1 A very brief history of random graphs

When told by mathematicians, the history of random graphs usually goes like this: one of their first notable use is due to Paul Erdős who, in 1947, used a strikingly simple method to provide a lower bound on Ramsey numbers [9]. In the following decade, several other papers related to random graphs were published. In particular, in 1959 Edgar Gilbert formally introduced the Erdős-Rényi random graph for the first time [13] and Erdős again used random graphs to prove graph-theoretic results that have nothing to do with randomness [10].

Finally, the theory was really born when, around 1960, Erdős and Alfred Rényi published a series of papers in which they did a comprehensive study of the model that now bears their name, and proved many of the classic and celebrated results of the theory [8, 11, 12]. That such a simple model could display a rich behavior was stunning, and the mathematical community was quick to recognize the potential of the underlying theory.

According to this account, random graphs were thus invented by mathematicians, for mathematicians.

While this is largely true, this obscures the fact that things were also happening outside of mathematics around the same period, and that the study of networks has been an interdisciplinary science from the beginning. For instance, physicists were already working on percolation theory in the late 1950s [3]. More strikingly, the Erdős-Rényi random graph had been introduced and studied as an interesting object on its own by the psychologist and mathematical biologist Anatol Rapoport, nearly a decade before the seminal papers of Erdős and Rényi [25]. In fact, in their 1951 paper Solomonoff and Rapoport had already correctly identified the phase transition for the giant component or the Erdős-Rényi random graph (it has to be said, though, that the comprehensiveness and rigour of their study was nowhere near that of Erdős and Rényi).

While Rapoport's ideas do not seem to have percolated to the mathematical community, they had a profound impact on social studies, explaining in part the pioneering role that they played in the development of network science.

Over the second half of the 20th century, random graph theory and network science developed at a steady pace, for the most part independently of each other. Things then took an interesting turn in the late 1990s, when better computers and the democratization of the Internet made it possible to study social and information networks on an unprecedented scale. Instead of simply describing the structure of networks, scientists started devising models to understand how networks were formed and evolved. Two major milestones in that respect were the introduction of small-world models by Watts and Strogatz in 1998 [27] and of preferential attachment models by Barabási and Albert in 1999 [1]. Since then, network science has been booming and random graphs have become an established part of the modeler's toolbox, finding applications in countless disciplines. Biology is no exception, and random graphs are now routinely used to study topics ranging from gene regulation to the wiring of the brain.

The historical account given in this section is based on [21] and [2]

1.2 A need for tractable random graphs in evolution

In the study of evolution, random graphs are used for a variety of purposes and at every level of description of evolutionary processes.

At the level of the gene, we now know that in many cases the effect of a gene cannot be considered independently of its interaction with others genes. A major challenge is thus to understand how gene regulatory networks are evolved – and how they affect the dynamics of evolution in return [15].

At the level of individuals and populations, the intra and inter-specific interactions of individuals have a crucial effect or their survival and reproduction. In particular, a substantial part of the literature on the evolution of social behavior is concerned with games on graphs that represent the structure of the population [22]. Other contexts in which random graphs are used include the evolution of food webs [17, 23, 24] and the study of the effect of population structure on gene flow and speciation [7].

At the level of species, random graphs have long played a central role in phylogeny, in the form of random trees used to model phylogenies [26, 16]. More recently, with the advent of DNA sequencing, evidence of the importance of horizontal gene transfers and hybridization on speciation has been accumulating rapidly. This has led several authors to call for a major change of paradigm in the way we think about phylogenies and to advocate the replacement of phylogenetic trees by phylogenetic networks [14].

Despite the great diversity of contexts in which random graphs are used in evolution, at least one general rule seems to have emerged: tractable models are rare. In fact, the majority of models of random graphs that are introduced are only studied through simulations or heuristics; and when in need of analytical tractability, authors often have no choice but to turn to the Erdős-Rényi random graph – which is often unsatisfying from a biological point of view, since this amounts to assuming the absence of any structure specific to the problem at hand.

In phylogeny and population genetics, the need for tractable model is perhaps even more palpable, as researchers have become accustomed to the benefits of working with highly tractable models such as the Kingman's coalescent or Markov branching models [16].

In this context, the aim of this thesis is to identify and study models of random graph that:

1. Emerge naturally from questions related to evolution.

2. Are highly tractable while retaining some mathematical interest.

The challenge of course is that there is a delicate trade-off for the second point: models that exhibit rich behavior tend to be intractable, while tractable models tend to be dull. Some models, such as the Erdős-Rényi random graph, manage to fall perfectly between these two extremes. But these are the exception rather than the rule.

The models that are introduced in this thesis are all highly tractable. I hope the reader will not find them dull.

1.3 Outline of the thesis

This thesis is structured in independent chapters, each corresponding to a different research project.

The aim of Chapter 2 is to give a mechanistic yet tractable model of speciation – or, more generally, a model to describe the structure and dynamics of interbreeding-potential networks. Indeed, it is a widely held view that species should be defined based on the capacity of individuals to interbreed, namely as clusters of populations with high interbreeding potential [5]. However, there are currently no data nor theoretical models to tell us what we should expect these clusters to look like, let alone how they should evolve through time. We thus introduce one such model, in the form of a dynamic random graph whose vertices represent populations and whose edges indicate interbreeding potential. In its definition, this random graph is reminiscent of classic models of protein interaction networks, but it turns out to be much more tractable and to have a very different behavior. In particular, this model produces dense clusters that are poorly connected to each other, in agreement with the biological intuition behind the definition of species and in contrast to most of the classic models of random graphs.

In Chapter 3, we introduce a very natural model of random forest. Although the idea for this model came from very specific biological questions, its interest lies in the simplicity of its definition and its connections to other random objects such as the Moran model [19], uniform random recursive trees [6] and uniform rooted labeled trees. Indeed, one of the main features of this random forest is that it can be built either from a graph-valued Markov chain, from a uniform attachment procedure, or from a uniform rooted labeled tree.

Chapter 4 is devoted to a specific model of phylogenetic networks. Tree-child networks form a class of phylogenetic networks that has gained traction among combinatorists and phylogeneticists in recent years [4, 18, 28, 28]. Unfortunately, they are not very tractable: for instance, it is not even known how to count them. By adding some appropriate structure to tree-child networks, we introduce a very tractable class of phylogenetic networks that are easy to count, sample and study analytically. These new phylogenetic networks also have the advantage of being more relevant from a biological point of view: indeed, their additional structure ensures that they could have resulted from evolutionary processes – unlike general tree-child networks, the majority of which could not have been produced by time-embedded processes compatible with evolution.

The result discussed in Chapter 5 is not tied to any particular biological setting and is a general property of randomly oriented graphs – namely, the positive association of the oriented percolation cluster. This generalizes and simplifies the main result of [20]. This technical result can be useful in applications, and to illustrate this a toy model of percolation is studied in detail.

Finally, the last chapter is a bit of an intruder. During my undergraduate studies, I became interested in structured populations and the various measures of generation time. I kept this interest during my PhD and this led to the work presented in Chapter 6, in which I point out serious problems associated with one of the most widely used measures of the mean age at reproduction and suggest an alternative way to quantify it. While this work is not directly related to random graphs and involves more biology than mathematics, I chose to include it in this thesis for two

reasons: (1) it was made possible by some basic knowledge about Poisson point processes which, having been trained as a biologist, I lacked before starting this PhD and learned as a beneficial side effect of my work on random graphs; and (2) it is representative of my research interests and of the fact that in addition to doing mathematics I want to keep working with biologists.

Literature cited the introduction

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] B. Bollobás. *Random graphs*. Cambridge University Press, 2001.

[3] S. R. Broadbent and J. M. Hammersley. Percolation processes: I. crystals and mazes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(3):629–641, 1957.

[4] G. Cardona, F. Rossello, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):552–569, 2009.

[5] J. A. Coyne and H. A. Orr. *Speciation*. Sinauer Associates, 2004.

[6] M. Drmota. *Random trees: an interplay between combinatorics and probability*. Springer-Verlag Vienna, 2009.

[7] R. J. Dyer and J. D. Nason. Population graphs: the graph theoretic shape of genetic structure. *Molecular ecology*, 13(7):1713–1727, 2004.

[8] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–61, 1960.

[9] P. Erdős. Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, 53(4):292–294, 1947.

[10] P. Erdős. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

[11] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[12] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1-2):261–267, 1961.

[13] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[14] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2005.

[15] E. V. Koonin, Y. I. Wolf, and G. P. Karev. *Power laws, scale-free networks and genome biology*. Springer US, 2006.

[16] A. Lambert. Probabilistic models for the (sub)tree(s) of life. *Brazilian Journal of Probability and Statistics*, 31(3):415–475, 2017.

[17] R. M. May. Network structure and the biology of populations. *Trends in Ecology & Evolution*, 21(7):394–399, 2006.

[18] C. McDiarmid, C. Semple, and D. Welsh. Counting phylogenetic networks. *Annals of Combinatorics*, 19(1):205–224, 2015.

[19] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

[20] B. Narayanan. Connections in randomly oriented graphs. *Combinatorics, Probability and Computing*, pages 1–5, 2016.

[21] M. Newman, A.-L. Barabási, and D. J. Watts. *The structure and dynamics of networks*, volume 12. Princeton University Press, 2011.

[22] M. A. Nowak. *Evolutionary dynamics*. Harvard University Press, 2006.

[23] M. Pascual, J. A. Dunne, et al. *Ecological networks: linking structure to dynamics in food webs*. Oxford University Press, 2006.

[24] S. R. Proulx, D. E. L. Promislow, and P. C. Phillips. Network thinking in ecology and evolution. *Trends in ecology & evolution*, 20(6):345–353, 2005.

[25] R. Solomonoff and A. Rapoport. Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117, 1951.

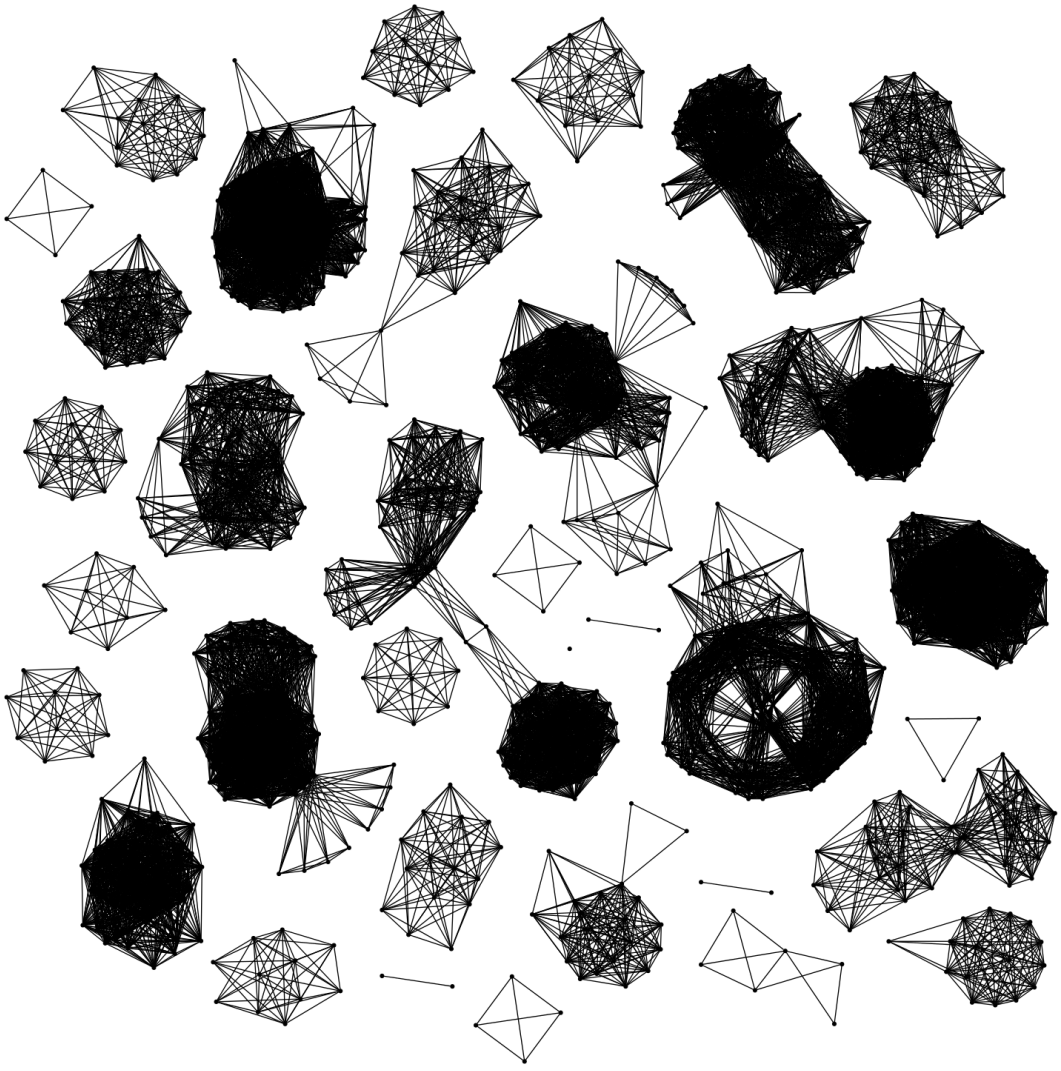
[26] M. Steel. *Phylogeny: discrete and random processes in evolution*. SIAM, 2016.

[27] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[28] S. J. Willson. Unique determination of some homoplasies at hybridization events. *Bulletin of mathematical biology*, 69(5):1709–1725, 2007.

Chapter 2

The split-and-drift random graph



This chapter is joint work with Florence Débarre and Amaury Lambert. It started during the research internship that preceded my thesis. At this time, Amaury was very interested in the biological species concept and in the geometry of interbreeding-potential networks (which phenomena such as ring species [10]). He therefore suggested I look for a biologically relevant model of random graph that would produce clusters.

After comparing several candidates through computer simulations, the model presented in this chapter seemed to stand out as the best compromise between simplicity and richness of behaviour.

Publication: This chapter has been published in *Stochastic Processes and their Applications* under the title "The split-and-drift random graph, a null model for speciation" [2].

Chapter contents

2.1 Introduction . . . 13
2.1.1 Biological context . . . 13
2.1.2 Formal description of the model . . . 14
2.1.3 Notation . . . 14
2.1.4 Statement of results . . . 15
2.2 Coalescent constructions of G_{n,r_n} . . . 17
2.2.1 The standard Moran process . . . 17
2.2.2 Backward construction . . . 18
2.2.3 Forward construction . . . 21
2.3 First and second moment methods . . . 22
2.3.1 First moments of graph invariants . . . 22
2.3.2 Identification of different regimes . . . 27
2.4 The degree distribution . . . 28
2.4.1 Ideas of the proof of Theorem 2.4.1 . . . 28
2.4.2 Formal proof of Theorem 2.4.1 . . . 30
2.5 Connected components in the intermediate regime . . . 33
2.5.1 Lower bound on the number of connected components . . . 33
2.5.2 Upper bound on the number of connected components . . . 34
2.6 Number of edges in the sparse regime . . . 36
2.6.1 Proof of the positive relation between the edges . . . 36
2.6.2 Proof of Theorem 2.6.1 . . . 40
Chapter references . . . 40
2.A Proofs of Propositions 2.2.4 and 2.2.6 and of Lemma 2.2.5 . . . 42
2.A.1 Proof of Propositions 2.2.4 and 2.2.6 . . . 42
2.A.2 Proof of Lemma 2.2.5 . . . 44
2.B Proofs of Proposition 2.3.5 and Corollary 2.3.6 . . . 45
2.B.1 Proof of Proposition 2.3.5 . . . 46
2.B.2 Proof of Corollary 2.3.6 . . . 47
2.C Proof of Theorem 2.4.2 . . . 48
2.C.1 Outline of the proof . . . 48
2.C.2 Step 1 . . . 49
2.C.3 Step 2 . . . 49

2.1 Introduction

In this chapter, 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 [5, 11, 20, 23] 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 2.1.2, let us briefly explain the motivation behind its introduction.

2.1.1 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 [18], was most famously promoted by Ernst Mayr [16] and has been most influential in biology [6]. 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 2.1. 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.

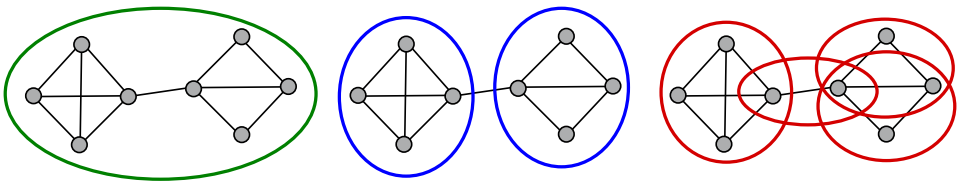


Figure 2.1: The vertices of the graph represent populations and its edges denote interbreeding potential (that is, individuals from two linked populations could interbreed, if given the chance). Even with such perfect information, it is not obvious how to delineate "groups of potentially interbreeding 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.

2.1.2 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 2.2.

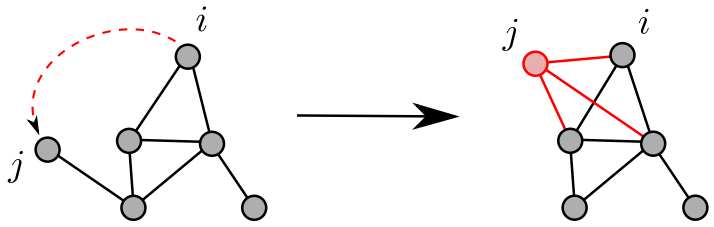


Figure 2.2: An illustration of vertex duplication. Here, i duplicates and replaces j. After the duplication, j is linked to i and to each of its neighbors.

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 chapter.

2.1.3 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

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

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 indicate that a property holds with probability that goes to 1 as n tends to infinity:

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

2.1.4 Statement of results

Table 2.1 lists the first moments of several graph invariants obtained in Section 2.3.1. These are then used to identify different regimes, depending on the asymptotic behavior of the parameter r_n, as stated in Theorem 2.3.10.

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, $\frac{2r}{(1+r)^2(3+r)(3+2r)}$ 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 2.1: 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.

Theorem 2.3.10. *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.*

In Section 2.4, 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 2.4.1. *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 2.4.2.

(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 2.5, where the following theorem is proved.

Theorem 2.5.1. *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*

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

where, for a positive sequence (u_n), $o_{\mathrm{p}}(u_n)$ denotes a 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 2.5.1 is rather crude, we formulate the following conjecture, which is well supported by simulations.

Conjecture 2.5.4.

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

Finally, in Section 2.6 we use the Stein–Chen method to show that the number of edges is Poissonian in the sparse regime, as shown by Theorem 2.6.1.

Theorem 2.6.1. *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 2.3.

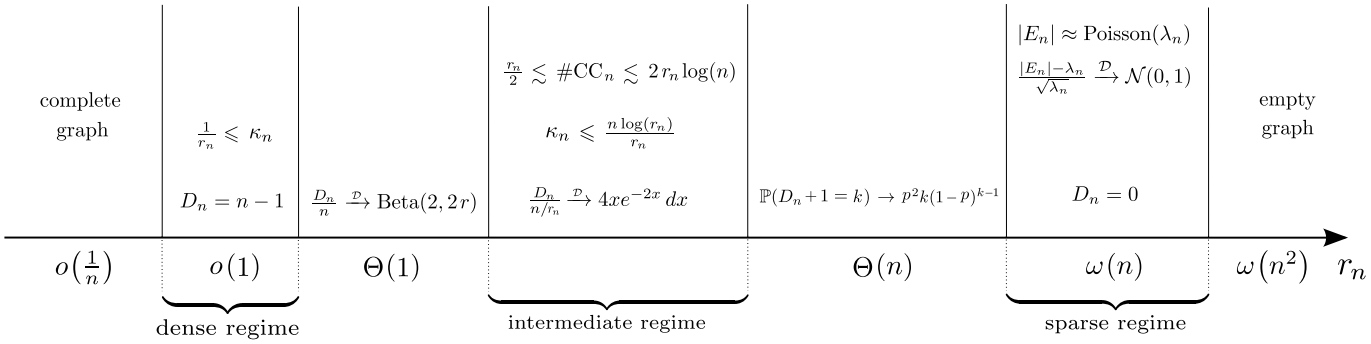


Figure 2.3: A graphical summary of the main results; D_n is the degree of a fixed vertex, $|E_n|$ the number of edges, $\#\mathrm{CC}_n$ the number of connected components, and κ_n the clique number. All equalities and inequalities are to be understood "asymptotically almost surely" (i.e. hold with probability that goes to 1 as n tends to infinity).

2.2 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 chapter. Let us start by recalling some results about the Moran model.

2.2.1 The standard Moran process

The Moran model [17] 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 chapter, we will use the Poissonian representation of the Moran process detailed in the next definition.

Definition 2.2.1. 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 [12]. 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 2.2.2. *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. [7] or [9].

2.2.2 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 2.2.3. *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*

$$P^\star_{\{ij\}} = \Big\{t \geqslant 0 : t \in P_{\{a_t(i)a_t(j)\}}\Big\}\,,$$

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

$$E = \Big\{\{ij\} \in V^{(2)} : P^\star_{\{ij\}} = \emptyset\Big\}\,.$$

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

Throughout the rest of this chapter, we will write G_{n,r_n} for the graph obtained by the procedure of Proposition 2.2.3.

Proof of Proposition 2.2.3. 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 [8]), 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 2.2.3 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 2.2.3. The proof of this result is straightforward and can be found in Section 2.A of the Appendix.

Proposition 2.2.4. *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*

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

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

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

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

We finish this section with a technical lemma that will be useful in the calculations of Section 2.3.1. Again, the proof of this result has no interest in itself and can be found in Section 2.A of the Appendix.

Lemma 2.2.5. *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, $\forall t \in I$, $\{a_t(i)\, a_t(j)\} \neq \{a_t(k)\, a_t(\ell)\}$,*

we have that $(P^\star_{\{ij\}} \cap I,\, \{ij\} \in S)$, 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 2.4. 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.

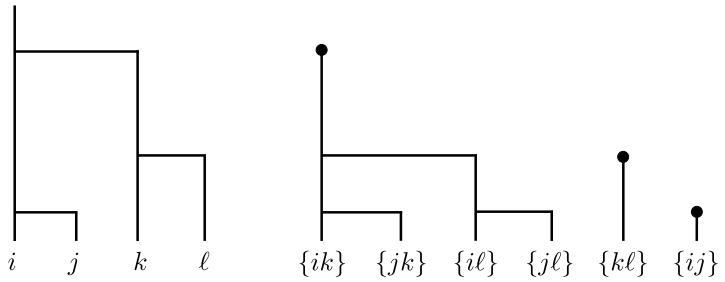


Figure 2.4: On the left, a genealogy on $\{i, j, k, \ell\}$ and on the right the corresponding genealogy of the pairs. Edge removal events occur at constant rate along the lineages of pairs of vertices, and a pair of vertices is an edge of G_{n,r_n} if and only if there is no atom on its lineage.

2.2.3 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

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

Proposition 2.2.6. *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 2.2.4 to a forward construction such as Proposition 2.2.6 is common in coalescent theory. The proofs, though straightforward, are somewhat tedious. They can be found in Section 2.A of the Appendix, and we save the rest of this section to comment on the forward construction.

Proposition 2.2.6 shows that, for any given sequence (r_n), for any $n \geqslant 2$, $\Phi_n(G_{r_n}^\star(n)) \sim G_{n,r_n}$. 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 $(\Phi_n(G_r^\star(n)))_{n\geqslant 2}$ 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 $G_r^\star(n)$ in Proposition 2.2.6, 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 $G_r^\star(n)$ 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 $G_r^\star(n)$ and $\Phi_n(G_r^\star(n))$ are isomorphic, when studying properties that are invariant under graph isomorphism (such as the number of connected components in Section 2.5 or the positive association of the edges in Section 2.6), we can work directly on $G_r^\star(n)$.

2.3 First and second moment methods

In this section, we apply Proposition 2.2.3 and Lemma 2.2.5 to obtain the expressions presented in Table 2.1. 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 2.2.5, 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.

2.3.1 First moments of graph invariants

Degree and number of edges

Proposition 2.3.1. *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 2.3.2. *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 2.2.3,

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

Reasoning conditionally on $T_{\{ij\}}$ and applying Lemma 2.2.5 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 $T_{\{ij\}} \sim \mathrm{Exp}(1)$,

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

where $e_1 = \inf P^{\star}_{\{ij\}}$ is an exponential variable with rate r_n that is independent of $T_{\{ij\}}$. This concludes the proof of the proposition.

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

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

Corollary 2.3.4. *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 2.2.2 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 2.2.3, $\{ij\}$ 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 2.2.5, $(P^\star_{\{ij\}} \cap [0, \tau_1[) \cup (P^\star_{\{ik\}} \cap [0, \tau_1[)$ 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 $D^{(v)}_n = \sum_{i \neq v} \mathbb{1}_{\{i \leftrightarrow v\}}$, which gives

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

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

Proposition 2.3.5. *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 2.3.6. *Let $D^{(i)}_n$ and $D^{(j)}_n$ 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*

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

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 2.3.5 and its corollary are conceptually identical to the proofs of Propositions 2.3.1 and 2.3.3 and their corollaries, but the calculations are more tedious and so they have been relegated to Section 2.B of the Appendix.

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 2.5.

In this section we establish the following results.

Proposition 2.3.7. *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 2.3.8. *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*

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

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 $r_n = O(n/\log(n))$, $\kappa_n = O_{\mathrm{p}}(n/r_n)$, 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 2.3.7. 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,

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

where S is any fixed set of k vertices. Using the notation of Proposition 2.2.3,

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

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 2.2.5 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 2.2.5 that, conditional on I_ℓ, $(P^\star_{\{ij\}} \cap I_\ell)$, $\{ij\} \in \widetilde{A}_\ell^{(2)}$, are i.i.d. Poisson point processes with rate r_n on I_ℓ. Therefore,

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

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

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

and Proposition 2.3.7 follows. □

Proof of Corollary 2.3.8. The first part of the corollary is a direct consequence of Proposition 2.3.7. 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 $n \to +\infty$,

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

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 $n\coloneqq\min\{n_k : n_k\geqslant N\}$ such that $\mathbb{P}(\kappa_n>g_nn/r_n)\geqslant\varepsilon$. □

A natural pendant to Proposition 2.3.7 and Corollary 2.3.8 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

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

where $p=1/(1+r_n)^{k-1}$. This shows that when $r_n\to0$ 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 2.5.1 and Conjecture 2.5.4), and since the degree of a typical vertex is approximately n/r_n in that regime, it seems reasonable to conjecture

Conjecture 2.3.9. *In the intermediate regime, i.e. when $r_n\to+\infty$ and $r_n=o(n)$,*

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

2.3.2 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 chapter (namely, Theorems 2.4.1 and 2.6.1), but no subsequent result depends on it, avoiding cyclic dependencies. While this section could have been placed at the end of the chapter, it makes more sense to present it here because it relies mostly on Section 2.3.1 and because it helps structure the rest of the chapter.

Theorem 2.3.10. *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 2.3.7 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 2.4.1, 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 2.3.8, 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 2.6.1, 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 2.3.11. 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. ◇

2.4 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 2.4.1 (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 2.4.2 (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 2.4.1 by coupling the degree to the number of individuals descended from a founder in a branching process with immigration. Theorem 2.4.2 is then easily deduced by a standard study that has been relegated to Section 2.C of the Appendix.

2.4.1 Ideas of the proof of Theorem 2.4.1

Before jumping to the formal proof of Theorem 2.4.1, 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 2.5.

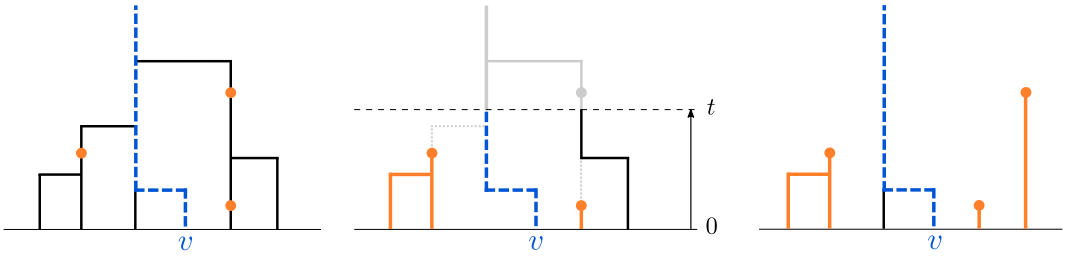


Figure 2.5: Illustration of the procedure used to find the neighborhood of v. On the left, the genealogy of the vertices. The dashed blue line represents the lineage of the focal vertex v, and a dot on lineage k corresponds to a point in $P_{\{ka_t(v)\}}$. In the middle, we uncover the genealogy and 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 [3, 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.

2.4.2 Formal proof of Theorem 2.4.1

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,

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

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 2.4.3. 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. ◇

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 2.4.1. Recall the notation of Proposition 2.2.3. 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 $(a_{t-}(k), a_{t-}(\ell)) = (i, j)$,
$$\vec{E}_t = \vec{E}_{t-} \cup \{(k, \ell)\}.$$
- 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

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

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 $(V_t^\circ, V_t^\bullet, \vec{E}_t)$ to
  - $\left(V_t^\circ,\, V_t^\bullet,\, \vec{E}_t \cup \{(i,j)\}\right)$ at rate $1/2$, for all i, j in $R(F_t)$
  - $\left(V_t^\circ \setminus T_{F_t}(i),\, V_t^\bullet \cup T_{F_t}(i),\, \vec{E}_t\right)$ 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

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

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

$$\begin{aligned}\mathbb{P}\left(\widetilde{F}_n = f_n\right) &= \frac{(1/2)^{n-|R(f_n)|}\ r_n^{|R(f_n)|-1}}{\prod_{k=2}^{n}\left(k(k-1)/2 + (k-1)r_n\right)} \\ &= \frac{1}{(n-1)!} \times \frac{(2\,r_n)^{|R(f_n)|-1}}{\prod_{k=2}^{n}\left(k + 2\,r_n\right)} \qquad (2.1)\end{aligned}$$

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

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

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

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

— $(W^\circ \cup \{N\},\, W^\bullet,\, \vec{E} \cup \{(N,c)\},\, N)$ at rate 1

— $(W^\circ,\, W^\bullet \cup \{N\},\, \vec{E},\, c)$ 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

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

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

$$\mathbb{P}\Big(\widetilde{Z}_n = (W_n^\circ, W_n^\bullet, \vec{E}_n, c_n)\Big) = \frac{(2\,r_n)^{|R(f_n)|-1}}{\prod_{k=1}^{n-1}(k+1+2\,r_n)}\,. \tag{2.2}$$

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 $\frac{q+2r}{p+1+q+2r_n}$.

Relabeling and end of proof

The last step before finishing the proof of Theorem 2.4.1 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 $\Big\{(\Phi_{(\tilde{c}_n,v)}(i), \Phi_{(\tilde{c}_n,v)}(j)) : (i,j) \in \widetilde{E}_n\Big\}$.

With all these elements, the proof of Theorem 2.4.1 goes as follows. First, from equations (2.1) and (2.2) 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

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

and on the other hand we have

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

Since $\left|\widetilde{W}_n^\circ\right|$ 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

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

from which the expression of Theorem 2.4.1 can be deduced through elementary calculations.

2.5 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 2.3.1). 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 2.5.1. *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*

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

where, for a positive sequence (u_n), $o_{\mathrm{p}}(u_n)$ denotes a sequence of random variables (X_n) such that $X_n/u_n \to 0$ in probability.

2.5.1 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 2.2.2 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 $\mathbb{P}(\#\mathrm{CC}(G^\star_{r_n}(m_n)) \geqslant \ell_n) \to 1$ 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 2.1. We see that, if r_n and m_n are both $\omega(1)$ and $o(n)$,

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

By Chebychev's inequality,

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

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

$$|E^\star_{m_n}| = \mathbb{E}(|E^\star_{m_n}|) + o_{\mathrm{p}}(r_n)\,.$$

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

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

The right-hand side is maximal for $\alpha = 1$ and is then $r_n/2 + o_{\mathrm{p}}(r_n)$.

2.5.2 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 2.5.2. 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 2.2.3. By construction, $\forall \{ij\} \in \mathcal{F}$, $P^\star_{\{ij\}} = P_{\{ij\}} \cap [0, T_{\{ij\}}]$. It follows that

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

and, as a consequence,

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

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_nL_n. Moreover, it is known [22] 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,

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

and

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

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 $\#(P\cap[0,L_n]) - 2\,r_n\log n = o_{\mathrm{p}}(r_n\log n)$ and finishes the proof.

Remark 2.5.3. 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 2.5.4.

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

Remark 2.5.5. After this work was published, the following argument supporting Conjecture 2.5.4 was given to me by Gustave Emprin: let $|\mathrm{CC}(v)|$ denote the size of the connected component containing vertex v. Then we have

$$\#\mathrm{CC}_n \;=\; \sum_v \frac{1}{|\mathrm{CC}(v)|} \;\leqslant\; \sum_v \frac{1}{D_n^{(v)}+1} \;\sim\; 2r_n$$

where the asymptotic equivalent follows from point (ii) of Theorem 2.4.1 ◇

2.6 Number of edges in the sparse regime

From the expressions obtained in section 2.3.1 and recapitulated in Table 2.1, 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 2.6.1. *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 2.6.1 is a standard application of the Stein–Chen method [21, 4]. A reference on the topic is [1], and another excellent survey is given in [19]. 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) $(X_1^{(i)}, \dots, X_N^{(i)}) \sim (X_1, \dots, X_N) \mid X_i = 1$.

(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 [19]). Finally, we will need the following classic theorem, which appears, e.g., as Theorem 4.20 in [19].

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,*

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

2.6.1 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.

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 2.6.2. *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

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

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 2.6.3. *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 $(X_j)_{j\neq i}$ and (T, τ_i) are independent, $(X_j^{(i)})_{j\neq i} \sim ((X_j(T))_{j\neq i} \mid X_i(T) = 1)$. This shows that $X_1(T), \dots, X_N(T)$ are positively related. □

Remark 2.6.4. Lemma 2.6.3 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. ◇

Stein–Chen coupling

Proposition 2.6.5. *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 2.2.3 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.

$$\begin{aligned} &\forall i \leqslant N, \exists \mathbf{Y}^{(i)} = (Y_1^{(i)}, \dots, Y_N^{(i)}) \text{ such that} \\ &\quad \text{(i)} \quad \mathbf{Y}^{(i)} \sim (\mathbf{X} \mid X_i = 1) \\ &\quad \text{(ii)} \quad \forall k \leqslant N,\ Y_k^{(i)} \geqslant X_k \end{aligned} \tag{2.3}$$

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 $X'_k = \mathbb{1}_{\{k \in E^\star_{n+1}\}}$, 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 2.6.3 (see also Remark 2.6.4), 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 (2.3).
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,

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

Similarly,

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

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 2.6.2, 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 $\mathbf{Y}^{(i)}$, $\mathbf{Y}'^{(i)}_k \geqslant \mathbf{X}'_k$. □

2.6.2 Proof of Theorem 2.6.1

Applying Theorem A to $|E_n| = \sum_{\{ij\}} \mathbb{1}_{\{i \leftrightarrow j\}}$ and using the expressions in Table 2.1, 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 2.6.1.

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., [1], page 17, where this is recovered as a consequence of inequality (1.39).

Literature cited in this chapter

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

[2] F. Bienvenu, F. Débarre, and A. Lambert. The split-and-drift random graph, a null model for speciation. *Stochastic Processes and their Applications*, 129(6):2010–2048, 2019.

[3] B. Chauvin, A. Rouault, and A. Wakolbinger. Growing conditioned trees. *Stochastic Processes and their Applications*, 39(1):117–130, 1991.

[4] L. H. Y. Chen. Poisson approximation for dependent trials. *Annals of Probability*, 3(3):534–545, 1975.

[5] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas. Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687, 2003.

[6] J. A. Coyne and H. A. Orr. *Speciation*. Sinauer Associates, 2004.

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

[8] R. Durrett. *Probability: theory and examples*. Cambridge University Press, 4th edition, 2010.

[9] A. Etheridge. *Some mathematical models from population genetics. École d'été de probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer-Verlag Berlin Heidelberg, 2011.

[10] D. E. Irwin, J. H. Irwin, and T. D. Price. Ring species as bridges between microevolution and speciation. In *Microevolution rate, pattern, process*, pages 223–243. Springer Netherlands, 2001.

[11] I. Ispolatov, P. Krapivsky, and A. Yuryev. Duplication-divergence model of protein interaction network. *Physical Review E*, 71(6):061911, 2005.

[12] J. F. C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[13] A. Lambert. Probabilistic models for the (sub)tree(s) of life. *Brazilian Journal of Probability and Statistics*, 31(3):415–475, 2017.

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

[15] M. Manceau and A. Lambert. The species problem from the modeler's point of view. *Bulletin of Mathematical Biology*, 81:878–898, 2019.

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

[17] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

[18] E. B. Poulton. What is a species? In *Proceedings of the Entomological Society of London 1903*, pages lxxvii–cxvi, 1904.

[19] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:201–293, 2011.

[20] R. Solé, R. Pastor-Satorras, E. Smith, and T. B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(1):43–54, 2002.

[21] C. M. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 583–602. University of California Press, 1972.

[22] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164, 1984.

[23] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2003.

Appendices to Chapter 2

2.A Proofs of Propositions 2.2.4 and 2.2.6 and of Lemma 2.2.5

2.A.1 Proof of Propositions 2.2.4 and 2.2.6

Proposition 2.2.4. *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*

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

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

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

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

Proof. Let $(a_t)_{t\geqslant 0}$ and $\mathcal{P}^\star$ be as in Proposition 2.2.3, and let

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

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,

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

With this formalism, writing

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

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

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

which terminates the proof. □

Proposition 2.2.6. *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 $\mathcal{P}^\dagger = (P^\dagger_{\{ij\}})_{\{ij\}\in V^{(2)}}$ 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

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

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

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

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

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

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 2.2.4, 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. [13]).

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 $U_i \sim \mathrm{Uniform}(\{1,\dots,n-i\})$, 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 $i \geqslant 1$,

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

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. □

2.A.2 Proof of Lemma 2.2.5

Lemma 2.2.5. *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, $\forall t \in I$, $\{a_t(i), a_t(j)\} \neq \{a_t(k), a_t(\ell)\}$,*

we have that $(P^\star_{\{ij\}} \cap I,\, \{ij\} \in S)$, 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.

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

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

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

and

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

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. □

2.B Proofs of Proposition 2.3.5 and Corollary 2.3.6

Proposition 2.3.5. *Let i, j, k and ℓ be four distinct vertices of G_{n,r_n}. We have*

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

Corollary 2.3.6. *Let $D^{(i)}_n$ and $D^{(j)}_n$ 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*

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

and

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

2.B.1 Proof of Proposition 2.3.5

The proof of Proposition 2.3.5 parallels that of Proposition 2.3.3, 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

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

so that, conditionally on I_1 and I_2, by Lemma 2.2.5,

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

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

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

and thus

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

Given a realization of the topology of the genealogy in 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 2.6.

Putting the pieces together, we find that

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

and Proposition 2.3.5 follows, since

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

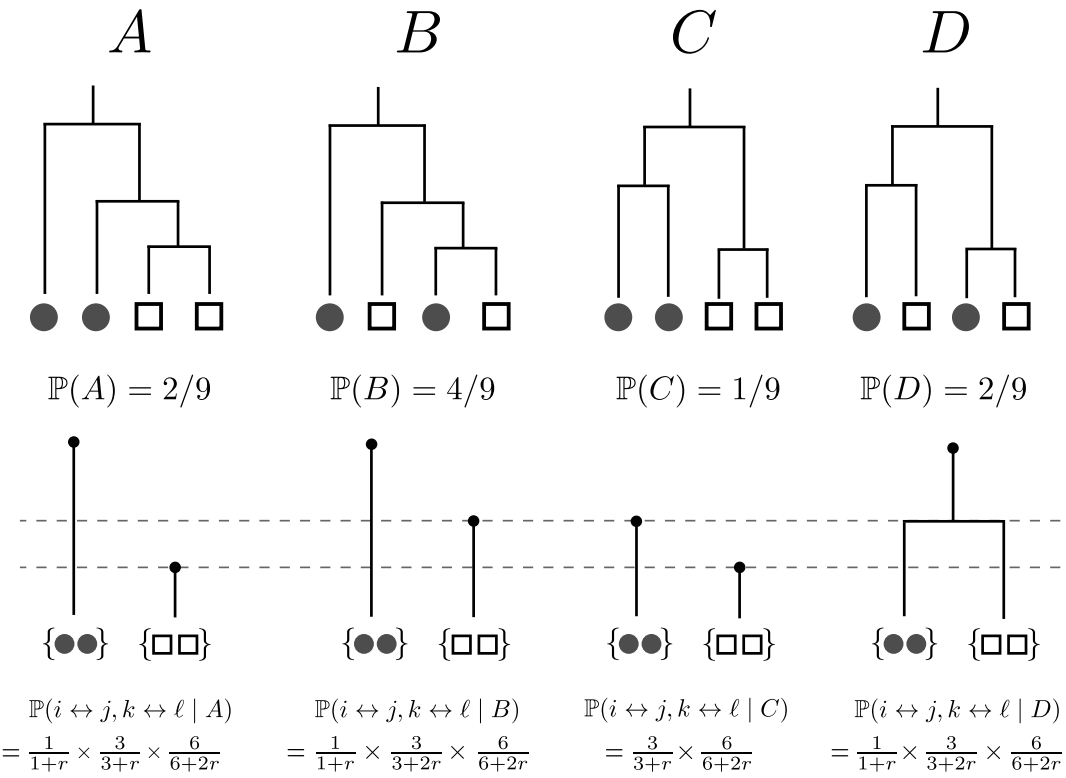


Figure 2.6: The four cases that we consider to compute $\mathbb{P}(i \leftrightarrow j, k \leftrightarrow \ell)$. Top, the "aggregated" genealogies of vertices and their probability. Each of these correspond to several genealogies on $\{i, j, k, \ell\}$, which are obtained by labeling symbols in such 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 $\{i \leftrightarrow j, k \leftrightarrow \ell\} \Leftrightarrow \{\square \leftrightarrow \square, \bullet \leftrightarrow \bullet\}$.

2.B.2 Proof of Corollary 2.3.6

Corollary 2.3.6 is proved by standard calculations. First,

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

Remembering from Proposition 2.3.1 that $\mathrm{Var}(\mathbb{1}_{\{i\leftrightarrow j\}}) = r_n/(1+r_n)^2$ and from Proposition 2.3.3 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 2.3.5, we find that

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

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,

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

2.C Proof of Theorem 2.4.2

In this section, we prove Theorem 2.4.2.

Theorem 2.4.2 (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 $\rho/(1+\rho)$.*

First, note that the proof of (iii) is immediate: indeed, by Theorem 2.4.1,

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

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

Let us now focus on the proof of (i) and (ii).

2.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

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

2.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.

2.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

$$f_n(x) = \frac{N_n\,2\,r_n\,(2\,r_n+1)}{(n+2\,r_n)(n-1+2\,r_n)}\,(\lfloor N_n x \rfloor + 1) \times P_n(x)\,, \tag{2.4}$$

where

$$P_n(x) = \frac{\Gamma(n)\,\Gamma(n-\lfloor N_n x \rfloor + 2\,r_n - 1)}{\Gamma(n-\lfloor N_n x \rfloor)\,\Gamma(n+2\,r_n-1)}\,. \tag{2.5}$$

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

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 (2.5), 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 $\mathrm{Beta}(2,2r)$ 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 $g \colon x \mapsto c\, x\, \mathbb{1}_{[0,1[}(x)$.

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 (2.5),

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

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 $Q_n \to e^{-2x}$:

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

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 (2.4), 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

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

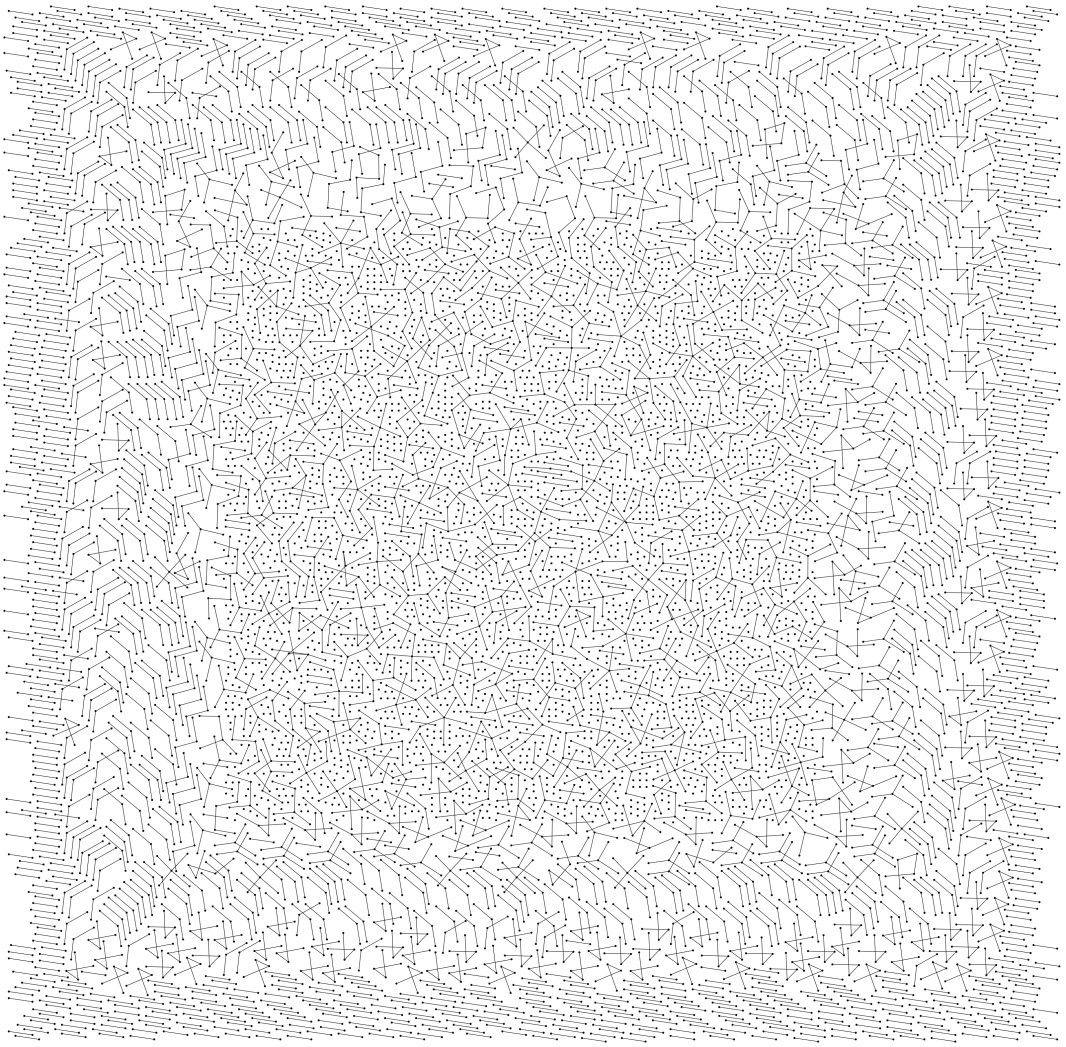
for n large enough. Finally,

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

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

Chapter 3

The Moran forest



The model presented in this chapter arose in two very different contexts: the first one was when talking with Guilhem Doulcier about the phenomenon of snowflake yeasts, where a daughter cell remains glued to its mother after mitosis [18]. The Moran forest provides a very natural way to model this situation, but because none of us had any real incentive to start a project on snowflake yeasts, we did not go any further than running a few simulations.

The second occasion in which this model came up was when Amaury Lambert, Félix Foutel-Rodier and I tried to think of a dynamic, forward-in-time model of population structure that would make it possible to track the movements of genes backwards in time. Even though the Moran forest turned out to be of no use for this, it was a very natural candidate so Félix and I started thinking about it. After realizing that the forward-in-time version of the coalescent construction of the model was so simple and elegant, we decided to start working on it and were quickly joined by Jean-Jil Duchamps.

So what exactly is the Moran forest? According to the Internet[1],

> The Moran Forest is a thick forest filled with magic that no human could pass easily [...]. It has even come to be called by humans as the "Forest of Illusion" due to its magical nature.

While I would agree that there is some magic in this forest, I am clearly biased in that regard and the paragraph above seems a bit excessive. In particular, it is hard to argue that it is "thick" when it mostly consists of very small trees, the largest of which is of size $\propto \log n$.

Publication: This chapter has been submitted for publication in *Random Structures and Algorithms* under the title "The Moran forest".

Chapter contents

3.1 Introduction . . . 55
3.1.1 The model . . . 55
3.1.2 Main results . . . 55
3.2 Sampling of the stationary distribution . . . 57
3.2.1 Backward construction . . . 57
3.2.2 Uniform attachment construction . . . 58
3.3 Number of trees . . . 60
3.3.1 Law of the number of trees . . . 60
3.3.2 Link with uniform labeled trees . . . 61
3.4 Degrees . . . 63
3.4.1 Degree of a fixed vertex . . . 63
3.4.2 Largest degree . . . 66
3.5 Tree sizes . . . 70
3.5.1 A discrete-time Yule process . . . 70
3.5.2 Size of some random trees . . . 73
3.5.3 Size of the largest tree . . . 75
Chapter references . . . 78
3.A Proof of point (ii) of Proposition 3.4.4 . . . 80
3.B Technical lemmas used in the proof of Theorem 3.5.8 . . . 82

[1] https://suikoden.fandom.com/wiki/Moran_Forest

3.1 Introduction

3.1.1 The model

Consider a Markov chain on the space of graphs on $\{1, \dots, n\}$ whose transition probabilities are defined as follows: at each time-step,

1. Choose an ordered pair of distinct vertices (u, v) uniformly at random.
2. Disconnect v from all of its neighbors, then connect it to u.

Note that if u is the only neighbor of v at time t, then the graph is unchanged at time $t + 1$. A simple example illustrating the dynamics of this Markov chain is depicted in Figure 3.1.

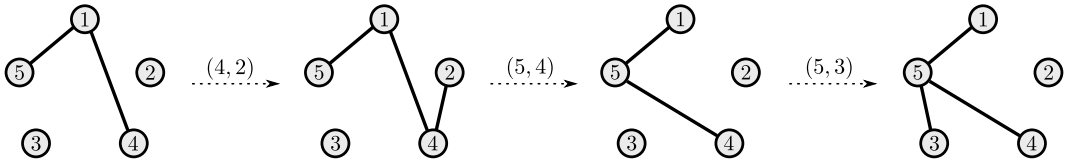


Figure 3.1: Example of four successive transitions of the Markov chain. Starting from the left-most graph, transitions are represented by dashed arrows decorated with the pair (u, v) which is chosen uniformly at each step.

This Markov chain has a stationary distribution whose support is the set of non-empty forests on $\{1, \dots, n\}$. Indeed,

- The graph cannot be empty because there is always an edge between the two vertices involved in the last transition.
- Starting from any graph, the chain will eventually reach a forest (for instance, the sequence of transitions $(1, 2), (1, 3), \dots, (1, n)$ will at some point turn the graph into the star graph centered on vertex 1).
- The chain cannot leave the set of non-empty forests because its transitions cannot create cycles.
- Any non-empty forest is accessible from any other graph (if not clear, this will become apparent in Section 3.2).
- The chain is aperiodic because it can stay in the same state.

The stationary distribution of this chain is the random forest model that we study in this paper. We denote it by $\mathscr{F}_n$ and call it the *Moran forest*, in reference to the Moran model of population genetics, where at each time step two distinct individuals are sampled uniformly at random, and the second one is replaced by a copy of the first [17, 9, 11].

3.1.2 Main results

Our first result, which we detail in Section 3.2, is that there is a simple way to sample $\mathscr{F}_n$. This construction enables us to study several of its statistics, such as its number of trees (Section 3.3.1), its degree distribution (Section 3.4.1), and the typical size of its trees (Section 3.5.2). Some of these results are presented in Table 3.1.

Notation	Variable	Distribution
N_n	Number of trees	$\sum_{\ell=1}^{n} I_\ell$, where $I_\ell \sim \mathrm{Ber}\left(\frac{\ell}{n-1}\right)$
D	Asymptotic degree distribution	$\mathrm{Ber}(1-U) + \mathrm{Poisson}(U)$, where $U \sim \mathrm{Unif}([0,1])$
T^U	Asymptotic size of a uniform tree	$\mathrm{Geometric}(e^{-X})$, where $X \sim 2x dx$ on $[0,1]$

Table 3.1: Some statistics of the Moran forest, for fixed n in the case of the number of trees, and as $n \to \infty$ for the degree and the size of a uniform tree. Note that the degree also has a simple, explicit distribution for fixed n (see Proposition 3.4.1). The Bernoulli variables I_ℓ used to describe the distribution of N_n are independent and, conditional on U, so are the Bernoulli and Poisson variables used for the distribution of D.

In Section 3.3.2, we show that the Moran forest is closely linked to uniform rooted labeled trees. Specifically, we prove the following theorem.

Theorem 3.3.4. *Let $\mathcal{T}$ be a uniform rooted tree on $\{1,\dots,n-1\}$. From this tree, build a forest $\mathcal{F}$ on $\{1,\dots,n\}$ according to the following procedure:*

1. *Remove all decreasing edges from $\mathcal{T}$ (that is, edges $\vec{uv}$ pointing away from the root such that $u > v$).*
2. *Add a vertex labeled n and connect it to a uniformly chosen vertex of $\mathcal{T}$*
3. *Relabel vertices according to a uniform permutation of $\{1,\dots,n\}$.*

Then, the resulting forest $\mathcal{F}$ has the law of the Moran forest $\mathscr{F}_n$.

Finally, we study the asymptotic concentration of the largest degree and of the size of the largest tree of $\mathscr{F}_n$. The following theorems are proved in Sections 3.4.2 and 3.5.3, respectively.

Theorem 3.4.5. *Let $D_n^{\max}$ denote the largest degree of $\mathscr{F}_n$. Then,*

$$D_n^{\max} = \frac{\log n}{\log\log n} + (1+o_{\mathrm{p}}(1))\,\frac{\log n \,\log\log\log n}{(\log\log n)^2},$$

where $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

Theorem 3.5.8. *Let $T_n^{\max}$ denote the size of the largest tree of $\mathscr{F}_n$. Then,*

$$T_n^{\max} = \alpha\big(\log n - (1+o_{\mathrm{p}}(1))\log\log n\big),$$

where $\alpha = (1-\log(e-1))^{-1} \approx 2.18019$.

3.2 Sampling of the stationary distribution

3.2.1 Backward construction

Consider an i.i.d. sequence $((V_t, W_t),\, t \in \mathbb{Z})$, where (V_t, W_t) is uniformly distributed on the set of ordered pairs of distinct elements of $\{1, \dots, n\}$. These variables are meant to encode the transitions of the chain: W_t represents the vertex that is disconnected at step t, and V_t the vertex to which W_t is then connected. We now explain how to construct a chain $(\mathscr{F}_n(t),\, t \in \mathbb{Z})$ of forests from the sequence $((V_t, W_t),\, t \in \mathbb{Z})$, by looking at it backwards in time.

Fix a focal time $t \in \mathbb{Z}$. For each vertex v, let us denote by

$$\tau_t(v) \coloneqq \max\{s \leqslant t : W_s = v\}$$

the last time before t when v was chosen to be disconnected, and define

$$m_t(v) \coloneqq V_{\tau_t(v)}$$

to be the vertex to which it was then reconnected. We refer to the time $\tau_t(v)$ as the *birth time* of v, and to the vertex $m_t(v)$ as its *mother*. Note that the variables $(\tau(v),\, 1 \leqslant v \leqslant n)$ are independent of $(m(v),\, 1 \leqslant v \leqslant n)$.

Now, for each $s \leqslant t$, let the vertices be in one of two states, *active* or *inactive*, as follows: vertex v is active at times s such that $\tau_t(v) \leqslant s \leqslant t$, and inactive at times $s < \tau_t(v)$. Finally, let $\mathscr{F}_n(t)$ be the forest obtained by connecting each vertex v to its mother if the mother is active at the time of birth of v, that is,

$$v \text{ is connected to } m_t(v) \iff \tau_t(m_t(v)) < \tau_t(v).$$

This procedure is illustrated in Figure 3.2.

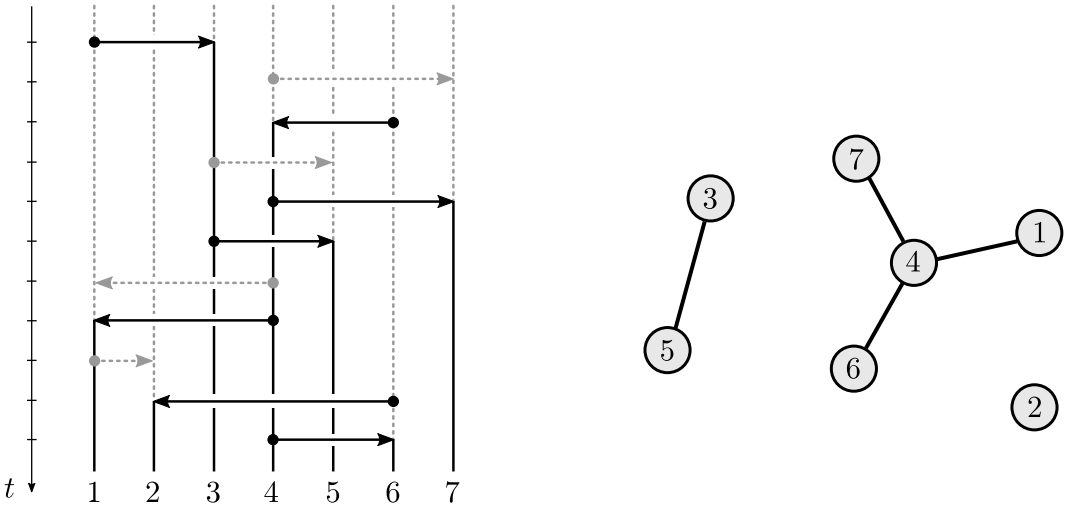


Figure 3.2: Illustration of the backward construction. Each vertex corresponds to a vertical line. A pair (V_t, W_t) is represented by an arrow $V_t \to W_t$. The line representing a vertex is solid black when that vertex is active, and dashed grey when it is inactive. Arrows pointing to inactive vertices are represented in dashed grey because they have no impact on the state of the graph at the focal time: their effect has been erased by subsequent arrows.

Let us show that the chain $(\mathscr{F}_n(t),\, t \in \mathbb{Z})$ has the same transitions as the chain described in the introduction. First, note that for $v \neq W_t$ we have $\tau_t(v) = \tau_{t-1}(v)$, and thus $m_t(v) = m_{t-1}(v)$. As a result, edges that do not involve W_t are the same

in $\mathscr{F}_n(t)$ and in $\mathscr{F}_n(t-1)$. Now, $\tau_t(W_t) = t$, so that W_t is always inactive as a mother in the construction of $\mathscr{F}_n(t)$, and $m_t(W_t) = V_t$ with $\tau_t(V_t) < t$, so that W_t is linked to V_t in $\mathscr{F}_n(t)$. In other words, $\mathscr{F}_n(t)$ is obtained from $\mathscr{F}_n(t-1)$ by disconnecting W_t from its neighbors, and then connecting it to V_t. This corresponds to the transitions of the chain described in the introduction.

Finally, $(\mathscr{F}_n(t),\, t \in \mathbb{Z})$ is stationary by construction, and thus $\mathscr{F}_n(t)$ is distributed as the Moran forest for all time $t \in \mathbb{Z}$.

3.2.2 Uniform attachment construction

We now give a forward-in-time variant of the construction described in the previous section. This forward-in-time procedure, which we call the *uniform attachment construction* (UA construction for short), is our main tool to study $\mathscr{F}_n$ and will be used throughout the rest of the paper.

Let $(U_n(\ell),\, 1 \leqslant \ell \leqslant n)$ be a vector of independent variables such that $U_n(\ell)$ is uniformly distributed on $\{1, \dots, n\} \setminus \{\ell\}$. Consider the forest $\mathscr{F}_n^*$ on $\{1, \dots, n\}$ obtained by setting

$$k \text{ is connected to } \ell, \text{ with } k < \ell \iff U_n(\ell) = k.$$

We will show that, after relabeling the vertices of $\mathscr{F}_n^*$ according to a uniform permutation of $\{1, \dots, n\}$, we obtain the Moran forest. Before this let us make a few remarks.

First, it will be helpful to think of the construction of $\mathscr{F}_n^*$ as a sequential process where, starting from a single vertex labeled 1, for $\ell = 2, \dots, n$ we add a new vertex labeled ℓ and connect it to $U_n(\ell)$ if $U_n(\ell) < \ell$. See Figure 3.3. This will make the link with some well-known stochastic processes more intuitive. This also explains that we speak of the *ℓ-th vertex in the UA construction* to refer to vertex ℓ in $\mathscr{F}_n^*$.

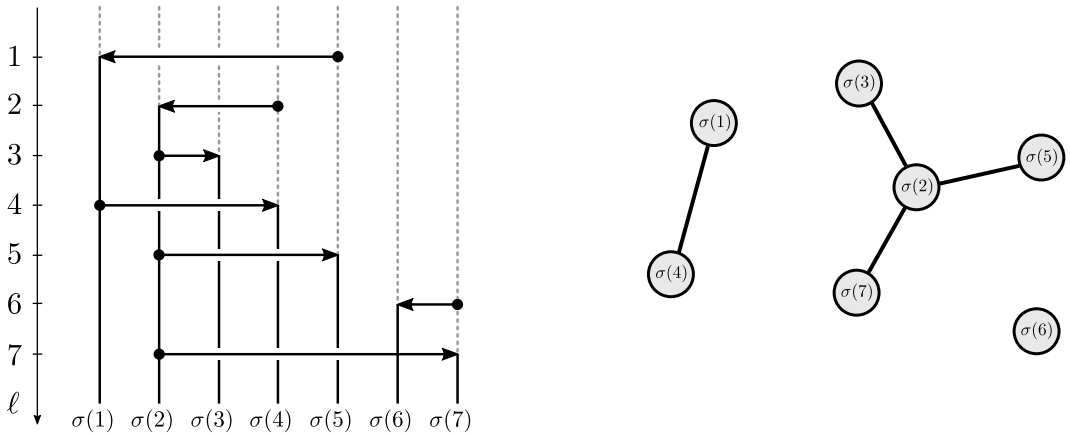


Figure 3.3: Illustration of the uniform attachment construction for $n = 7$ and the vector $(U_n(1), \dots, U_n(n)) = (5, 4, 2, 1, 2, 7, 2)$. The ℓ-th vertical line from the left corresponds to vertex $\sigma(\ell)$ (i.e, in the sequential vision, to the ℓ-th vertex that is added). $U_n(\ell)$ is represented by the arrow pointing from the $U_n(\ell)$-th line to the ℓ-th one at time ℓ. Compare this with Figure 3.2: the vertical lines corresponding to the vertices have been reordered in increasing order of their birth time, and the grey arrow that left no trace on the graph at the focal time has been removed.

Second, the edges of $\mathscr{F}_n^*$ are by construction increasing, in the sense that if we root every tree of $\mathscr{F}_n^*$ by letting the root of a tree be its smallest vertex, then each edge $u\vec{v}$ pointing away from the root of its tree is such that $u < v$.

Rooted trees that have only increasing edges are known as *recursive trees* [8], and forests of recursive trees have been called *recursive forests* [4]. Recursive trees have been studied extensively [5, 16, 15]. In particular, the uniform attachment tree, which corresponds to the uniform distribution over the set of recursive trees, has received much attention, see [8] for an overview. However, the random forest $\mathscr{F}_n^*$ does not seem to correspond to any previously studied model of random recursive forest (in particular, it is not uniformly distributed over the set of recursive forests).

Proposition 3.2.1. *The random forest obtained by relabeling the vertices of $\mathscr{F}_n^*$ according to a uniform permutation of $\{1,\dots,n\}$ is distributed as the Moran forest.*

Proof. Consider the forest $\mathscr{F}_n(0)$ built from the variables $((V_t, W_t),\, t \in \mathbb{Z})$ in the previous section. To ease notation, we will omit the subscript in τ_0 and m_0.

Let us relabel the vertices in increasing order of their birth time: since the variables $(\tau(v),\, 1 \leqslant v \leqslant n)$ are all distinct, there exists a unique permutation σ of $\{1,\dots,n\}$ such that

$$\tau(\sigma(1)) < \cdots < \tau(\sigma(n)).$$

In words, $\sigma(\ell)$ is the ℓ-th vertex that was born in the construction of $\mathscr{F}_n(0)$. Using the new labeling, let us denote its birth time by $\tau^*(\ell) = \tau(\sigma(\ell))$ and its mother by $m^*(\ell) = \sigma^{-1}(m(\sigma(\ell)))$.

Now, for every vertex $v = \sigma(\ell)$,

$$\begin{aligned} v \text{ is connected to } m(v) \text{ in } \mathscr{F}_n(0) &\iff \tau(m(v)) < \tau(v) \\ &\iff \tau^*(m^*(\ell)) < \tau^*(\ell) \\ &\iff m^*(\ell) < \ell\,. \end{aligned}$$

Thus, if we set $U_n(\ell) = m^*(\ell)$ in the construction of $\mathscr{F}_n^*$ then ℓ is connected to $m^*(\ell)$ if and only if $v = \sigma(\ell)$ is connected to $m(v) = \sigma(m^*(\ell))$ in $\mathscr{F}_n(0)$. Therefore, to finish the proof we have to show that:

(i) The variables $(m^*(\ell),\, 1 \leqslant \ell \leqslant n)$ are independent and such that $m^*(\ell)$ is uniformly distributed on $\{1,\dots,n\} \setminus \{\ell\}$.

(ii) The permutation σ is uniform and independent of $(m^*(\ell),\, 1 \leqslant \ell \leqslant n)$.

First, note that by construction the variables $(m(v),\, 1 \leqslant v \leqslant n)$ clearly satisfy the analog of the first point above, i.e. that those variables are independent and that for each v, $m(v)$ is uniformly distributed on $\{1,\dots,n\} \setminus \{v\}$. Since the permutation σ depends only on the variables $(\tau(v),\, 1 \leqslant v \leqslant n)$, which are independent of $(m(v),\, 1 \leqslant v \leqslant n)$, we see that σ is independent of $(m(v),\, 1 \leqslant v \leqslant n)$. Moreover, the variables $(\tau(v),\, 1 \leqslant v \leqslant n)$ are exchangeable so the permutation σ is uniform. Now, for any fixed permutation π of $\{1,\dots n\}$ and any fixed map $f : \{1,\dots,n\} \to \{1,\dots,n\}$ such that $f(\ell) \neq \ell$ for all ℓ,

$$\begin{aligned} \mathbb{P}(\sigma = \pi,\, m^* = f) &= \mathbb{P}\Big(\sigma = \pi,\, m = \pi \circ f \circ \pi^{-1}\Big) \\ &= \frac{1}{n!}\frac{1}{(n-1)^n}\,, \end{aligned}$$

concluding the proof. □

3.3 Number of trees

3.3.1 Law of the number of trees

In the UA construction, let $I_\ell = \mathbb{1}_{\{U_n(\ell)<\ell\}}$ be the indicator variable of the event "the ℓ-th vertex was linked to a previously added vertex". The variables $(I_1, \dots, I_n)$ are thus independent Bernoulli variables such that

$$I_\ell \;\sim\; \text{Bernoulli}\Big(\tfrac{\ell-1}{n-1}\Big)\,.$$

With this notation, the number of edges $|E_n|$ and the number of trees N_n are

$$|E_n| = \sum_{\ell=1}^{n} I_\ell \quad \text{and} \quad N_n = \sum_{\ell=1}^{n} (1 - I_\ell)\,.$$

Moreover, since $I_\ell \overset{d}{=} 1 - I_{n-\ell+1}$, we see that

$$\mathbb{P}(N_n = k) \;=\; \mathbb{P}(N_n = n-k) \;=\; \mathbb{P}(|E_n| = k)\,,$$

that is, the number of trees and the number of edges have the same, symmetric distribution. In consequence, from now on we only use the notation N_n and refer to it as the number of trees of $\mathscr{F}_n$ when stating our results—even though we sometimes work with the number of edges in the proofs.

From the representation of N_n as a sum of independent Bernoulli variables, we immediately get the following result.

Proposition 3.3.1. *Let N_n denote the number of trees of $\mathscr{F}_n$.*

(i) $\mathbb{E}(N_n) = \dfrac{n}{2}$

(ii) $\mathrm{Var}(N_n) = \dfrac{n(n-2)}{6(n-1)}$.

(iii) $G_{N_n}(z) \coloneqq \mathbb{E}\Big(z^{N_n}\Big) = \displaystyle\prod_{k=1}^{n-1} \Big(1 + \tfrac{k}{n-1}(z-1)\Big)$.

The representation of N_n as a sum of independent Bernoulli variables also makes it straightforward to get the following central limit theorem.

Proposition 3.3.2. *Let N_n denote the number of trees of $\mathscr{F}_n$. Then,*

$$\frac{N_n - n/2}{\sqrt{n/6}} \;\xrightarrow[n\to\infty]{d}\; \mathcal{N}(0,1)\,.$$

Proof. This is an immediate consequence of the Lyapunov CLT for triangular arrays of independent random variables. Indeed, $\mathbb{E}\Big(|I_\ell - \mathbb{E}(I_\ell)|^3\Big) \leqslant 1$. Therefore,

$$\frac{1}{n^{3/2}} \sum_{\ell=1}^{n} \mathbb{E}\Big(|I_\ell - \mathbb{E}(I_\ell)|^3\Big) \leqslant \frac{1}{\sqrt{n}} \;\xrightarrow[n\to\infty]{}\; 0\,,$$

and the result follows, e.g., from Corollary 11.1.4 in [3]. □

3.3.2 Link with uniform labeled trees

As announced in the introduction, there is a strong connection between the Moran forest and uniform labeled trees. Our starting point is the following observation about the probability generating function of N_n. First,

$$\begin{aligned}\sum_{k\geqslant 0} \mathbb{P}(N_n = k)\, z^k &= \prod_{k=1}^{n-1} \left(1 + \tfrac{k}{n-1}(z-1)\right) \\ &= \frac{z}{(n-1)^{n-2}} \prod_{k=1}^{n-2} (n-1-k+kz) \\ &= \sum_{k=0}^{n-2} \frac{a(n-1,k)}{(n-1)^{n-2}}\, z^{k+1},\end{aligned}$$

where

$$\sum_{k=0}^{n-2} a(n-1,k)\, z^k = \prod_{k=1}^{n-2} (n-1-k+kz).$$

Second, the coefficients of this polynomial have a simple combinatorial interpretation: $a(n-1,k)$ is the number of rooted trees on $\{1,\dots,n-1\}$ with k increasing edges, where an edge $u\vec{v}$ pointing away from the root is said to be increasing if $u < v$. This fact is known in the literature as a consequence of the more general Theorem 1.1 of [10] (see also Example 1.7.2 in [7] and Theorem 9.1 in [12]).

This simple observation already gives us the following proposition.

Proposition 3.3.3. *The probability mass function of the number of trees of $\mathscr{F}_n$ is*

$$\mathbb{P}(N_n = k) = \frac{a(n-1,k-1)}{(n-1)^{n-2}},$$

where $a(n,k)$ is the number of rooted trees on $\{1,\dots n\}$ with k increasing edges (sequence A067948 of the On-Line Encyclopedia of Integer Sequences [1]).

Looking for a bijective proof of Proposition 3.3.3 naturally leads to the following more general result about the link between the Moran forest and uniform rooted labeled trees.

Theorem 3.3.4. *Let $\mathcal{T}$ be a uniform rooted tree on $\{1,\dots,n-1\}$. From this tree, build a forest $\mathcal{F}$ on $\{1,\dots,n\}$ according to the following procedure:*

1. *Remove all decreasing edges from $\mathcal{T}$ (that is, edges $u\vec{v}$ pointing away from the root such that $u > v$).*
2. *Add a vertex labeled n and connect it to a uniformly chosen vertex of $\mathcal{T}$*
3. *Relabel vertices according to a uniform permutation of $\{1,\dots,n\}$.*

Then, the resulting forest $\mathcal{F}$ has the law of the Moran forest $\mathscr{F}_n$.

Proof. In the UA construction, let $F_{|n-1}$ denote the forest obtained after the addition of $n-1$ vertices, before their relabeling. After this, the n-th vertex will be linked to a uniformly chosen vertex of $F_{|n-1}$. As a result, to prove the theorem it suffices to show that $F_{|n-1}$ has the same law as the forest obtained from $\mathcal{T}$ by removing its decreasing edges.

To do so, we couple $F_{|n-1}$ and $\mathcal{T}$ in such a way that the edges of $F_{|n-1}$ are exactly the increasing edges of $\mathcal{T}$. Formally, $F_{|n-1}$ is a deterministic function of the random vector $\mathbf{U} = (U_n(2), \dots, U_n(n-1))$. Moreover, $\mathbf{U}$ is uniform on the set

$$\mathscr{S}^{\star}_{n-1} \;=\; \Big\{\mathbf{u} \in \{1,\dots,n\}^{\{2,\dots,n-1\}} \,:\, u_\ell \neq \ell\Big\}\,.$$

Thus, to end the proof it is sufficient to find a bijection Φ from $\mathscr{S}^{\star}_{n-1}$ to the set of rooted trees on $\{1,\dots,n-1\}$ and such that

$$k\ell \in F_{|n-1}(\mathbf{u}) \iff k\ell \text{ is an increasing edge of } \Phi(\mathbf{u})\,.$$

First, let

$$\mathscr{S}_{n-1} \;=\; \{1,\dots,n-1\}^{\{2,\dots,n-1\}}$$

and consider the bijection $\Theta : \mathscr{S}^{\star}_{n-1} \to \mathscr{S}_{n-1}$ defined by

$$\Theta\mathbf{u} : \ell \mapsto u_\ell - \mathbb{1}_{\{u_\ell > \ell\}}\,.$$

Importantly, note that Θ does not modify the entries of $\mathbf{u}$ that correspond to edges of $F_{|n-1}(\mathbf{u})$, that is, for all $k < \ell$,

$$k\ell \in F_{|n-1}(\mathbf{u}) \iff u_\ell = k \iff (\Theta\mathbf{u})(\ell) = k\,.$$

As a result, it remains to find a bijection Ψ from $\mathscr{S}_{n-1}$ to the set of rooted trees on $\{1,\dots,n-1\}$ such that

$$u_\ell < \ell \iff u_\ell \text{ and } \ell \text{ are linked by an increasing edge in } \Psi(\mathbf{u})\,.$$

This bijection will essentially be that used in [10], which can itself be seen as a variant of Joyal's bijection [14, 2].

Let $\mathscr{G}_{\mathbf{u}}$ be the directed graph on $\{1,\dots,n-1\}$ obtained by putting a directed edge going from u_ℓ to ℓ for all $\ell \geqslant 2$.

If $\mathscr{G}_{\mathbf{u}}$ has no cycle or self-loop, then it is a tree. Moreover, the orientation of its edges uniquely identify vertex 1 as its root. Thus we set $\Psi(\mathbf{u}) = \mathscr{G}_{\mathbf{u}}$.

If $\mathscr{G}_{\mathbf{u}}$ is not a tree, set $\mathscr{C}_0 = \{1\}$ and let $\mathscr{C}_1, \dots, \mathscr{C}_k$ denote the cycles of $\mathscr{G}_{\mathbf{u}}$, taken in increasing order of their largest element and treating self-loops as cycles of length 1. Note that because each vertex has exactly one incoming edge, except for vertex 1 which has none, these cycles are vertex-disjoint and directed.

To turn $\mathscr{G}_{\mathbf{u}}$ into a tree, set $s_0 = 1$ and for $i \geqslant 1$ let m_i denote the largest element of $\mathscr{C}_i$ and $m_i\vec{s}_i$ its out-going edge in $\mathscr{C}_i$. With this notation, for $i = 1, \dots, k$ remove the edge $m_i\vec{s}_i$ from $\mathscr{G}_{\mathbf{u}}$ and replace it by $m_i\vec{s}_{i-1}$. Note that

- This turns $\mathscr{C}_0 \sqcup \cdots \sqcup \mathscr{C}_k$ into a directed path $\mathscr{P}$ going from s_k to 1.
- Because $m_i = \max \mathscr{C}_i$ and that $1 < m_1 < \cdots < m_k$, every edge $m_i\vec{s}_i$ was non-increasing and has been replaced by the decreasing edge $m_i\vec{s}_{i-1}$.

Therefore, this procedure turns $\mathscr{G}_{\mathbf{u}}$ into a tree $\Psi(\mathbf{u})$ rooted in s_k, without modifying its increasing edges. Consequently, the increasing edges of $\Psi(\mathbf{u})$ are exactly the pairs $k\ell$ for which $k = u_\ell < \ell$.

To see that Ψ is a bijection, it suffices to note that the cycles $\mathscr{C}_0, \dots, \mathscr{C}_m$ can be recovered unambiguously from the path $\mathscr{P}$ going from the root to vertex 1. Indeed, writing this path as the word $1m_1 \cdots s_1 m_2 \cdots s_k$, the m_i are exactly the left-to-right maxima of that word.

Setting $\Phi = \Psi \circ \Theta$ thus gives us the bijection that we were looking for, concluding the proof. □

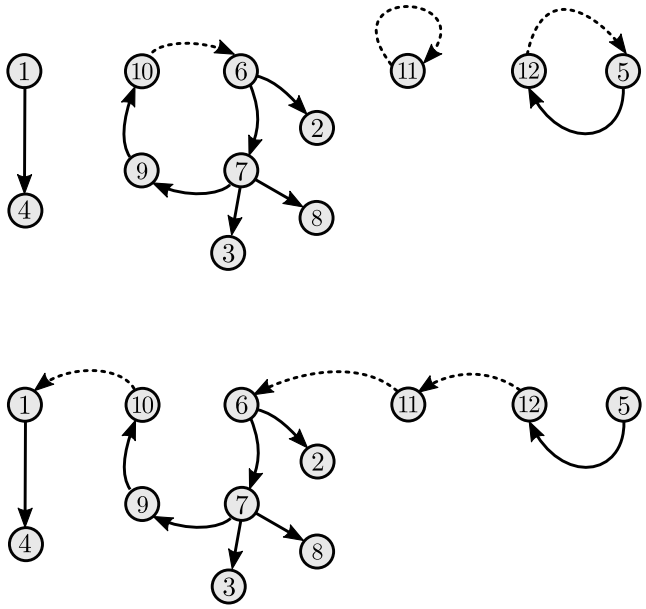


Figure 3.4: Example of construction of $\Phi(\mathbf{u})$, for $\mathbf{u} = (7, 8, 1, 13, 11, 6, 7, 7, 9, 12, 5)$. Applying Θ yields $\mathbf{u}' = \Theta\mathbf{u} = (6, 7, 1, 12, 10, 6, 7, 7, 9, 11, 5)$. The directed graph $\mathscr{G}_{\mathbf{u}'}$ encoding $\mathbf{u}'$ is represented on top. Its cycles are $\mathscr{C}_1 = (10, 6, 7, 9)$, $\mathscr{C}_2 = (11)$ and $\mathscr{C}_3 = (12, 5)$, and we set $\mathscr{C}_0 = (1)$. The edges $m_i\vec{s}_i$ are dashed. Rewiring them as described in the main text turns $\mathscr{G}_{\mathbf{u}'}$ into the rooted tree $\Psi(\mathbf{u}')$ represented on bottom. No information is lost when turning the cycles $(1)(10, 6, 7, 9)(11)(12, 5)$ into the path going from 5 to 1 encoded by the word $(1, 10, 6, 7, 9, 11, 12, 5)$, because the left-to-right maxima of that word—here 1, 10, 11 and 12—each mark the start of a new cycle.

3.4 Degrees

3.4.1 Degree of a fixed vertex

Using the UA construction and the notation from Section 3.2.2, let us denote by

- $I_\ell = \mathbb{1}_{\{U_n(\ell)<\ell\}}$ the indicator variable of the event "the ℓ-th vertex is linked to a previously added vertex".
- $X_\ell^{(v)} = \mathbb{1}_{\{U_n(\ell)=\sigma^{-1}(v)\}}$ the indicator variable of the event "the ℓ-th vertex is linked to vertex v".
- $B_v = \sigma^{-1}(v)$ the step of the construction at which vertex v is added.

With this notation, the degree of vertex v is

$$D_n^{(v)} \;=\; I_{B_v} \;+\; \sum_{\ell=B_v+1}^{n} X_\ell^{(v)}.$$

Moreover, conditional on $\{B_v = b\}$, $(X_{b+1}^{(v)}, \dots, X_n^{(v)})$ are i.i.d. Bernoulli variables with parameter $1/(n-1)$ and I_b is a Bernoulli variable with parameter $\frac{b-1}{n-1}$ that is independent of $(X_{b+1}^{(v)}, \dots, X_n^{(v)})$. As a result, conditional on B_v and writing L_v for $n - B_v$,

$$D_n^{(v)} \;\overset{d}{=}\; \mathrm{Ber}\Big(1 - \tfrac{L_v}{n-1}\Big) \;+\; \mathrm{Bin}\Big(L_v,\, \tfrac{1}{n-1}\Big)\,,$$

where the Bernoulli and the binomial variables are independent conditional on L_v. Using that L_v is uniformly distributed on $\{0, \dots, n-1\}$, the mean, variance and probability generating function of $D_n^{(v)}$ are obtained by routine calculations.

Proposition 3.4.1. *Let D_n be the degree of a fixed vertex of $\mathscr{F}_n$. Then,*

(i) $\mathbb{E}(D_n) = 1$.

(ii) $\mathrm{Var}(D_n) = \dfrac{2(n-2)}{3(n-1)}$.

(iii) $G_{D_n}(z) \coloneqq \mathbb{E}(z^{D_n}) = \dfrac{1}{n}\displaystyle\sum_{\ell=0}^{n-1}\Big(1 + (1 - \tfrac{\ell}{n-1})(z-1)\Big)\Big(1 + \tfrac{1}{n-1}(z-1)\Big)^{\ell}$.

(iii') $G_{D_n}(z) = 2\Big(1 - \dfrac{1}{n}\Big)\dfrac{\Big(1 + \frac{z-1}{n-1}\Big)^n - 1}{z \,-\, 1} - 1$.

Remark 3.4.2. Note that we also have $\mathbb{E}(D_n^{(v)} \mid L_v) = 1$, that is, the average degree of a vertex is independent of the step at which it was added in the UA construction. ◇

Proposition 3.4.3. *The degree D_n of a fixed vertex of $\mathscr{F}_n$ converges in distribution to the variable D satisfying:*

(i) $D \sim \mathrm{Ber}(1-U) + \mathrm{Poisson}(U)$, *where U is uniform on $[0,1]$ and the Bernoulli and Poisson variables are independent conditional on U.*

(ii) $G_D(z) \coloneqq \mathbb{E}\Big(z^D\Big) = \displaystyle\int_0^1 (1 + (1-x)(z-1))e^{x(z-1)}dx = 2\,\frac{e^{z-1} - 1}{z-1} - 1$.

(iii) *For all $p \geqslant 1$,* $\mathbb{E}\big(D(D-1)\cdots(D-p+1)\big) = \dfrac{2}{p+1}$.

(iv) $\mathbb{P}(D = 0) = 1 - 2/e$ *and, for $k \geqslant 1$,*

$$\mathbb{P}(D = k) \;=\; \frac{2}{e}\sum_{j>k}\frac{1}{j!}\,.$$

Proof. First, for all $z \in \mathbb{C} \setminus \{1\}$,

$$G_{D_n}(z) \;=\; 2\Big(1 - \frac{1}{n}\Big)\frac{\Big(1 + \frac{z-1}{n-1}\Big)^n - 1}{z \,-\, 1} - 1 \;\xrightarrow[n\to\infty]{}\; 2\,\frac{e^{z-1} - 1}{z-1} - 1.$$

This pointwise convergence of the probability generating function of D_n proves convergence in distribution of D_n to a random variable D satisfying (ii). Point (i) then follows immediately from the integral expression of G_D.

To compute the factorial moments of D, note that

$$G_D(z) \;=\; 2\sum_{k\geqslant 0}\frac{(z-1)^k}{(k+1)!} \;-\; 1.$$

As a result, for $p \geqslant 1$ the p-th derivative of G_D is

$$G_D^{(p)}(z) \;=\; 2\sum_{k\geqslant 0}\frac{(z-1)^k}{(k+1+p)k!}\,,$$

and, in particular, $\mathbb{E}(D(D-1)\cdots(D-p+1)) = G_D^{(p)}(1) = \frac{2}{p+1}$, proving (iii).

Finally, to prove (iv), using (i) we see that

$$\mathbb{P}(D=0) \;=\; \int_0^1 x e^{-x}\,dx \;=\; 1-\frac{2}{e}$$

and that, for $k \geqslant 1$,

$$\mathbb{P}(D=k) \;=\; \frac{1}{k!}\int_0^1 \Big(kx^{k-1}-kx^k+x^{k+1}\Big)\,e^{-x}\,dx.$$

Noting that $\big(kx^{k-1}-kx^k+x^{k+1}\big)e^{-x} = 2x^k e^{-x} + \frac{d}{dx}\Big((x^k-x^{k+1})e^{-x}\Big)$, one gets

$$\mathbb{P}(D=k) = \frac{2}{k!}\int_0^1 x^k e^{-x}\,dx,$$

and an easy integration by parts yields

$$\mathbb{P}(D=k+1) \;=\; \mathbb{P}(D=k) \;-\; \frac{2}{e(k+1)!}\,,$$

from which (iv) follows by induction. □

Before closing this section, let us give an asymptotic equivalent of the tail of D_n. We will need it in the proof of Theorem 3.4.5 on the largest degree.

Proposition 3.4.4. *Let D_n be the degree of a fixed vertex of $\mathcal{F}_n$ and let D have the asymptotic distribution of D_n.*

(i) *For all $k \geqslant 1$,*

$$\frac{2/e}{(k+1)!} \;\leqslant\; \mathbb{P}(D \geqslant k) \;\leqslant\; \Big(1+\frac{1}{k}\Big)^2 \frac{2/e}{(k+1)!}\,.$$

(ii) *For all $K_n = o(\sqrt{n})$, there exists $\varepsilon_n = o(1)$ such that, for all $k \leqslant K_n$,*

$$|\mathbb{P}(D_n \geqslant k) - \mathbb{P}(D \geqslant k)| \;\leqslant\; \varepsilon_n\,\mathbb{P}(D \geqslant k)\,.$$

(iii) *For all $k_n \to +\infty$ and $K_n \geqslant k_n$ such that $K_n = o(\sqrt{n})$,*

$$\mathbb{P}(D_n \geqslant k) \;\sim\; \frac{2/e}{(k+1)!}\,,$$

uniformly in k such that $k_n \leqslant k \leqslant K_n$.

Proof. First, observe that

$$\frac{1}{(\ell+1)!} \;\leqslant\; \frac{1}{\ell\cdot\ell!} - \frac{1}{(\ell+1)\cdot(\ell+1)!}\,,$$

so that

$$\sum_{\ell>i}\frac{1}{\ell!} \;\leqslant\; \frac{1}{i\cdot i!} \;=\; \Big(1+\frac{1}{i}\Big)\frac{1}{(i+1)!}\,.$$

Recalling from Proposition 3.4.3 that

$$\mathbb{P}(D \geqslant k) \;=\; \frac{2}{e}\sum_{i\geqslant k}\sum_{\ell>i}\frac{1}{\ell!}\,,$$

point (i) follows readily.

The proof of (ii) is somewhat technical so we only outline it here and refer the reader to Section 3.A of the Appendix for the detailed calculations.

Consider the function

$$\Delta_n(z) \;=\; \sum_{i\geqslant 0} \big(\mathbb{P}(D \geqslant i) - \mathbb{P}(D_n \geqslant i)\big) z^i.$$

With this function, (ii) can be re-expressed as

$$\Delta_n^{(k)}(0) \;=\; \frac{\varepsilon_n}{k+1} \quad \text{for all } k \leqslant K_n = o(\sqrt{n}),$$

where $\Delta_n^{(k)}$ denotes the k-th derivative of Δ_n. But Δ_n can be expressed in terms of the generating functions of D and D_n, namely as

$$\Delta_n(z) \;=\; \left(1 + \frac{1}{z-1}\right)\big(G_D(z) - G_{D_n}(z)\big).$$

The expressions of G_D and G_{D_n} obtained in Propositions 3.4.1 and 3.4.3 thus make it straightforward to obtain a power series expansion of Δ_n at $z = 1$, and this expansion can be used to bound $\Delta_n^{(k)}(0)$ and conclude the proof.

Finally, (iii) is a direct consequence of (i) and (ii). □

3.4.2 Largest degree

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

Theorem 3.4.5. *Let $D_n^{\max} = \max_v D_n^{(v)}$ denote the largest degree of $\mathcal{F}_n$. Then,*

$$D_n^{\max} = \frac{\log n}{\log\log n} \;+\; (1 + o_{\mathrm{p}}(1))\,\frac{\log n \,\log\log\log n}{(\log\log n)^2},$$

where $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

Our proof uses a first and second moment method which will also be used in the proof of Theorem 3.5.8 concerning the size of the largest tree. In order to avoid repeating ourselves, we isolate this classic part of our reasoning as a lemma, whose proof we recall for the sake of completeness.

Lemma 3.4.6. *For all integers n, let $(X_n^{(1)}, \dots, X_n^{(n)})$ be a vector of exchangeable random variables and*

$$X_n^{\max} = \max\Big\{X_n^{(i)} : i = 1, \dots, n\Big\}.$$

Write $p_n(k)$ for $\mathbb{P}(X_n^{(i)} \geqslant k)$, and suppose that there exists a sequence (m_n) and a constant β such that, for all $\varepsilon > 0$, as $n \to \infty$,

(i) $np_n((\beta+\varepsilon)m_n) \to 0$.

(ii) $np_n((\beta-\varepsilon)m_n) \to +\infty$.

(iii) $\mathbb{P}\big(X_n^{(1)} \geqslant (\beta-\varepsilon)m_n,\, X_n^{(2)} \geqslant (\beta-\varepsilon)m_n\big) \sim p_n((\beta-\varepsilon)m_n)^2$.

Then for all $\varepsilon > 0$,

$$\mathbb{P}\big(X_n^{\max} \geqslant (\beta+\varepsilon)\, m_n\big) \to 0 \quad \textit{and} \quad \mathbb{P}\big(X_n^{\max} \geqslant (\beta-\varepsilon) m_n\big) \to 1,$$

which can also be written

$$X_n^{\max} = \big(\beta + o_{\mathrm{p}}(1)\big)\, m_n,$$

where $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

Proof. First,

$$\begin{aligned}\mathbb{P}\big(X_n^{\max} \geqslant (\beta+\varepsilon) m_n\big) &= \mathbb{P}\left(\bigcup_{i=1}^{n} \Big\{X_n^{(i)} \geqslant (\beta+\varepsilon) m_n\Big\}\right) \\ &\leqslant n p_n\big((\beta+\varepsilon) m_n\big)\,,\end{aligned}$$

which goes to zero by (i). Now, denote by

$$Z_n = \sum_{i=1}^{n} \mathbb{1}_{\{X_n^{(i)} \geqslant (\beta-\varepsilon) m_n\}}$$

the number of variables $X_n^{(i)}$ that are greater than or equal to $(\beta-\varepsilon)m_n$. By the Cauchy–Schwartz inequality,

$$\mathbb{P}\big(X_n^{\max} \geqslant (\beta-\varepsilon) m_n\big) \;=\; \mathbb{P}(Z_n > 0) \;\geqslant\; \frac{\mathbb{E}(Z_n)^2}{\mathbb{E}(Z_n^2)}\,.$$

Moreover,

$$\begin{aligned}\mathbb{E}\Big(Z_n^2\Big) \;&=\; n p_n\big((\beta-\varepsilon) m_n\big) \\ &\quad+\; n(n-1)\mathbb{P}\Big(X_n^{(1)} \geqslant (\beta-\varepsilon) m_n,\, X_n^{(2)} \geqslant (\beta-\varepsilon) m_n\Big)\,,\end{aligned}$$

and so, by (ii) and (iii), $\mathbb{E}(Z_n)^2/\mathbb{E}(Z_n^2) \to 1$ as $n \to \infty$. □

Remark 3.4.7. Note that under assumption (ii) of this lemma, for any $\varepsilon > 0$, letting $n \to \infty$ in $\mathbb{E}(Z_n)^2/\mathbb{E}(Z_n^2) \leqslant 1$ shows that

$$p_n\big((\beta-\varepsilon) m_n\big)^2 \;\leqslant\; \mathbb{P}\Big(X_n^{(1)} \geqslant (\beta-\varepsilon) m_n,\, X_n^{(2)} \geqslant (\beta-\varepsilon) m_n\Big)\big(1+o(1)\big).$$

Therefore, to prove (iii) it suffices to show

$$\text{(iii')}\;\; \mathbb{P}\Big(X_n^{(1)} \geqslant (\beta-\varepsilon) m_n,\, X_n^{(2)} \geqslant (\beta-\varepsilon) m_n\Big) \;\leqslant\; p_n\big((\beta-\varepsilon) m_n\big)^2\big(1+o(1)\big). \qquad \diamond$$

We now turn to the proof of Theorem 3.4.5.

Proof of Theorem 3.4.5. Instead of proving the theorem directly for the variables $(D_n^{(1)}, \dots, D_n^{(n)})$, we prove it for some auxiliary variables $(\tilde{D}_n^{(1)}, \dots, \tilde{D}_n^{(n)})$ whose maximum has the same asymptotic behavior as $D_n^{\max}$. The point in doing this is that the tails of the variables $\tilde{D}_n^{(v)}$ are less correlated than those of the variables $D_n^{(v)}$, which makes it easier to study their maximum by the first and second moment method.

Remember from Section 3.4.1 that, in the UA construction,

$$D_n^{(v)} \;=\; I_{B_v} \;+\; \sum_{\ell=B_v+1}^{n} X_\ell^{(v)},$$

where B_v is the step at which vertex v was added, $X_\ell^{(v)}$ is the indicator of "the ℓ-th vertex is linked to vertex v", and I_ℓ is the indicator of "the ℓ-th vertex is linked to a previously added vertex". With this notation, set

$$\tilde{D}_n^{(v)} = \sum_{\ell=B_v+1}^{n} X_\ell^{(v)}$$

and $\tilde{D}_n^{\max} = \max\{\tilde{D}_n^{(v)} : v = 1, \dots, n\}$. Since $\tilde{D}_n^{\max}$ and $D_n^{\max}$ differ by at most 1, for any $m_n \to +\infty$,

$$D_n^{\max} - \tilde{D}_n^{\max} \;=\; o_{\mathrm{p}}(m_n)\,,$$

i.e. $(D_n^{\max} - \tilde{D}_n^{\max})/m_n$ goes to 0 in probability. Thus, to prove the theorem we apply Lemma 3.4.6 to the variables

$$\left(\tilde{D}_n^{(1)} - \frac{\log n}{\log\log n}, \dots,\; \tilde{D}_n^{(n)} - \frac{\log n}{\log\log n}\right),$$

with $m_n = (\log n)(\log\log\log n)/(\log\log n)^2$ and $\beta = 1$.

Using Proposition 3.4.4 and Stirling's formula, we see that for any $k_n = o(\sqrt{n})$,

$$\log\left(\mathbb{P}(D_n \geqslant k_n)\right) \;=\; -k_n \log k_n \;+\; k_n \;+\; O(\log k_n)\,.$$

Writing $\tilde{D}_n$ to refer to the common distribution of the the variables $\tilde{D}_n^{(v)}$, since

$$\mathbb{P}(D_n \geqslant k_n + 1) \;\leqslant\; \mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big) \;\leqslant\; \mathbb{P}(D_n \geqslant k_n)\,,$$

we also have

$$\log\left(\mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)\right) \;=\; -k_n \log k_n \;+\; k_n \;+\; O(\log k_n)\,.$$

In particular, for $k_n = (\log n)/(\log\log n) + \gamma\, m_n$ with

$$m_n = \frac{\log n\;\log\log\log n}{(\log\log n)^2}\,,$$

this gives

$$\log\left(\mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)\right) \;=\; -\log n \;-\; (\gamma - 1)\,\frac{\log n\;\log\log\log n}{\log\log n} \;+\; O\!\left(\frac{\log n}{\log\log n}\right) \quad (3.1)$$

As a result, for all $\varepsilon > 0$,

(i) $\quad n\,\mathbb{P}\Big(\tilde{D}_n - \dfrac{\log n}{\log\log n} \geqslant (1+\varepsilon)m_n\Big) \to 0.$

(ii) $\quad n\,\mathbb{P}\Big(\tilde{D}_n - \dfrac{\log n}{\log\log n} \geqslant (1-\varepsilon)m_n\Big) \to +\infty.$

Thus, to apply Lemma 3.4.6 and finish the proof it suffices to show that

$$\mathbb{P}\Big(\tilde{D}_n^{(1)} \geqslant k_n,\, \tilde{D}_n^{(2)} \geqslant k_n\Big) \;\sim\; \mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)^2$$

whenever $k_n = (\log n)/(\log\log n) + (1-\varepsilon)m_n$. More precisely, using Remark 3.4.7 it is sufficient to show that

$$\mathbb{P}\Big(\tilde{D}_n^{(1)} \geqslant k_n,\, \tilde{D}_n^{(2)} \geqslant k_n\Big) \;\leqslant\; \mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)^2 + o\Big(\mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)^2\Big)$$

First let us fix $b_1 \neq b_2 \in \{1,\dots,n\}$. Conditional on $\{B_1 = b_1, B_2 = b_2\}$, recall that the variables $(X_\ell^{(2)},\, b_2+1 \leqslant \ell \leqslant n)$ are independent Bernoulli variables with parameter $1/(n-1)$. By further conditioning on the variables $X_\ell^{(1)}$, the independence of $(X_\ell^{(2)},\, b_2+1 \leqslant \ell \leqslant n)$ still holds but their distribution is changed. Indeed, choose $(x_\ell,\, \ell \neq b_1) \in \{0,1\}^{n-1}$ and consider the event

$$A \coloneqq \{B_1 = b_1, B_2 = b_2,\; \forall \ell \neq b_1, X_\ell^{(1)} = x_\ell\}.$$

Then by construction, for all $\ell \notin \{b_1, b_2\}$, we have

$$\mathbb{P}\Big(X_\ell^{(2)} = 1 \,\Big|\, A\Big) = \begin{cases} 0 & \text{if } x_\ell = 1 \\ \frac{1}{n-2} & \text{if } x_\ell = 0. \end{cases}$$

Consequently $X_\ell^{(2)}$ is always stochastically dominated by a Bernoulli($\frac{1}{n-2}$) random variable, and so we bound the distribution of $\tilde{D}_n^{(2)} = \sum_{\ell > b_2} X_\ell^{(2)}$ conditional on A by

$$(\tilde{D}_n^{(2)} \mid A) \;\overset{d}{\leqslant}\; \text{Binomial}\left(n - b_2, \frac{1}{n-2}\right).$$

To get a bound on the distribution of $\tilde{D}_n^{(2)}$ conditional on $\tilde{D}_n^{(1)} = i$ for some i, first note that summing over all configurations $b_1, b_2, (x_\ell, \ell \neq b_1)$ such that $\sum_{\ell > b_1} x_\ell = i$ gives

$$(\tilde{D}_n^{(2)} \mid B_1 = b_1, B_2 = b_2,\, \tilde{D}_n^{(1)} = i) \;\overset{d}{\leqslant}\; \text{Binomial}\left(n - b_2, \frac{1}{n-2}\right).$$

Let us now write for conciseness $L_1 = n - B_1$ and $L_2 = n - B_2$. Note that L_2 is not independent of $\{\tilde{D}_n^{(1)} = i\}$ because they are linked by L_1. Indeed, L_1 is positively correlated to $\tilde{D}_n^{(1)}$ and we always have $L_2 \neq L_1$. Nevertheless, since conditional on L_1, L_2 is independent of $\tilde{D}_n^{(1)}$ and uniform on $\{0,\dots,n-1\} \setminus L_1$, we have the following stochastic ordering:

$$(L_2 \mid B_1 = b_1,\, \tilde{D}_n^{(1)} = i) \;\overset{d}{\leqslant}\; \bar{L}_2\,,$$

where $\bar{L}_2$ is uniformly distributed on $\{1,\dots,n-1\}$. Summing over b_1 and b_2, one thus get

$$(\tilde{D}_n^{(2)} \mid \tilde{D}_n^{(1)} = i) \;\overset{d}{\leqslant}\; \text{Binomial}\left(\bar{L}_2, \frac{1}{n-2}\right).$$

Let us define a random variable $M_n \sim \text{Bin}\left(\bar{L}_2, \frac{1}{n-2}\right)$. As the previous bound is uniform in i, we have

$$\mathbb{P}\Big(\tilde{D}_n^{(2)} \geqslant k_n \,\Big|\, \tilde{D}_n^{(1)} \geqslant k_n\Big) \;\leqslant\; \mathbb{P}(M_n \geqslant k_n).$$

To conclude, it is sufficient to show that $\mathbb{P}(M_n \geqslant k_n) \sim \mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)$ since this would imply

$$\mathbb{P}\left(\tilde{D}_n^{(1)} \geqslant k_n,\, \tilde{D}_n^{(2)} \geqslant k_n\right) \leqslant \mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)\mathbb{P}(M_n \geqslant k_n) \sim \mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)^2.$$

For this, define on the same probability space as the variables $\bar{L}_2$ and M_n the variable

$$\underline{L}_2 \coloneqq \bar{L}_2 \mathbb{1}_{\{\bar{L}_2 \leqslant n-2\}}.$$

$\underline{L}_2$ is then uniformly distributed on $\{0, \dots, n-2\}$, and we have the equality in distribution

$$M_n \mathbb{1}_{\{\bar{L}_2 \leqslant n-2\}} \overset{d}{=} \tilde{D}_{n-1} \sim \text{Binomial}\left(\underline{L}_2, \frac{1}{n-2}\right).$$

As the two variables M_n and $M_n \mathbb{1}_{\{\bar{L}_2 \leqslant n-2\}}$ differ on an event of probability no greater than $1/(n-1)$, we have

$$\mathbb{P}(M_n \geqslant k_n) = \mathbb{P}\left(\tilde{D}_{n-1} \geqslant k_n\right) + O(\tfrac{1}{n}),$$

and finally (3.1) with $\gamma = (1-\varepsilon)$ allows us to conclude that this expression is indeed equivalent to $\mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)$. □

3.5 Tree sizes

In this section, we study the size of the trees composing the Moran forest. Section 3.5.2 is concerned with the typical size of these trees, while Section 3.5.3 focuses on the asymptotics of the size of the largest tree. But before going any further we need to introduce a process that will play a central role throughout the rest of this paper.

3.5.1 A discrete-time Yule process

Let $\Upsilon_n = (\Upsilon_n(\ell),\, \ell \geqslant 0)$ be the Markov chain defined by $\Upsilon_n(0) = 1$ and the following transition probabilities:

$$\mathbb{P}\left(\Upsilon_n(\ell+1) = j \mid \Upsilon_n(\ell) = i\right) = \begin{cases} \frac{i}{n-1} & \text{if } j = i+1 \\ 1 - \frac{i}{n-1} & \text{if } j = i, \end{cases}$$

and stopped when reaching n.

The reason why this process will play an important role when studying the trees of $\mathscr{F}_n$ is the following: let $\mathscr{T}_n^{(v)}$ denote the tree containing v, and $\tilde{\mathscr{T}}_n^{(v)}$ the subtree descending from v in the UA construction—that is, letting $m(v)$ denote the mother of v and $\mathscr{T}_n^{(v)} \setminus \{vm(v)\}$ the forest obtained by removing the edge between v and $m(v)$ from $\mathscr{T}_n^{(v)}$ (if that edge existed), $\tilde{\mathscr{T}}_n^{(v)}$ is the tree of $\mathscr{T}_n^{(v)} \setminus \{vm(v)\}$ containing v. Recalling that L_v denotes the number of steps after vertex v was added in the UA construction and letting $\tilde{T}_n^{(v)} = |\tilde{\mathscr{T}}_n^{(v)}|$ be the size of $\tilde{\mathscr{T}}_n^{(v)}$, we have

$$\tilde{T}_n^{(v)} \overset{d}{=} \Upsilon_n(L_v),$$

where Υ_n is independent of L_v. In particular, the size of a tree created at step $n-h$ of the UA construction is distributed as $\Upsilon_n(h)$.

In the rest of this section, we list a few basic properties of Υ_n that will be used in subsequent proofs.

Lemma 3.5.1. *For all* $0 \leqslant \ell \leqslant n-1$,

$$\mathbb{E}(\Upsilon_n(\ell)) \;=\; \Big(1+\frac{1}{n-1}\Big)^{\ell}.$$

Proof. For $0 \leqslant \ell < n-1$, we have $\Upsilon_n(\ell) < n$ almost surely, therefore we can write

$$\begin{aligned}\mathbb{E}(\Upsilon_n(\ell+1) \,|\, \Upsilon_n(\ell)) \;&=\; \frac{\Upsilon_n(\ell)}{n-1}(\Upsilon_n(\ell)+1) \;+\; \Big(1-\frac{\Upsilon_n(\ell)}{n-1}\Big)\Upsilon_n(\ell)\\ &=\; \Upsilon_n(\ell)\Big(1+\frac{1}{n-1}\Big),\end{aligned}$$

and the result follows by induction. □

We now compare the discrete-time process Υ_n to the Yule process. By Yule process, we refer to the continuous-time Markov chain $(Y(t),\, t \geqslant 0)$ that jumps from i to $i+1$ at rate i (see e.g. [19], Section 5.3).

Lemma 3.5.2. *As* $n \to \infty$,

$$(\Upsilon_n(\lfloor tn \rfloor),\, t \geqslant 0) \;\Longrightarrow\; (Y(t),\, t \geqslant 0),$$

where " $\Longrightarrow$ *" denotes convergence in distribution in the Skorokhod space [6], and* $(Y(t),\, t \geqslant 0)$ *is a Yule process.*

Proof. Since both processes only have jumps of size 1, it suffices to prove that the sequence of jump times of $(\Upsilon_n(\lfloor tn \rfloor),\, t \geqslant 0)$ converges in distribution to that of the Yule process. For $1 \leqslant i \leqslant n$, let

$$t_n(i) = \inf\{\ell \geqslant 0 : \Upsilon_n(\ell) = i\}$$

be the jump times of the chain Υ_n. By the strong Markov property, the variables $(t_n(i+1) - t_n(i),\, 1 \leqslant i \leqslant n-1)$ are independent, and $t_n(i+1) - t_n(i) \sim \text{Geometric}(\frac{i}{n-1})$. Therefore,

$$\Big(\tfrac{1}{n}\big(t_n(i+1) - t_n(i)\big),\, 1 \leqslant i \leqslant n-1\Big) \;\xrightarrow[n\to\infty]{d}\; (\mathscr{E}(i),\, i \geqslant 1),$$

where the variables $(\mathscr{E}(i),\, i \geqslant 1)$ are independent and $\mathscr{E}(i) \sim \text{Exponential}(i)$. This concludes the proof. □

Lemma 3.5.3. *For all integers* $0 \leqslant k \leqslant \ell \leqslant n-1$,

$$\mathbb{P}\Big(Y\big(\tfrac{\ell-k+1}{n-1}\big) > k\Big) \;\leqslant\; \mathbb{P}(\Upsilon_n(\ell) > k) \;\leqslant\; \mathbb{P}\Big(Y\big(\lambda_n(k)\tfrac{\ell}{n-1}\big) > k\Big),$$

where

$$\lambda_n(k) = -\frac{n-1}{k}\log\Big(1-\frac{k}{n-1}\Big).$$

Proof. Let us start with the upper bound, and write $\lambda := \lambda_n(k)$ for simplicity. Note that, for all $t \geqslant 0$ and $i \geqslant 1$,

$$\mathbb{P}\Big(Y\big(t+\tfrac{\lambda}{n-1}\big) = i \;\Big|\; Y(t) = i\Big) = e^{-\frac{i\lambda}{n-1}},$$

and that we have chosen λ such that if $i \leqslant k$ then

$$e^{-\frac{i\lambda}{n-1}} \leqslant 1 - \frac{i}{n-1} = \mathbb{P}\big(\Upsilon_n(\ell+1) = i \,\big|\, \Upsilon_n(\ell) = i\big).$$

Thus, until it reaches $k+1$ individuals, the process Υ_n is dominated by the Markov chain $(Y(\frac{\lambda\ell}{n-1}),\, 0 \leqslant \ell \leqslant n-1)$. This shows that

$$\mathbb{P}(\Upsilon_n(\ell) > k) \leqslant \mathbb{P}\Big(Y(\tfrac{\lambda\ell}{n-1}) > k\Big),$$

proving the second inequality of the lemma.

To prove the first inequality, we couple Υ_n with a "censored" Yule process Y_{c}. Intuitively, this censoring consists in ignoring births that occur less than $1/(n-1)$ unit of time after another birth.

Formally, we define Y_{c} by specifying the sequence $t_0 = 0 < t_1 < t_2 < \dots$ of times corresponding to births in the population. Let $(\mathscr{E}_i,\, i \geqslant 1)$ be an independent sequence of exponential random variables where $\mathscr{E}_i \sim \text{Exponential}(i)$. Set $t_0 = 0$ and, for each $i \geqslant 1$,

$$t_i \coloneqq \mathscr{E}_1 + \sum_{j=2}^{i} \Big(\frac{1}{n-1} + \mathscr{E}_i\Big) = \frac{i-1}{n-1} + \sum_{j=1}^{i} \mathscr{E}_i. \tag{3.2}$$

We now define, for all $t \geqslant 0$,

$$Y_{\mathrm{c}}(t) \coloneqq 1 + \sum_{i \geqslant 1} \mathbb{1}_{\{t_i \leqslant t\}} = \sum_{i \geqslant 1} i\, \mathbb{1}_{\{t_{i-1} \leqslant t < t_i\}}.$$

The censoring of the Yule process after birth events implies that for any time $t \geqslant 0$, the random variable $Y_{\mathrm{c}}(t + \frac{1}{n-1}) - Y_{\mathrm{c}}(t)$ takes values in $\{0, 1\}$. Furthermore, for any $i \in \mathbb{N}$,

$$\mathbb{P}\big(Y_{\mathrm{c}}(t + \tfrac{1}{n-1}) = i+1 \,\big|\, Y_{\mathrm{c}}(t) = i\big) \leqslant 1 - e^{-\frac{i}{n-1}} \leqslant \frac{i}{n-1}.$$

Therefore, we can couple $(\Upsilon_n(\ell),\, 0 \leqslant \ell \leqslant n-1)$ and $(Y_{\mathrm{c}}(t),\, t \geqslant 0)$ in such a way that, for all $0 \leqslant \ell \leqslant n-1$,

$$Y_{\mathrm{c}}\Big(\tfrac{\ell}{n-1}\Big) \leqslant \Upsilon_n(\ell).$$

Now, by construction, the sequence $(t_i - \frac{i-1}{n-1},\, i \geqslant 1)$ has the distribution of the sequence of jump times of a Yule process. Therefore,

$$\begin{aligned}
\mathbb{P}(\Upsilon_n(\ell) > k) &\geqslant \mathbb{P}\Big(Y_{\mathrm{c}}(\tfrac{\ell}{n-1}) > k\Big) \\
&= \mathbb{P}\Big(t_k \leqslant \tfrac{\ell}{n-1}\Big) \\
&= \mathbb{P}\Big(t_k - \tfrac{k-1}{n-1} \leqslant \tfrac{\ell-k+1}{n-1}\Big) \\
&= \mathbb{P}\Big(Y(\tfrac{\ell-k+1}{n-1}) > k\Big),
\end{aligned}$$

which yields the lower bound of the lemma. □

3.5.2 Size of some random trees

In this section, we study the size of some typical trees of $\mathscr{F}_n$. In particular, we study the asymptotics of size $T_n^{(1)}$ of the tree containing vertex 1 and of the size T_n^U of a tree sampled uniformly at random among the trees composing $\mathscr{F}_n$. Our main result is the following theorem.

Theorem 3.5.4.

(i) *Let T_n^U be the size of a uniform tree of $\mathscr{F}_n$. Then,*

$$\mathbb{P}\left(T_n^U = k\right) \xrightarrow[n\to\infty]{} 2\int_0^1 xe^{-x}(1-e^{-x})^{k-1}dx,$$

that is, $T_n^U \xrightarrow{d} T^U$ where $T^U \sim \text{Geometric}(e^{-X})$, and $X \sim 2xdx$ on $[0,1]$.

(ii) *Let $T_n^{(1)}$ be the size of the tree containing vertex 1 in $\mathscr{F}_n$. Then,*

$$\mathbb{P}\left(T_n^{(1)} = k\right) \xrightarrow[n\to\infty]{} k\int_0^1 xe^{-x}(1-e^{-x})^{k-1}dx,$$

that is, $T_n^{(1)}$ converges in distribution to the size-biasing of T^U.

Remark 3.5.5. Note that even though the limit distribution of $T_n^{(1)}$ is the size-biased limit distribution of T_n^U, for finite n the distribution of $T_n^{(1)}$ is *not* the size-biased distribution of T_n^U. ◇

We start by giving the distribution of $T_n^{(1)}$ in terms of the process Υ_n defined in Section 3.5.1. For this, we first need to introduce some notation. Let $\mathscr{T}_n^{(v)}$ be the tree containing vertex v in $\mathscr{F}_n$. We denote by $H_n^{(v)}$ the number of steps after the root of $\mathscr{T}_n^{(v)}$ was added in the UA construction. Recalling the notation from Section 3.2.2, where $\sigma^{-1}(v) \in \{1,\dots,n\}$ denotes the step of the UA construction at which vertex v was added, we thus have

$$H_n^{(v)} \;=\; n - \min\left\{\sigma^{-1}(u) : u \in \mathscr{T}_n^{(v)}\right\}.$$

Proposition 3.5.6. *Let $T_n^{(1)}$ be the size of the tree containing vertex 1 in $\mathscr{F}_n$, and denote by $H_n^{(1)}$ the number of steps after the root of that tree was added in the UA construction. Then,*

(i) *For $0 \leqslant h \leqslant n-1$, $\mathbb{P}(H_n^{(1)} = h) = \dfrac{h}{n(n-1)}\left(1+\dfrac{1}{n-1}\right)^h$.*

(ii) *Conditional on $\{H_n^{(1)} = h\}$, $T_n^{(1)}$ is distributed as the size-biasing of $\Upsilon_n(h)$.*

Remark 3.5.7. The size-biasing of $\Upsilon_n(h)$ can be easily represented as follows. Consider the Markov chain $\Upsilon_n^* = (\Upsilon_n^*(\ell),\, 0 \leqslant \ell \leqslant n-1)$ defined by $\Upsilon_n^*(0) = 1$ and the following transition probabilities:

$$\mathbb{P}\left(\Upsilon_n^*(\ell+1) = j \mid \Upsilon_n^*(\ell) = i\right) = \begin{cases} \frac{i+1}{n} & \text{if } j = i+1 \\ 1 - \frac{i+1}{n} & \text{if } j = i. \end{cases}$$

A straightforward induction on ℓ shows that $\Upsilon_n^*(\ell)$ is distributed as the size-biasing of $\Upsilon_n(\ell)$. ◇

Proof. First, note that $H_n^{(1)} = h$ if and only if a new tree is created at step $n-h$, and vertex 1 belongs to this tree. Now, the probability that a new tree is created at step $n-h$ is $\frac{h}{n-1}$, and the size of this tree is then distributed as $\Upsilon_n(h)$. Moreover, at the end of the UA construction, the labels are assigned to the vertices uniformly. As a result, conditional on a tree having size i, the probability that it contains vertex 1 is i/n. We thus have

$$\mathbb{P}\Big(H_n^{(1)} = h,\, T_n^{(1)} = i\Big) \;=\; \frac{h}{n-1} \cdot \frac{i}{n}\, \mathbb{P}(\Upsilon_n(h) = i).$$

Summing over i and using Lemma 3.5.1 yields

$$\mathbb{P}\Big(H_n^{(1)} = h\Big) \;=\; \frac{h}{n(n-1)}\Big(1 + \frac{1}{n-1}\Big)^h.$$

Finally,

$$\mathbb{P}\Big(T_n^{(1)} = i \;\Big|\; H_n^{(1)} = h\Big) = i\,\mathbb{P}(\Upsilon_n(h) = i)\Big(1 + \frac{1}{n-1}\Big)^{-h},$$

which concludes the proof. □

We can now turn to the proof of our main result.

Proof of Theorem 3.5.4. (i) First recall the notation of the UA construction and Section 3.4, and note that conditional on the event

$$\{I_{n-h} = 0\} = \{\text{a new tree is created at step } n-h \text{ of the UA construction}\},$$

the total number of trees has distribution

$$(N_n \mid I_{n-h} = 0) \;\overset{d}{=}\; 1 + \sum_{\substack{\ell=1 \\ \ell \neq n-h}}^{n} (1 - I_\ell),$$

where $I_\ell \sim \mathrm{Ber}(\frac{\ell-1}{n-1})$ are independent random variables. From this, it is clear that uniformly in h,

$$\mathbb{E}(N_n \mid I_{n-h} = 0) \sim \frac{n}{2}\,. \tag{3.3}$$

On the event $\{I_{n-h} = 0\}$, let us denote by $\mathcal{T}_{n,h}$ the size of the tree created at step $n-h$. Note that the marginal distribution of $\mathcal{T}_{n,h}$ is simply $\Upsilon_n(h)$. Let us now compute

$$\begin{aligned}\mathbb{P}\Big(T_n^U = k,\, H_n^U = h\Big) &= \frac{h}{n-1}\mathbb{P}\Big(\mathcal{T}_{n,h} = k,\, H_n^U = h \;\Big|\; I_{n-h} = 0\Big) \\ &= \frac{h}{n-1}\mathbb{E}\Big(\frac{1}{N_n}\mathbb{1}_{\{\mathcal{T}_{n,h}=k\}} \;\Big|\; I_{n-h} = 0\Big),\end{aligned}$$

and note that

$$\begin{aligned}\frac{n}{2}\,\mathbb{E}\Big(\frac{1}{N_n}\mathbb{1}_{\{\mathcal{T}_{n,h}=k\}} \;\Big|\; I_{n-h} = 0\Big) \;&=\; \mathbb{P}(\Upsilon_n(h) = k) \\ &\quad+\; \mathbb{E}\Big(\Big(\frac{n/2}{N_n} - 1\Big)\mathbb{1}_{\{\mathcal{T}_{n,h}=k\}} \;\Big|\; I_{n-h} = 0\Big).\end{aligned} \tag{3.4}$$

The last term in this display goes to zero as $n \to \infty$, uniformly in h. Indeed, using (3.3) and applying Hoeffding's inequality [13] to N_n, which is a sum of n independent Bernoulli random variables, we get, for $\varepsilon > 0$ and uniformly in h,

$$\mathbb{P}\Big(\Big|\tfrac{N_n}{n/2} - 1\Big| > \varepsilon \;\Big|\; I_{n-h} = 0\Big) \;\leqslant\; 2e^{-Cn},$$

where C is a positive constant that depends only on ε. Using that for any $0 < \varepsilon < 1/2$ and positive x, we have $\left|\frac{1}{x} - 1\right| > 2\varepsilon \implies |x-1| > \varepsilon$, we may bound

$$\begin{aligned}\mathbb{E}\Big(\Big|\tfrac{n/2}{N_n} - 1\Big| \,\Big|\, I_{n-h} = 0\Big) &\leqslant 2\varepsilon + \frac{n}{2}\mathbb{P}\Big(\Big|\tfrac{n/2}{N_n} - 1\Big| > 2\varepsilon \,\Big|\, I_{n-h} = 0\Big)\\ &\leqslant 2\varepsilon + \frac{n}{2}\mathbb{P}\Big(\Big|\tfrac{N_n}{n/2} - 1\Big| > \varepsilon \,\Big|\, I_{n-h} = 0\Big)\\ &\leqslant 2\varepsilon + ne^{-Cn}.\end{aligned}$$

This shows that the last term in (3.4) goes to zero uniformly in h. We thus get

$$\mathbb{P}\Big(T_n^U = k,\, H_n^U = h\Big) = \frac{2h}{n^2}(\mathbb{P}(\Upsilon_n(h) = k) + o(1)),$$

and so using Lemma 3.5.2, summing over h yields

$$\begin{aligned}\mathbb{P}\Big(T_n^U = k\Big) &= \frac{2}{n}\sum_{h=0}^{n-1}\frac{h}{n}\mathbb{P}(\Upsilon_n(h) = k) + o(1)\\ &\xrightarrow[n\to\infty]{} 2\int_0^1 x\mathbb{P}(Y(x) = k)\,dx.\end{aligned}$$

Recalling the well-known fact that $Y(x)$ has a Geometric(e^{-x}) distribution (see for instance Section 5.3 in [19]) proves the first point.

(ii) We know from Proposition 3.5.6 that

$$\begin{aligned}\mathbb{P}\Big(T_n^{(1)} = k\Big) &= \frac{1}{n}\sum_{h=0}^{n-1}\frac{h}{n-1}\mathbb{E}\Big(\Upsilon_n(h)\mathbb{1}_{\{\Upsilon_n(h)=k\}}\Big)\\ &= \frac{k}{n}\sum_{h=0}^{n-1}\frac{h}{n-1}\mathbb{P}(\Upsilon_n(h) = k).\end{aligned}$$

Again, using Lemma 3.5.2 and dominated convergence, we have

$$\frac{k}{n}\sum_{h=0}^{n-1}\frac{h}{n-1}\mathbb{P}(\Upsilon_n(h) = k) \xrightarrow[n\to\infty]{} k\int_0^1 x\mathbb{P}(Y(x) = k)dx,$$

which yields the result. □

3.5.3 Size of the largest tree

The goal of this section is to derive asymptotics for $T_n^{\max} := \max_v T_n^{(v)}$, the size of the largest tree in the Moran forest on n vertices, when $n \to \infty$.

Theorem 3.5.8. *Let $T_n^{\max}$ denote the size of the largest tree in $\mathcal{F}_n$. Then*

$$T_n^{\max} = \alpha\big(\log n - (1 + o_{\mathrm{p}}(1))\log\log n\big),$$

where $\alpha = (1 - \log(e-1))^{-1} \approx 2.18019$ and $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

As in Section 3.5.1, for any vertex v let us define $\tilde{\mathcal{T}}_n^{(v)} \subset \mathcal{T}_n^{(v)}$ as the subtree descending from v in the UA construction. For our purpose, it will be sufficient

to study the size $\tilde{T}_n^{(v)} := |\tilde{\mathcal{T}}_n^{(v)}|$ of those subtrees instead of that of the trees $\mathcal{T}_n^{(v)}$. Indeed, observe that

$$T_n^{\max} = \max_v \tilde{T}_n^{(v)},$$

so that applying Lemma 3.4.6 with $m_n = \alpha \log\log n$ and $\beta = -1$ to the exchangeable variables $(\tilde{T}_n^{(1)} - \alpha \log n, \dots, \tilde{T}_n^{(n)} - \alpha \log n)$ will prove the theorem. Again, we omit the superscript and denote by $\tilde{T}_n$ a random variable with distribution equal to that of $\tilde{T}_n^{(1)}$.

For the rest of the section, we thus study the tail probabilities of the variable $\tilde{T}_n$. Recall from the UA construction that the number L of steps after a fixed vertex was added is uniformly distributed on $\{0, \dots, n-1\}$, and from Section 3.5.1 that, conditional on $\{L = \ell\}$,

$$\tilde{T}_n \stackrel{d}{=} \Upsilon_n(\ell)\,.$$

Proposition 3.5.9. *For any sequence of integers $k_n \to \infty$ with $k_n = o(\sqrt{n})$,*

$$\mathbb{P}\Big(\tilde{T}_n > k_n\Big) \sim \frac{e}{k_n}(1 - e^{-1})^{k_n+1}.$$

Proof. Using the upper bound in Lemma 3.5.3 and the fact that L is uniform on $\{0, \dots, n-1\}$, we have

$$\begin{aligned}
\mathbb{P}\Big(\tilde{T}_n > k_n\Big) &\leqslant \frac{1}{n}\sum_{\ell=0}^{n-1} \mathbb{P}\Big(Y\big(\lambda_n(k)\tfrac{\ell}{n-1}\big) > k_n\Big) \\
&= \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\big(\lambda_n(k_n)\tfrac{\lfloor x(n-1)\rfloor}{n-1}\big) > k_n\Big)\,dx + \frac{1}{n}\mathbb{P}(Y(\lambda_n(k_n)) > k_n) \\
&\leqslant \int_0^1 \mathbb{P}(Y(\lambda_n(k_n)x) > k_n)\,dx + \frac{1}{n}(1 - e^{-\lambda_n(k_n)})^{k_n} \\
&= \int_0^1 \big(1 - e^{-\lambda_n(k_n)x}\big)^{k_n}\,dx + \frac{1}{n}(1 - e^{-\lambda_n(k_n)})^{k_n}.
\end{aligned}$$

Now recall that $\lambda_n(k_n) = -\frac{n-1}{k_n}\log\big(1 - \frac{k_n}{n-1}\big) = 1 + O(\frac{k_n}{n})$, so uniformly in $x \in [0,1]$,

$$e^{-\lambda_n(k_n)x} = e^{-x} + O(\frac{k_n}{n}).$$

Since $k_n = o(\sqrt{n})$, we have $k_n/n = o(1/k_n)$ and thus Lemma 3.B.1 from the Appendix gives

$$\int_0^1 \big(1 - e^{-\lambda_n(k_n)x}\big)^{k_n}\,dx \sim \frac{e}{k_n}(1 - e^{-1})^{k_n+1}.$$

Elementary calculations also show that when $k_n = o(\sqrt{n})$, we have

$$\frac{1}{n}(1 - e^{-\lambda_n(k_n)})^{k_n} \sim \frac{1}{n}(1 - e^{-1})^{k_n} = o\Big(\frac{(1 - e^{-1})^{k_n}}{k_n}\Big).$$

It remains to examine the lower bound in Lemma 3.5.3. As above, we get an integral

$$\begin{aligned}
\mathbb{P}\Big(\tilde{T}_n > k_n\Big) &\geqslant \frac{1}{n}\sum_{\ell=0}^{n-1} \mathbb{P}\Big(Y\big(\tfrac{\ell - k_n + 1}{n-1}\big) > k_n\Big) \\
&\geqslant \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\big(\tfrac{\lceil x(n-1)\rceil - k_n}{n-1}\big) > k_n\Big)\,dx \\
&\geqslant \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\big(x - \tfrac{k_n}{n-1}\big) > k_n\Big)\,dx.
\end{aligned}$$

Since

$$\mathbb{P}\Big(Y(x-\tfrac{k_n}{n-1}) > k_n\Big) = \big(1-\exp(-x+\tfrac{k_n}{n-1})\big)^{k_n} = \big(1-e^{-x}+O(k_n/n)\big)^{k_n},$$

using Lemma 3.B.1 again, we get

$$\mathbb{P}\Big(\tilde{T}_n > k_n\Big) \geqslant \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y(x-\tfrac{k_n}{n-1}) > k_n\Big)\,dx \sim \frac{e}{k_n}(1-e^{-1})^{k_n+1},$$

which completes the proof. □

Note that if k_n is not integer-valued, then

$$\mathbb{P}\Big(\tilde{T}_n > k_n\Big) = \mathbb{P}\Big(\tilde{T}_n > \lfloor k_n\rfloor\Big) \sim \frac{e}{k_n}(1-e^{-1})^{\lfloor k_n\rfloor+1},$$

which is not necessarily equivalent to $\frac{e}{k_n}(1-e^{-1})^{k_n+1}$ since $k_n - \lfloor k_n\rfloor$ may oscillate between 0 and 1. However, we do have $\mathbb{P}(\tilde{T}_n > k_n) = \Theta((1-e^{-1})^{k_n}/k_n)$, where the Bachmann–Landau notation $u_n = \Theta(v_n)$ indicates that there exist positive constants c, C such that $c\,v_n \leqslant u_n \leqslant C v_n$ for n large enough. This approximation is sufficient for our purpose.

We may now prove Theorem 3.5.8 using the first and second moment method that we already used for the largest degree.

Proof of Theorem 3.5.8. We apply Lemma 3.4.6 to the exchangeable variables

$$(X_n^{(1)},\dots,X_n^{(n)}) = (\tilde{T}_n^{(1)} - \alpha\log n,\dots,\tilde{T}_n^{(n)} - \alpha\log n)\,,$$

with $m_n = \alpha\log\log n$ and $\beta = -1$. The first two points of the lemma are readily checked, since Proposition 3.5.9 tells us that for $\alpha = (1-\log(e-1))^{-1} = -(\log(1-e^{-1}))^{-1}$ and any $\gamma > 0$, we have for $k_n := \alpha(\log n - \gamma\log\log n)$

$$\mathbb{P}\Big(\tilde{T}_n - \alpha\log n \geqslant -\gamma\alpha\log\log n\Big) = \mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big) = \Theta\Big(\frac{(\log n)^{\gamma-1}}{n}\Big).$$

Thus, for all $\varepsilon > 0$,

(i) $n\mathbb{P}\Big(\tilde{T}_n - \alpha\log n \geqslant (-1+\varepsilon)\alpha\log\log n\Big) \to 0.$

(ii) $n\mathbb{P}\Big(\tilde{T}_n - \alpha\log n \geqslant (-1-\varepsilon)\alpha\log\log n\Big) \to +\infty.$

All that remains to check is the third point of the lemma. From now we fix $k_n = \alpha(\log n - (1+\varepsilon)\log\log n)$ for some $\varepsilon > 0$, and for the sake of readability, we set $R_n := \mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big)$. With this notation, given Remark 3.4.7 we need to show that

$$\mathbb{P}\,(\tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n) \;\leqslant\; R_n^2 + o(R_n^2). \tag{3.5}$$

Since this is rather technical, we defer the complete proof to Lemma 3.B.2 in Appendix 3.B, and only outline the main ideas of the proof here. As in the study of the largest degree, we prove this by showing that the law of $\tilde{T}_n^{(2)}$ conditional on $\{\tilde{T}_n^{(1)} \geqslant k_n\}$ is close to its unconditional law. We first prove that

$$\mathbb{P}\,(A_n,\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n) \;=\; o(R_n^2),$$

where $A_n := \{\tilde{\mathcal{T}}_n^{(2)} \subset \tilde{\mathcal{T}}_n^{(1)}\} \sqcup \{\tilde{\mathcal{T}}_n^{(1)} \subset \tilde{\mathcal{T}}_n^{(2)}\}$ is the event that one of the two vertices 1 and 2 is an ancestor of the other in the UA construction. We then show that

$$\mathbb{P}\left(A_n^{\mathrm{c}}, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \leqslant R_n^2 + o(R_n^2)\,,$$

where A_n^{c} denotes the complement of A_n. This is done by showing that, conditional on $\{\tilde{T}_n^{(1)} = i\}$, on the event A_n^{c} the process counting the number of vertices of the tree $\mathcal{T}_n^{(2)}$ in the UA construction behaves as a modified Υ_n process, which we essentially bound from above by Υ_{n-i}. Therefore, $\tilde{T}_n^{(2)}$ can be compared with an independent variable with distribution $\tilde{T}_{n-i}$. Finally, we show that

$$\sum_{i \geqslant k_n} \mathbb{P}\left(\tilde{T}_n^{(1)} = i\right)\mathbb{P}\left(\tilde{T}_{n-i} \geqslant k_n\right) \leqslant R_n^2 + o(R_n^2),$$

thereby proving (3.5) and concluding the proof of Theorem 3.5.8. □

Literature cited in this chapter

[1] The On-Line Encyclopedia of Integer Sequences, published electronically at http://oeis.org, 2019.

[2] M. Aigner and G. M. Ziegler. *Proofs from THE BOOK*. Springer-Verlag Berlin, 6th edition, 2018.

[3] K. B. Athreya and S. N. Lahiri. *Measure theory and probability theory*. Springer Science+Business Media, 2006.

[4] K. T. Balińska, L. V. Quintas, and J. Szymański. Random recursive forests. *Random Structures & Algorithms*, 5(1):3–12, 1994.

[5] F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In *CAAP '92*, pages 24–48. Springer Berlin Heidelberg, 1992.

[6] P. Billingsley. *Convergence of Probability Measures*. Wiley, 2nd edition, 1999.

[7] B. Drake. *An inversion theorem for labeled trees and some limits of areas under lattice paths*. PhD thesis, Brandeis University, 2008.

[8] M. Drmota. *Random trees: an interplay between combinatorics and probability*. Springer-Verlag Vienna, 2009.

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

[10] Ö. Eğecioğlu and J. B. Remmel. Bijections for Cayley trees, spanning trees, and their q-analogues. *Journal of Combinatorial Theory, Series A*, 42(1):15–30, 1986.

[11] A. Etheridge. *Some mathematical models from population genetics. École d'été de probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer-Verlag Berlin Heidelberg, 2011.

[12] I. M. Gessel and S. Seo. A refinement of Cayley's formula for trees. *The electronic journal of combinatorics*, 11(2):R27, 2006.

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar. 1963.

[14] A. Joyal. Une théorie combinatoire des séries formelles. *Advances in mathematics*, 42(1):1–82, 1981.

[15] H. M. Mahmoud and R. T. Smythe. On the distribution of leaves in rooted subtrees of recursive trees. *The Annals of Applied Probability*, 1(3):406–418, 1991.

[16] A. Meir and J. W. Moon. Cutting down recursive trees. *Mathematical Biosciences*, 21(3):173–181, 1974.

[17] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

[18] W. C. Ratcliff, J. D. Fankhauser, D. W. Rogers, D. Greig, and M. Travisano. Origins of multicellular evolvability in snowflake yeast. *Nature Communications*, 6:6102, 2015.

[19] S. M. Ross. *Stochastic Processes*. Wiley, 2nd edition, 1995.

Appendices to Chapter 3

3.A Proof of point (ii) of Proposition 3.4.4

We want to prove that, for all $K_n = o(\sqrt{n})$, there exists $\varepsilon_n = o(1)$ such that, for all $k \leqslant K_n$,

$$|\mathbb{P}(D_n \geqslant k) - \mathbb{P}(D \geqslant k)| \;\leqslant\; \varepsilon_n \, \mathbb{P}(D \geqslant k)\,.$$

Doing this directly from the expressions of D_n and D involves unappealing calculations. To somewhat circumvent this, we make use of the simple expressions of the probability generating functions G_{D_n} and G_D. For this, let

$$\Delta_n(z) \;:=\; \sum_{i \geqslant 0} \big(\mathbb{P}(D \geqslant i) - \mathbb{P}(D_n \geqslant i)\big) z^i,$$

so that the k-th derivative of Δ_n evaluated at $z = 0$ is

$$\Delta_n^{(k)}(0) \;=\; k! \, \big(\mathbb{P}(D \geqslant k) - \mathbb{P}(D_n \geqslant k)\big).$$

Since $\mathbb{P}(D \geqslant k) \geqslant \frac{2/e}{(k+1)!}$, we have to show that for any given sequence $K_n = o(\sqrt{n})$,

$$\Delta_n^{(k)}(0) \;=\; \frac{\varepsilon_n}{k+1}$$

for some $\varepsilon_n \to 0$ and all $k \leqslant K_n$. Now, since for any non-negative integer-valued random variable X,

$$\sum_{i \geqslant 0} \mathbb{P}(X \geqslant i)\, z^i \;=\; \frac{z\,\mathbb{E}\left(z^X\right) - 1}{z - 1}\,,$$

we can express Δ_n in terms of the generating functions of D and D_n, that is,

$$\Delta_n(z) \;=\; \left(1 + \frac{1}{z-1}\right)\big(G_D(z) - G_{D_n}(z)\big)\,.$$

Moreover, we know from Proposition 3.4.3 that

$$G_D(z) = \; 2\,\frac{e^{z-1} - 1}{z-1} - 1 \;=\; 2 \sum_{i \geqslant 0} \frac{(z-1)^i}{(i+1)!} - 1$$

and from Proposition 3.4.1 that

$$\begin{aligned} G_{D_n}(z) &= 2\Big(1-\frac{1}{n}\Big)\frac{\Big(1+\frac{z-1}{n-1}\Big)^n-1}{z-1} \;-\; 1 \\ &= 2\Big(1-\frac{1}{n}\Big)\sum_{i=0}^{n-1}\binom{n}{i+1}\Big(\frac{1}{n-1}\Big)^{i+1}(z-1)^i \;-\; 1 \\ &= 2\sum_{i=0}^{n-1}\Big(\prod_{\ell=1}^{i}\frac{n-\ell}{n-1}\Big)\frac{(z-1)^i}{(i+1)!} \;-\; 1\,, \end{aligned}$$

where the empty product is 1. Therefore,

$$G_D(z)-G_{D_n}(z)=\sum_{i\geqslant 0}A(n,i)\frac{(z-1)^i}{(i+1)!}\,,$$

where

$$A(n,i)=2\Big[1-\Big(\prod_{\ell=1}^{i}\frac{n-\ell}{n-1}\Big)\mathbb{1}_{\{i\leqslant n-1\}}\Big]\,.$$

Using that $A(n,0)=A(n,1)=0$ and rearranging a bit, we obtain the following expansion of Δ_n at $z=1$:

$$\Delta_n(z) \;=\; \sum_{i\geqslant 1}\Big(A(n,i)+\frac{A(n,i+1)}{i+2}\Big)\frac{(z-1)^i}{(i+1)!}\,,$$

from which we get

$$\Delta_n^{(k)}(0) \;=\; \sum_{i\geqslant k}\Big(A(n,i)+\frac{A(n,i+1)}{i+2}\Big)\frac{(-1)^{i-k}}{(i-k)!\,(i+1)}\,.$$

Now, pick any $J_n=o(\sqrt{n})$ such that $K_n=o(J_n)$. For all $i<J_n$,

$$\Big|A(n,i)+\frac{A(n,i+1)}{i+2}\Big| \;\leqslant\; 4\Big(1-\prod_{\ell=1}^{J_n}\frac{n-\ell}{n-1}\Big) \;=\; \varepsilon_n\,,$$

with $\varepsilon_n\to 0$, since

$$\prod_{\ell=1}^{J_n}\frac{n-\ell}{n-1} \;\geqslant\; \Big(\frac{n-J_n}{n-1}\Big)^{J_n} \;=\; \exp\Big(-\tfrac{J_n^2}{n}+o\Big(\tfrac{J_n^2}{n}\Big)\Big)\,.$$

For $i\geqslant J_n$, we have

$$\Big|A(n,i)+\frac{A(n,i+1)}{i+2}\Big| \;\leqslant\; 4\,.$$

Combining these two upper bounds, we get

$$\begin{aligned} \big|\Delta_n^{(k)}(0)\big| &\leqslant \sum_{i=k}^{J_n-1}\frac{\varepsilon_n}{(i-k)!\,(i+1)} \;+\; \sum_{i\geqslant J_n}\frac{4}{(i-k)!\,(i+1)} \\ &\leqslant \frac{\varepsilon_n\,C_1}{(k+1)} \;+\; \frac{C_2}{(J_n+1)}\,. \end{aligned}$$

Finally, since $K_n=o(J_n)$, we have for all $k\leqslant K_n$,

$$\frac{1}{J_n+1} \;\leqslant\; \frac{1}{k+1}\cdot\frac{K_n+1}{J_n+1}\,,$$

with $(K_n+1)/(J_n+1)=o(1)$. This concludes the proof.

Note that although we have been quite crude in that we have used the triangle inequality on an alternating series, a more careful analysis would show that the $o(\sqrt{n})$ requirement on K_n is in fact optimal.

3.B Technical lemmas used in the proof of Theorem 3.5.8

Lemma 3.B.1. *For any sequence $k_n \to \infty$ and any sequence of measurable maps $f_n : [0,1] \to \mathbb{R}$ such that for all $x \in [0,1]$, $(1 - e^{-x} + f_n(x)) \geqslant 0$ and $\sup_x |f_n(x)| = o(1/k_n)$, we have*

$$\int_0^1 \left(1 - e^{-x} + f_n(x)\right)^{k_n} dx \sim \frac{e}{k_n}(1 - e^{-1})^{k_n+1}.$$

Proof. Let us compute

$$\begin{aligned}\int_0^1 \frac{(1 - e^{-x} + f_n(x))^{k_n}}{(1 - e^{-1})^{k_n}} k_n \, dx &= \int_0^1 \left(1 - \frac{e^{1-x} - 1}{e - 1} + \frac{e}{e-1} f_n(x)\right)^{k_n} k_n \, dx \\ &= \int_0^{k_n} \left(1 - \frac{y}{k_n} + g_n(y)\right)^{k_n} \frac{e-1}{1 + (e-1)\frac{y}{k_n}} \, dy,\end{aligned}$$

where we used the change of variable $y = k_n(e^{1-x} - 1)(e-1)^{-1}$, and defined the map g_n as

$$g_n(y) = \frac{e}{e-1} f_n\left(1 - \log\left(1 + \frac{y}{k_n}(e-1)\right)\right),$$

Now since $(1 - \frac{y}{k_n} + g_n(y))^{k_n} \leqslant \exp(-y + \frac{e}{e-1} k_n \sup_x f_n(x))$, it follows from dominated convergence that

$$\int_0^1 \frac{(1 - e^{-x} + f_n(x))^{k_n}}{(1 - e^{-1})^{k_n}} k_n \, dx \xrightarrow[n\to\infty]{} \int_0^\infty e^{-y}(e-1) \, dy = e - 1,$$

concluding the proof. □

Lemma 3.B.2. *Let $\tilde{T}_n^{(v)}$ denote the size of the subtree descending from v in the UA construction of $\tilde{\mathcal{F}}_n$. Then, for $\alpha = -1/\log(1 - e^{-1})$ and any $\varepsilon > 0$, letting $k_n = \alpha(\log n - (1+\varepsilon)\log\log n)$ and $R_n = \mathbb{P}\left(\tilde{T}_n \geqslant k_n\right)$,*

$$\mathbb{P}\left(\tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \leqslant R_n^2 + o(R_n^2).$$

Proof. We have $R_n = \Theta(\frac{(\log n)^\varepsilon}{n})$.

Let us denote by $A_n := \{\tilde{\mathcal{T}}_n^{(2)} \subset \tilde{\mathcal{T}}_n^{(1)}\} \sqcup \{\tilde{\mathcal{T}}_n^{(1)} \subset \tilde{\mathcal{T}}_n^{(2)}\}$ the event that one of the vertices 1 and 2 is an ancestor of the other. We start by showing that

$$\mathbb{P}\left(A_n, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) = o(R_n^2). \tag{3.6}$$

By exchangeability, we have

$$\begin{aligned}&\mathbb{P}\left(A_n, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \\ &\quad = 2\,\mathbb{P}\left(\tilde{\mathcal{T}}_n^{(2)} \subset \tilde{\mathcal{T}}_n^{(1)}, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \\ &\quad = \sum_{i \geqslant k_n} \mathbb{P}\left(\tilde{\mathcal{T}}_n^{(2)} \subset \tilde{\mathcal{T}}_n^{(1)}, \tilde{T}_n^{(2)} \geqslant k_n \mid \tilde{T}_n^{(1)} = i\right) \mathbb{P}\left(\tilde{T}_n = i\right).\end{aligned}$$

Let us call the *height* of a vertex the number of steps after it was added in the UA construction. Conditional on $\{\tilde{T}_n^{(1)} = i\}$ and on the heights of the vertices of $\tilde{\mathcal{T}}_n^{(1)}$ being $\ell_1 > \ldots > \ell_i$, the height L_2 of vertex 2 is uniformly distributed on $\{0, \ldots n-1\} \setminus \{\ell_1\}$. Moreover, in order to have

$$\{\tilde{\mathcal{T}}_n^{(2)} \subset \tilde{\mathcal{T}}_n^{(1)}, \tilde{T}_n^{(2)} \geqslant k_n\},$$

the height of vertex 2 must belong to $\{\ell_2, \dots, \ell_{i-(k_n-1)}\}$, which happens with probability $\frac{i-k_n}{n-1}$. Therefore,

$$\begin{aligned}\mathbb{P}\left(A_n, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) &\leqslant \sum_{i \geqslant k_n} \mathbb{P}\Big(\tilde{T}_n = i\Big) \frac{i-k_n}{n-1} \\ &= \frac{1}{n-1} \sum_{i > k_n} \mathbb{P}\Big(\tilde{T}_n \geqslant i\Big).\end{aligned}$$

To show that this is small enough, we let $K_n \coloneqq k_n + \alpha(\log n)^\delta$ with $0 < \delta < \min(1, \varepsilon)$, and $K_n' \coloneqq 2\alpha \log n$, and crudely bound

$$\sum_{i > k_n} \mathbb{P}\Big(\tilde{T}_n \geqslant i\Big) \leqslant (K_n - k_n)\mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big) + K_n' \,\mathbb{P}\Big(\tilde{T}_n \geqslant K_n\Big) + n\,\mathbb{P}\Big(\tilde{T}_n \geqslant K_n'\Big).$$

Now let us show that these three terms are negligible compared to nR_n^2. Recalling that $R_n = \Theta(\frac{(\log n)^\varepsilon}{n})$, we have $nR_n^2 = \Theta((\log n)^{2\varepsilon}/n)$ and therefore

- $(K_n - k_n)\mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big) \sim \alpha(\log n)^\delta R_n = \Theta\Big(\dfrac{(\log n)^{\delta+\varepsilon}}{n}\Big) = o(nR_n^2)$.
- $K_n' \,\mathbb{P}\Big(\tilde{T}_n \geqslant K_n\Big) = \Theta\big(\log n\, R_n e^{-(\log n)^\delta}\big) = o(R_n) = o(nR_n^2)$.
- $n\,\mathbb{P}\Big(\tilde{T}_n \geqslant K_n'\Big) = \Theta\Big(n\dfrac{n^{-2}}{\log n}\Big) = o(1/n) = o(nR_n^2)$.

Therefore (3.6) is proven, and it remains to show that

$$\mathbb{P}\left(A_n^{\mathrm{c}}, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \leqslant R_n^2 + o(R_n^2),$$

where A_n^{c} denotes the complement of A_n. We now fix $n \geqslant 1$, $i \geqslant k_n$, and a finite sequence $n-1 \geqslant \ell_1 > \ldots > \ell_i \geqslant 0$. Let us write B for the event that $\tilde{\mathcal{T}}_n^{(1)}$ contains exactly the vertices with heights $\ell_1 > \ldots > \ell_i$. Conditional on B, let us examine the distribution of $\tilde{\mathcal{T}}_n^{(2)}$. Recall that the height L_2 of vertex 2 is uniformly distributed on $\{0, \dots n-1\} \setminus \{\ell_1\}$. Now in the UA construction, define $\mathbb{T}$ as the tree obtained by starting from a root arrived at height L_2 and allowing the attachment of a vertex with height ℓ to $\mathbb{T}$ only if $\ell \notin \{\ell_1, \dots \ell_i\}$. Then, on the event A_n^{c}, this tree must coincide with $\mathcal{T}_n^{(2)}$, and so

$$\mathbb{P}\left(A_n^{\mathrm{c}}, \tilde{T}_n^{(2)} \geqslant k_n \mid B\right) \;=\; \mathbb{P}\left(A_n^{\mathrm{c}}, |\mathbb{T}| \geqslant k_n \mid B\right).$$

From the UA construction, for any $\ell \notin \{\ell_1, \dots, \ell_i\}$, conditional on $B \cap \{L_2 = \ell\}$, we can describe $|\mathbb{T}|$ using a modified process Υ_n, which we denote by $(\widetilde{\Upsilon}_\ell(m),\, 0 \leqslant m \leqslant \ell)$, and define by

- $\widetilde{\Upsilon}_\ell(0) = 1$.
- For all $0 < m \leqslant \ell$, $\widetilde{\Upsilon}_\ell(m) - \widetilde{\Upsilon}_\ell(m-1) \in \{0, 1\}$ and, conditional on $\widetilde{\Upsilon}_\ell(m-1) = j$, $\widetilde{\Upsilon}_\ell(m) = j+1$ with probability

$$\begin{cases} \dfrac{j}{n-1-J_m} & \text{if } \ell - m \notin \{\ell_1, \dots, \ell_i\} \\ 0 & \text{if } \ell - m \in \{\ell_1, \dots, \ell_i\}, \end{cases}$$

where

$$J_m = |\{\ell_1, \dots, \ell_i\} \cap \{\ell - m, \dots, n\}|$$

is the number of vertices of $\tilde{T}_n^{(1)}$ with height greater than $\ell - m$ in the UA construction.

With this definition, for any $\ell \notin \{\ell_1, \dots, \ell_i\}$, conditional on $B \cap \{L_2 = \ell\}$, we have by construction $|\mathbb{T}| \overset{d}{=} \tilde{\Upsilon}_\ell(\ell)$. Now, note that the probability of increasing is always bounded by $j/(n-1-i)$. Therefore the modified process $\tilde{\Upsilon}_\ell$ can be coupled with Υ_{n-i} in such a way that, for all $0 \leqslant m \leqslant \ell < n - i$,

$$\tilde{\Upsilon}_\ell(m) \leqslant \Upsilon_{n-i}(m).$$

For $\ell \geqslant n - i$, we use instead the crude bound $\mathbb{P}(\tilde{\Upsilon}_\ell(\ell) \geqslant k_n) \leqslant \mathbb{E}(\tilde{\Upsilon}_\ell(\ell))/k_n$ in that case. Using the same the same reasoning as in Lemma 3.5.1, note that $\mathbb{E}(\tilde{\Upsilon}_\ell(\ell)) \leqslant (1 + \frac{1}{n-i-1})^{n-i-1} \leqslant e$. We thus get

$$\begin{aligned}
\mathbb{P}(A_n^{\mathrm{c}}, |\mathbb{T}| \geqslant k_n \mid B) &\leqslant \mathbb{P}(L_2 \notin \{\ell_1, \dots, \ell_i\}, |\mathbb{T}| \geqslant k_n \mid B) && (3.7)\\
&= \frac{1}{n-1} \sum_{\substack{\ell=0 \\ \ell \notin \{\ell_1, \dots \ell_i\}}}^{n-1} \mathbb{P}(\tilde{\Upsilon}_\ell(\ell) \geqslant k_n), \\
&\leqslant \frac{ei}{k_n(n-1)} + \frac{1}{n-1} \sum_{\ell=0}^{n-i-1} \mathbb{P}(\Upsilon_{n-i}(\ell) \geqslant k_n) && (3.8)\\
&= \frac{ei}{k_n(n-1)} + \frac{n-i}{n-1} \mathbb{P}\left(\tilde{T}_{n-i} \geqslant k_n\right). && (3.9)
\end{aligned}$$

As this bound depends on the set $\{\ell_1, \dots, \ell_i\}$ only via its cardinality i, one can integrate with respect to the distribution of $\mathcal{T}_n^{(1)}$ to get

$$\mathbb{P}(A_n^c, |\mathbb{T}| \geqslant k_n \mid \tilde{T}_n^{(1)} = i) \leqslant \frac{ei}{k_n(n-1)} + \frac{n-i}{n-1} \mathbb{P}\left(\tilde{T}_{n-i} \geqslant k_n\right),$$

Finally, because $\Upsilon_{n-(i+1)}(\ell) \overset{d}{\geqslant} \Upsilon_{n-i}(\ell)$, the expression (3.8) (and therefore (3.9)) is nondecreasing in i, and we have

$$\begin{aligned}
&\mathbb{P}(A_n^{\mathrm{c}}, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n) \\
&\quad \leqslant \sum_{i \geqslant k_n} \mathbb{P}\left(\tilde{T}_n = i\right) \left(\frac{ei}{k_n(n-1)} + \frac{n-i}{n-1} \mathbb{P}\left(\tilde{T}_{n-i} \geqslant k_n\right) \right) \\
&\quad \leqslant \sum_{i=k_n}^{K_n} \mathbb{P}\left(\tilde{T}_n = i\right) \left(\frac{e\,K_n}{k_n(n-1)} + \frac{n-K_n}{n-1} \mathbb{P}\left(\tilde{T}_{n-K_n} \geqslant k_n\right) \right) && (3.10)\\
&\qquad + \sum_{i \geqslant K_n} \mathbb{P}\left(\tilde{T}_n = i\right) \left(\frac{ei}{k_n(n-1)} + \frac{n-i}{n-1} \mathbb{P}\left(\tilde{T}_{n-i} \geqslant k_n\right) \right), && (3.11)
\end{aligned}$$

for any sequence $K_n \geqslant k_n$. Letting $K_n \coloneqq \alpha(\log n)^{1+\varepsilon/2}$, we then show that (3.10) is asymptotically no greater than R_n^2, and that (3.11) is negligible compared to R_n^2. Indeed, (3.10) is bounded from above by

$$R_n \left(\frac{e\,K_n}{k_n(n-1)} + \frac{n-K_n}{n-1} \mathbb{P}\left(\tilde{T}_{n-K_n} \geqslant k_n\right) \right).$$

Now note that $\frac{e\,K_n}{k_n(n-1)} = O(\frac{(\log n)^{\varepsilon/2}}{n}) = o(R_n)$, and that since $n - K_n \sim n$, we have $k_n = o(\sqrt{n - K_n})$. Therefore, by Proposition 3.5.9, $\mathbb{P}\left(\tilde{T}_{n-K_n} \geqslant k_n\right) \sim R_n$. Finally, up to a multiplicative constant, (3.11) is bounded from above by

$$\mathbb{P}\left(\tilde{T}_n \geqslant K_n\right) = \Theta\Big(\frac{n^{-(\log n)^{\varepsilon/2}}}{K_n}\Big) = o(n^{-2}) = o(R_n^2).$$

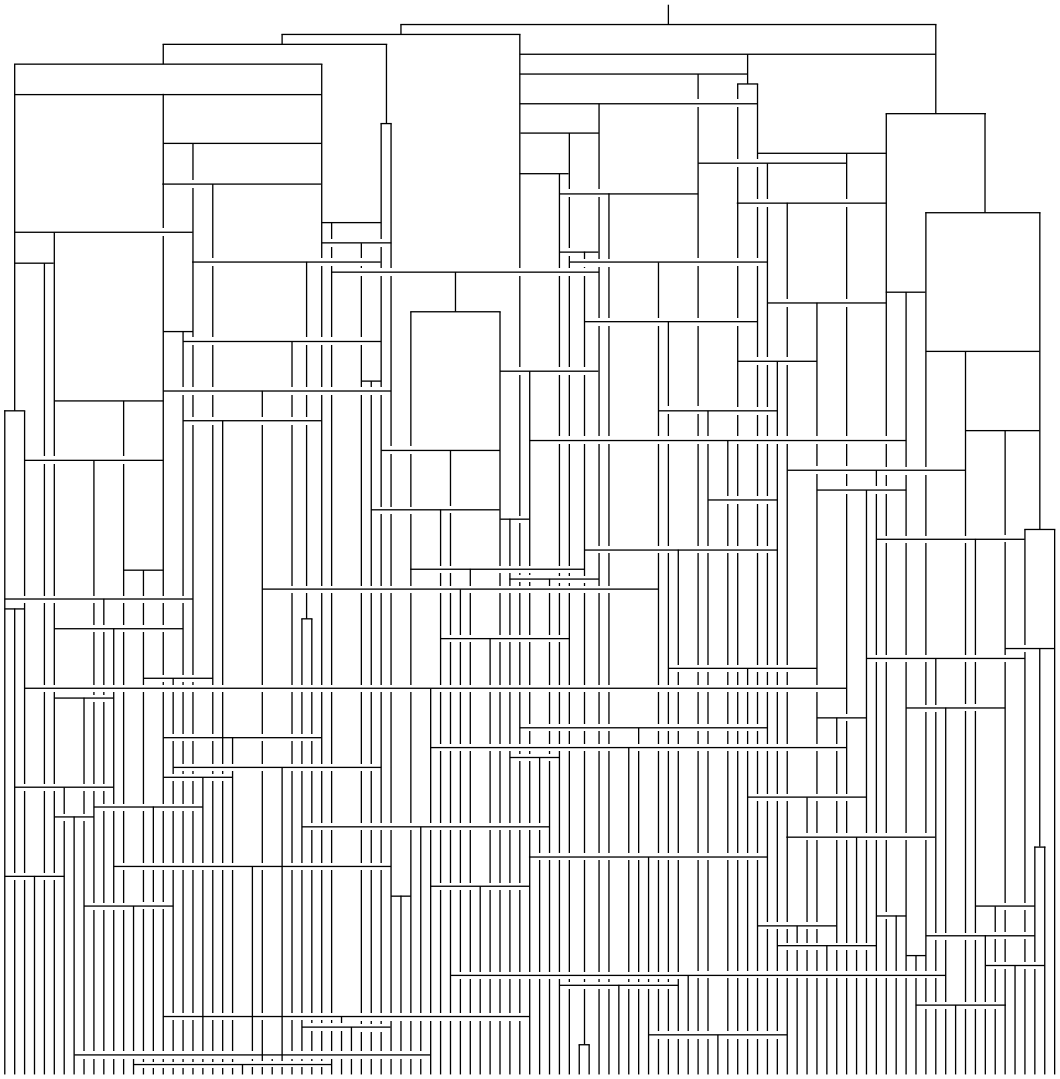
Putting everything together, we have proved that

$$\mathbb{P}\left(\tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \;\leqslant\; R_n^2 + o(R_n^2),$$

which concludes the proof. □

Chapter 4

Ranked tree-child networks



This project originated when Mike Steel and Charles Semple met with Amaury Lambert and told him about the difficulties posed by the enumeration of a class of phylogenetic networks known as tree-child networks. Amaury had the idea to rank the internal nodes of these networks in order to make them easier to study and generate via time-embedded processes, as is done for ranked trees. Nevertheless, the ranking that he used was in a way too "flexible" and the objects he studied remained quite complex.

I started working on the subject after meeting Mike with Amaury. I reused Amaury's idea to rank tree-child networks in order to make them more tractable, but used a different – more constraining yet biologically natural – notion of ranking. This made the resulting "ranked tree-child networks" more different from regular ones, but in a way more biologically relevant. This also made them much more tractable, to the point that I feared they might fall on the dull end of the spectrum. Fortunately, this also created connections with other combinatorial objects and some interesting questions remained. Mike gave me the opportunity to work on them with him in the University of Canterbury.

As of today, we are working on Conjecture 4.6.6 with Amaury and Mike to complete this project.

Chapter contents

4.1 Introduction . . . 88
4.1.1 Preliminaries . . . 88
4.1.2 Ranked tree-child networks . . . 89
4.1.3 Relation to other types of phylogenetic networks . . . 91
4.1.4 Main results . . . 91
4.2 Counting and generating RTCNs . . . 94
4.2.1 Backward-time construction of RTCNs . . . 94
4.2.2 Forward-time construction of RTCNs . . . 98
4.3 RTCNs and ranked trees . . . 100
4.3.1 Reticulation, unreticulation and base trees . . . 100
4.3.2 Uniform base trees of a uniform RTCN . . . 101
4.3.3 Sampling RTCNs conditional on displaying a tree . . . 102
4.4 Cherries and reticulated cherries . . . 104
4.4.1 Number of cherries . . . 105
4.4.2 Number of reticulated cherries . . . 107
4.5 Random paths between the roots and the leaves . . . 109
4.5.1 Length of a random walk from the root to a leaf . . . 110
4.5.2 Length of a random walk from a leaf to the root . . . 111
4.5.3 An alternative proof of Theorem 4.5.3 . . . 112
4.6 Number of lineages in the ancestry of a leaf . . . 114
4.6.1 Definition and characterization of $X^{(\ell)}$ . . . 114
4.6.2 Simulations . . . 116
4.6.3 The stochastic phase . . . 117
4.6.4 The deterministic phase . . . 119
Chapter references . . . 121
4.A Permutations and subexcedant sequences . . . 123
4.B Lemmas used in Section 4.4 . . . 125
4.C Variance of χ_ℓ . . . 126

4.1 Introduction

Tree-child networks are a class of directed acyclic graphs (DAGs) introduced by [4] as a way to model reticulated phylogenies (that is, phylogenies that take into account the possibility of hybridization or horizontal gene transfer). In addition to being biologically relevant, tree-child networks are mathematically interesting combinatorial structures and have thus gained attention recently to become one of the most studied classes of phylogenetic networks. However, they are also notoriously hard to study. For instance, their enumeration is still an open problem [10, 7] and there is no known algorithm to sample them uniformly (although a recursive procedure to enumerate them has recently been introduced [3]). As a result, very little is known about the properties of "typical" tree-child networks.

In this paper, we introduce a new class of phylogenetic networks that we term *ranked tree-child networks*, or RTCNs for short. These networks correspond to a subclass of tree-child networks that are endowed with an additional structure ensuring that they could have resulted from a time-embedded evolutionary process, something that is not required of tree-child networks.

Besides being arguably more biologically relevant than tree-child networks, one of the main advantages of RTCNs is that they are much easier to study. For instance, there are explicit formulas for the number of leaf-labeled RTCNs as well as simple procedures to sample them uniformly at random (or even uniformly at random subject to some natural constraints such as containing a fixed number of reticulations, or displaying a given tree). These make it possible to get some insight into the structure of uniform RTCNs.

4.1.1 Preliminaries

Let us start by recalling the definition of tree-child networks and introducing some vocabulary.

Definition 4.1.1. A *binary phylogenetic network* is a directed acyclic graph where each vertex has either

- in-degree 0 and out-degree 2 (the *root*)
- in-degree 1 and out-degree 0 (the *leaves*)
- in-degree 1 and out-degree 2 (*tree vertices*)
- in-degree 2 and out-degree 1 (*reticulation vertices*) ◇

If V is the vertex set of a binary phylogenetic network, we write ∂V for the set of its leaves. The vertices that are not leaves are called *internal vertices* and we denote their set by V°.

We refer to the elements of the set $\Gamma_{\text{in}}(v) = \{u : u \to v\}$ as the *parents* of v and to that of the set $\Gamma_{\text{out}}(v) = \{u : v \to u\}$ as the *children* of v. Two vertices are said to be *siblings* if they share a parent and *step-siblings* if they share a sibling.

Finally, an edge $\vec{uv}$ is called a *reticulation edge* if v is a reticulation vertex and a *tree edge* if v is a tree vertex or a leaf.

Definition 4.1.2. A *tree-child network* is a binary phylogenetic network such that every internal vertex has at least one child that is a tree vertex or a leaf. ◇

Note that there are other simple characterizations of tree-child networks. Consider for instance the following equivalent definition (see Lemma 2 in [4]).

Definition 4.1.3. A binary phylogenetic network is tree-child if and only if for every vertex v there exists a leaf such that every path going from the root to that leaf goes through v. ◇

4.1.2 Ranked tree-child networks

First, note that every DAG – and thus every tree-child network – is endowed with a partial order, which we refer to as the *genealogical order*, defined by

$$u \rightsquigarrow v \iff \text{there exists a directed path from } u \text{ to } v.$$

Let us now introduce the notion of *events* of a tree-child network.

Definition 4.1.4. Let N be a tree-child network. Define an equivalence relation $\mathcal{R}$ on the set of V° of internal vertices of N by

$$u \mathcal{R} v \iff u \text{ and } v \text{ are linked by a reticulation edge}.$$

The equivalence classes of $\mathcal{R}$ are called the *events* of N. Moreover, writing $\bar{u}$ for the equivalence class of a vertex u,

- either $\bar{u} = \{u\}$, in which case $\bar{u}$ is called a *branching event*;
- or $\bar{u} = \{u, v, w\}$, and $\bar{u}$ is called a *reticulation event*. ◇

Definition 4.1.5. A *ranked tree-child network* is an ordered pair $(N, \prec)$ where

- N is a tree-child network.
- The *chronological order* $\prec$ is a strict total order on the set of events of N that is compatible with the genealogical order – that is, for every internal vertices u and v,
$$u \rightsquigarrow v \implies \bar{u} \prec \bar{v} \text{ or } \bar{u} = \bar{v}.$$ ◇

Observe that in the case where N is a tree, Definition 4.1.5 agrees with the classical notion of the ranking of a tree: indeed, in that case every internal vertex is its own equivalence class and so the chronological order can be seen as a total strict order on V°.

Note that this ranking is very natural from a biological point of view. Indeed, real-world phylogenies are the end result of a time-embedded evolutionary process where lineages speciate and hybridize. In a tree-child network corresponding to a real-world phylogeny, each internal vertex can therefore be associated to one of these punctual evolutionary events. Now if to that vertex we associate a time-stamp t corresponding to the time at which the event occurred, then, under the assumption that no two events can occur simultaneously [1], by defining

$$\bar{u} = \bar{v} \iff t(u) = t(v) \quad \text{and} \quad \bar{u} \prec \bar{v} \iff t(u) < t(v)$$

[1] This means that the time-stamps of *events* are distinct; but the three vertices that form a reticulation will share the same time-stamp.

we obtain a valid partition into events and a valid ranking. Of course, this can only work if the tree-child network had been produced by a time-embedded process to start with; we will come back to this in a moment.

In other words, RTCNs are simply the combinatorial description of dated phylogenies, discarding the specifics of the times at which the events occurred but keeping the information about their relative order of occurrence.

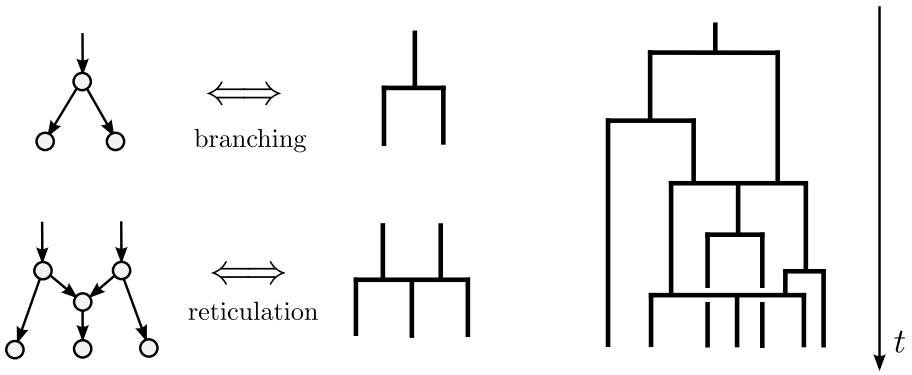


Figure 4.1: Graphical representation of a RTCN. Each "$\bot$" corresponds to a tree vertex (or the root), each "$\top$" to a reticulation vertex, and the tip of each dangling vertical line to a leaf. The vertical lines represent tree edges, and the horizontal ones events. Note that $a \prec b$ if and only if the horizontal line representing b is below that of a.

The notion of ranking raises two questions:

1. Can all tree-child networks be ranked?

2. How many rankings are there for a given tree-child network?

It is not hard to see that the answer to the first question is no. In fact, we will prove in the next section that almost no tree-child network can be ranked. While this rules out the possibility of using RTCNs to gain insight into the structure of tree-child networks, this does not make RTCNs irrelevant, because the tree-child networks that we can expect to see in nature should be rankable [2]. Thus, from a strictly biological point of view, rankable tree-child networks might be more relevant than tree-child networks.

Regarding the second question, Knuth proved in [8] that the number of ways to rank a tree is

$$\frac{2^{\kappa-s}\,(\ell-1)!}{\prod_{v\in V^\circ}(\lambda(v)-1)}\,,$$

where ℓ is the number of leaves of the tree, κ the number of cherries, s the number of symmetric vertices and $\lambda(v)$ the number of leaves subtended by v – that is, $\lambda(v) = \#\{u \in \partial V : v \rightsquigarrow u\}$. See e.g. Section 1.2.2 of [9]. However, the proof of this results relies on the recursive structure of trees, which tree-child networks lack. Thus, counting the number of ways to rank a given tree-child network remains an open question.

[2] This is not necessarily the case if some extinct lineages are not observed. But if we assume that the only possible evolutionary events are binary branchings and reticulations, then we can always assume that there exists an underlying ranked tree-child network.

4.1.3 Relation to other types of phylogenetic networks

Let us start by recalling the definition of temporal networks.

Definition 4.1.6. A binary phylogenetic network $N = (V, \vec{E})$ is *temporal* if it is possible to assign a time-stamp t to each vertex in such a way that

(i) If $\vec{uv}$ is a reticulation edge then $t(u) = t(v)$.

(ii) If $\vec{uv}$ is a tree edge then $t(u) < t(v)$. ◇

Every RTCN is a temporal network, because to get a valid temporal labeling t one can assign to every internal vertex the rank of the corresponding event (that is, $t(u) = k$ where $\bar{u}$ is the k-th event, and $t(\rho) = 0$ for the root). However, not all temporal tree-child networks are RTCNs because there is no requirement that the time-stamps of vertices that belong to different events be distinct in Definition 4.1.6.

Let us now recall the notion of normal network, as introduced by [16].

Definition 4.1.7. An edge $\vec{uv}$ is said to be *redundant* (or a *shortcut*) if there exists a directed path from u to v that does not contain $\vec{uv}$. A *normal network* is a tree-child network that has no redundant edges. ◇

It is well-known and not too hard to see that every temporal tree-child network is a normal network (see e.g. Proposition 10.12 in [13]). As a result, rankable tree-child networks are normal networks, in the sense that if $(N, \prec)$ is a RTCN then N is a normal network. Since by Theorem 1.4 in [10] the fraction of tree-child networks (leaf-labeled or vertex-labeled alike) that are also normal networks goes to zero as their number of vertices goes to infinity, this proves the next proposition.

Proposition 4.1.8. *The fraction of rankable tree-child networks with ℓ labeled leaves and the fraction of rankable tree-child networks with n labeled vertices both go to 0 as ℓ and n go to infinity.*

4.1.4 Main results

All the results presented in this paper are about leaf-labeled RTCNs.

In Section 4.2, we give two constructions of RTCNs: one in backward time and one in forward time. Each of these constructions yields a proof of the following result. Note that the number of reticulations of a RTCN with ℓ leaves and b branchings is $r = \ell - b - 1$.

Theorem 4.2.3. *The number of ranked tree-child networks with ℓ labeled leaves and b branchings is*

$$C_{\ell,b} = \begin{bmatrix} \ell - 1 \\ b \end{bmatrix} T_\ell \,,$$

where $T_\ell = \ell!\,(\ell-1)!/\,2^{\ell-1}$ is the number of ranked trees with ℓ labeled leaves and $\left[{\ell-1 \atop b}\right]$ is the number of permutations of $\{1, \ldots, \ell-1\}$ with b cycles (these quantities are known as the unsigned Stirling numbers of the first kind).

The backward-time and forward-time constructions also provide simple procedures to sample leaf-labeled RTCNs, be it:

- uniformly at random;
- uniformly at random conditional on their total number of reticulations;
- uniformly at random conditional on which events are reticulations.

The rest of our study focuses on the properties of uniform leaf-labeled RTCNs. One of their interesting characteristics is their intimate relation to uniform leaf-labeled trees. This is detailed in Section 4.3, where we also explain how to sample a uniform RTCN conditional on displaying a given tree.

Two of the most basic statistics of binary phylogenetic networks are their number of cherries and of reticulated cherries. While almost nothing is known about them in uniform tree-child networks, they prove very tractable in uniform RTCNs. Explicit expressions for their mean and variance are given in Section 4.4, where we also prove the following theorem.

Theorem 4.4.2. *Let κ_ℓ be the number of cherries of a uniform RTCN with ℓ labeled leaves and χ_ℓ be its number of reticulated cherries. Then, as $\ell \to \infty$,*

(i) $\kappa_\ell \xrightarrow{d} \text{Poisson}\left(\frac{1}{4}\right)$.

(ii) $\chi_\ell \,/\, \ell \xrightarrow{\mathbb{P}} \frac{1}{7}$.

Sections 4.5 and 4.6 contain our most informative results about the structure of uniform RTCNs: in Section 4.5 we study the length of some typical paths joining the root to the leaf set of uniform RTCNs and prove the following theorem.

Theorem 4.5.3. *Let ν be a uniform RTCN with ℓ labeled leaves and let*

- *$\gamma^{\downarrow}$ be the path taken by a random walk going from the root of ν to its leaves, respecting the direction of the edges.*
- *$\gamma^{\uparrow}$ be the path taken by a random walk going from a uniformly chosen leaf of ν to its root, following the edges in reverse direction.*

Then, letting length$(\cdot)$ *denote the length of these paths, not counting reticulation edges, there exist two constants $c^{\downarrow}$ and $c^{\uparrow}$ such that, as $\ell \to \infty$,*

(i) $\text{length}(\gamma^{\downarrow}) \approx \text{Poisson}(2\log\ell + c^{\downarrow})$

(ii) $\text{length}(\gamma^{\uparrow}) \approx \text{Poisson}(3\log\ell + c^{\uparrow})$

in the sense that the total variation distance between these distributions goes to 0.

Remark 4.1.9. Compare Theorem 4.5.3 with the analogous for uniform ranked trees, in which $\text{length}(\gamma^{\downarrow}) \approx \log\ell$ and $\text{length}(\gamma^{\uparrow}) \approx 2\log\ell$. ◇

Finally, in Section 4.6 we study the number of lineages in the ancestry of a leaf, i.e. in the subgraph consisting of all paths joining this leaf to the root, as illustrated in Figure 4.2. Starting from the leaves and going towards the root, one event of the RTCN after the other, this number of lineages will on average start by increasing but

will eventually decrease as we get nearer the root and the total number of lineages of the RTCN itself decreases.

Let $X^{(\ell)} = (X_k^{(\ell)}, 0 \leqslant k \leqslant \ell - 2)$ denote the process counting these lineages. What can we say about the behavior of $X^{(\ell)}$ as ℓ goes to infinity? As it turns out, $X_k^{(\ell)}$ remains quite small as long as $k = \lfloor x\ell \rfloor$ with $0 < x < 1$. However, when we reach the regime $k = \lfloor \ell - x\sqrt{\ell} \rfloor$, $X_k^{(\ell)}$ starts to increase substantially, until a point where it will decrease suddenly. Moreover, even though the properly rescaled trajectories of $X^{(\ell)}$ seem to become smooth as $\ell \to \infty$, they do not become deterministic. In fact, simulations presented in Section 4.6.2 suggest that they behave has a deterministic process with a random initial condition.

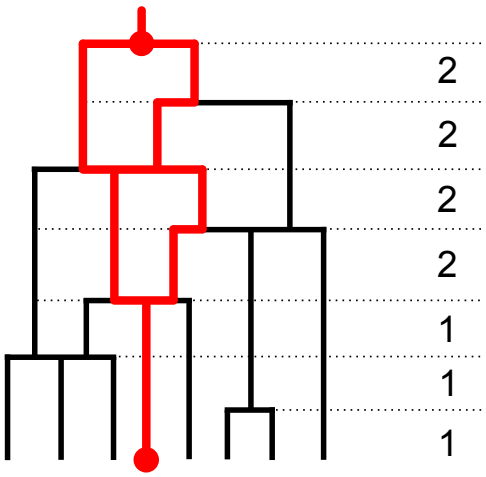


Figure 4.2: In red, the ancestry of a leaf and on the right the process counting the number of lineages in this ancestry.

Another natural way to study the asymptotics of $X^{(\ell)}$, instead of restricting ourselves to an appropriate time-window, is to consider the embedded process $\tilde{X}^{(\ell)}$, that is, to ignore the steps in which $X^{(\ell)}$ does not change. As illustrated in Figure 4.3, simulations suggest that that the scaling-limit of $\tilde{X}^{(\ell)}$ is a remarkably regular deterministic process with a random initial condition.

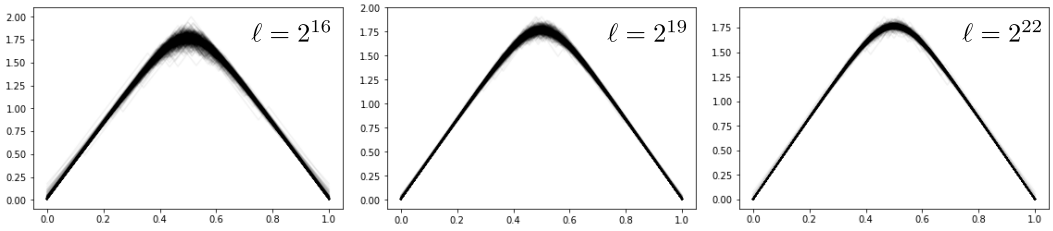


Figure 4.3: 500 superimposed trajectories of the process $\tilde{X}^{(\ell)}$, rescaled so that their domain is $[0, 1]$ and that the total area under each curve is equal to 1.

This part of our study is still work in progress, so our discussion will rely mostly on simulations and heuristics. Below are some of our results and conjectures so far.

Proposition 4.6.5. *For all $\ell \geqslant 2$ and k such that $0 \leqslant k \leqslant \ell - 2$,*

$$\frac{\ell}{\ell - k + 1}\left(1 - \frac{2k}{(\ell - k)(\ell - k - 1)}\right) \leqslant \mathbb{E}\left(X_k^{(\ell)}\right) \leqslant \frac{\ell}{\ell - k} .$$

As a result, for $M > 2$ and all ℓ large enough, for all $\varepsilon > 0$,

$$(1-\varepsilon)\frac{1}{M^2}\left(1-\frac{2}{M^2}\right) \leqslant \mathbb{E}\left(\tfrac{1}{\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell - M\sqrt{\ell}\rfloor}\right) \leqslant \frac{1}{M^2}\,,$$

so that the sequence of random variables $\frac{1}{M\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell - M\sqrt{\ell}\rfloor}$ is tight and bounded away from 0 in L^1.

Conjecture 4.6.6.

$$\tfrac{1}{M\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell - M\sqrt{\ell}\rfloor} \xrightarrow[\ell\to\infty]{d} W_M > 0 \qquad (\diamond)$$

Proposition 4.6.7. *If Conjecture 4.6.6 holds, then for all ε such that $0 < \varepsilon < 1$, as $\ell \to \infty$,*

$$\left(\tfrac{1}{M\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell - M\sqrt{\ell}(1-t)\rfloor},\ t\in[0,1-\varepsilon]\right) \implies \left(y(t,W_M),\ t\in[0,1-\varepsilon]\right)$$

where $\implies$ denotes convergence in distribution in the Skorokhod space and

$$y(t,W_M) \;=\; \frac{1-t}{C_M\cdot(1-t)^2+1}$$

where $C_M = W_M^{-1} - 1$.

4.2 Counting and generating RTCNs

4.2.1 Backward-time construction of RTCNs

Let us start by recalling that there is a simple way to label the internal vertices of any leaf-labeled tree-child network (or any leaf-labeled DAG), namely by labeling each internal vertex with the set of labels of its children.

Definition 4.2.1. Given a tree-child network with leaf set $\partial V = \{1,\dots,\ell\}$, the associated *canonical labeling* is the function ξ such that

(i) $\forall v \in \partial V$, $\xi(v) = v$.

(ii) $\forall v \in V^\circ$, $\xi(v) = \{\xi(u) : v \to u\}$. $\diamond$

While the canonical labeling of a tree-child network encodes it unambiguously (the whole network can be recovered from the label of the root), this is not the case for RTCNs because the information about the order of the events is missing. One way to retain this information is to encode RTCNs using the process $(P_k,\ 1 \leqslant k \leqslant \ell)$ defined as follows.

First, given a set $P = \{\xi_1,\dots,\xi_m\}$ of labels of vertices of $(N,\prec)$, define the two operations:

- $\mathrm{coal}(P,\{\xi_i,\xi_j\}) = (P\setminus\{\xi_i,\xi_j\}) \cup \{\{\xi_i,\xi_j\}\}$
- $\mathrm{ret}(P,\xi_i,\{\xi_j,\xi_k\}) = (P\setminus\{\xi_i,\xi_j,\xi_k\}) \cup \{\{\{\xi_i\},\xi_j\},\ \{\{\xi_i\},\xi_k\}\}$

Now, let $U_1 \prec \cdots \prec U_{\ell-1}$ denote the events of $(N, \prec)$. Then, starting from $P_1 = \{1, \ldots, \ell\}$ and going backwards in time, for $k = 1$ to $\ell - 1$:

- If $U_{\ell-k}$ is the coalescence of v and w, i.e. if there exists u such that $U_{\ell-k} = \{u\}$ and $\Gamma_{\text{out}}(u) = \{v, w\}$, then let $P_{k+1} = \text{coal}(P_k, \{\xi(v), \xi(w)\})$.
- If $U_{\ell-k}$ is the reticulation of u and v with the hybrid h, i.e. if $U_{\ell-k} = \{u', h', v'\}$ with $\Gamma_{\text{out}}(u') = \{u, h'\}$, $\Gamma_{\text{out}}(v') = \{v, h'\}$ and $\Gamma_{\text{out}}(h') = \{h\}$, then let $P_{k+1} = \text{ret}(P_k, \xi(h), \{\xi(u), \xi(v)\})$.

The result of procedure is illustrated in Figure 4.4. Note that the RTCN that produced a process $(P_k, 1 \leqslant k \leqslant \ell)$ can unambiguously be recovered from that process.

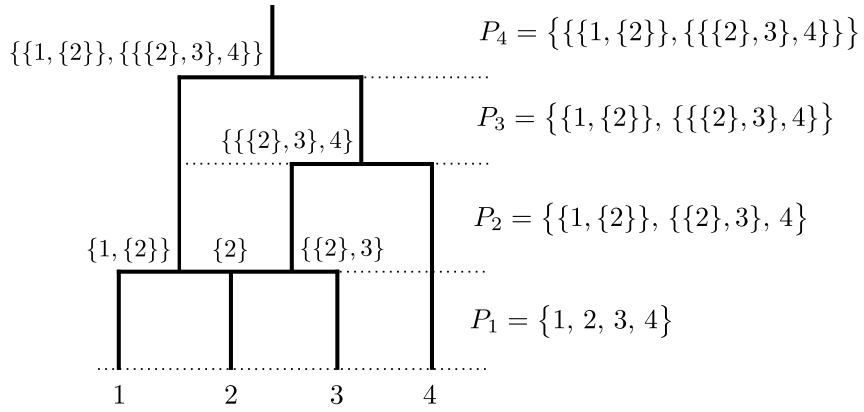


Figure 4.4: A RTCN with its canonical labeling and the associated process $(P_k, 1 \leqslant k \leqslant \ell)$.

In order to count RTCNs based on their number of reticulations (or, equivalently, branchings, since both are linked by $r + b = \ell - 1$), we need to introduce the notion of profile of a RTCN.

Definition 4.2.2. Let ν be a RTCN and $U_1 \prec \cdots \prec U_{\ell-1}$ its events. The *profile* of ν is the vector $\mathbf{q} = (q_1, \ldots, q_{\ell-1})$ defined by

$$q_k = \begin{cases} 1 \text{ if } U_k \text{ is a branching,} \\ 0 \text{ otherwise.} \end{cases} \qquad \diamond$$

We are now in position to prove our main result concerning the enumeration of RTCNs.

Theorem 4.2.3. *The number of RTCNs with profile* $\mathbf{q}$ *is*

$$F(\mathbf{q}) \;=\; \prod_{k=1}^{\ell-1} \left(q_k + (k-1)(1-q_k)\right) T_\ell \,,$$

where $T_\ell = \ell!\,(\ell-1)!\,/2^{\ell-1}$. *As a result, the number of RTCNs with* ℓ *labeled leaves and* b *branchings is*

$$C_{\ell,b} \;=\; \begin{bmatrix} \ell - 1 \\ b \end{bmatrix} T_\ell \,,$$

where the bracket denotes the unsigned Stirling numbers of the first kind.

Proof. Let us count the processes $(P_k,\, 1 \leqslant k \leqslant \ell)$ with profile $\mathbf{q}$. When going from $k+1$ to k lineages, i.e. when considering event U_k, there are:

- $\binom{k+1}{2}$ possible coalescence events;
- $(k-1)\binom{k+1}{2}$ possible reticulation events.

Therefore,

$$F(\mathbf{q}) \;=\; \prod_{k=1}^{\ell-1} \left(\binom{k+1}{2} q_k + (k-1)\binom{k+1}{2}(1-q_k) \right).$$

Factoring out the binomial coefficients, we get the first part of the theorem. Then, summing over all profiles with b branchings,

$$C_{\ell,b} \;=\; \left(\sum_{|\mathbf{q}|=b} \prod_{k=1}^{\ell-1} \left(q_k + (k-1)(1-q_k) \right) \right) \times T_\ell \,.$$

The first factor can be seen to be the coefficient of degree b of the polynomial $P(X) = (X+0)(X+1)\cdots(X+\ell-2)$. Since by definition of the unsigned Stirling numbers of the first kind,

$$P(X) = \sum_{b=0}^{\ell-1} \begin{bmatrix} \ell-1 \\ b \end{bmatrix} X^b \,,$$

this concludes the proof. □

Theorem 4.2.3 and its proof have several immediate corollaries. Let us start by pointing out the following procedure to sample uniform RTCNs conditional on a profile.

Corollary 4.2.4. *The following procedure yields a uniform RTCN with profile* $\mathbf{q}$*. Starting from* ℓ *labeled lineages, for* $k = \ell - 1$ *down to* 1*:*

- *If* $q_k = 1$*:* *let two lineages coalesce, uniformly at random among all* $\binom{k+1}{2}$ *possibilities.*
- *If* $q_k = 0$*:* *let three lineages reticulate, uniformly at random among all* $(k-1)\binom{k+1}{2}$ *possibilities.*

Second, let us lift the restriction on the number of reticulations.

Proposition 4.2.5. *For* $\ell \geqslant 2$*, the number of RTCNs with* ℓ *labeled leaves is*

$$C_\ell = \frac{\ell!\,(\ell-1)!^{\,2}}{2^{\ell-1}} \,.$$

Proposition 4.2.6. *Starting from* ℓ *labeled lineages, let pairs of lineages and triplets of lineages reticulate, choosing what to do uniformly among all possibilities at each step and stopping when there is only one lineage left. Then, the resulting RTCN has the uniform distribution on the set of RTCNs with* ℓ *labeled leaves.*

Proofs. Proposition 4.2.5 is obtained by recalling that $\sum_{b=0}^{\ell-1} \left[\begin{smallmatrix} \ell-1 \\ b \end{smallmatrix} \right] = (\ell-1)!$ and Proposition 4.2.6 by noting the realizations of the procedure correspond to those of the process $(P_k, 1 \leqslant k \leqslant \ell)$, which uniquely encodes every leaf-labeled RTCN. □

Finally, let us point out the following fact about the profile and the number of branchings of uniform RTCNs.

Corollary 4.2.7. *Let* $\mathbf{q}$ *be the profile of a uniform RTCN with* ℓ *labeled leaves. Then,*

$$\mathbf{q} \;\sim\; (X_1, \dots, X_{\ell-1})\,,$$

where $(X_1, \dots, X_{\ell-1})$ *are independent Bernoulli variables such that*

$$\mathbb{P}(X_k = 1) \;=\; \frac{1}{k}\,.$$

As a result, the number B_ℓ *of branchings of a uniform RTCN with* ℓ *leaves, which is distributed as the number of cycles of a uniform permutation of* $\{1, \dots, \ell-1\}$*, satisfies*

(i) $\mathbb{E}(B_\ell) = H_{\ell-1}$*, where* $H_{\ell-1} = \sum_{k=1}^{\ell-1} 1/k$ *is the* $(\ell-1)$*-th harmonic number.*

(ii) *As* $\ell \to \infty$*,* $d_{\mathrm{TV}}(B_\ell, \mathrm{Poisson}(H_{\ell-1})) \to 0$*. In particular,* $\frac{B_\ell - \log \ell}{\sqrt{\log \ell}} \xrightarrow{d} \mathcal{N}(0,1)$*.*

Proof. The first part of the proposition follows from the fact that the steps of the algorithm described in Proposition 4.2.6 are independent and that, when going from $k+1$ to k lineages there are $(k-1)\binom{k+1}{2}$ possible reticulations and $\binom{k+1}{2}$ possible coalescences, so that choosing uniformly among those the probability of picking a coalescence is $1/k$.

Points (i) and (ii) for B_ℓ are classic properties of the distribution of the number of cycles in a uniform permutation – see for instance Section 3.1 of [11] – that follow easily from its representation as a sum of independent Bernoulli variables. Indeed, (i) is immediate and for (ii) we can use the Stein-Chen bound on the total variation distance between a sum of independent Bernoulli variables and the corresponding Poisson distribution (recalled as Theorem B in Section 4.5) to get

$$d_{\mathrm{TV}}(B_\ell, \mathrm{Poisson}(H_{\ell-1})) \;\leqslant\; \min\{1, 1/H_{\ell-1}\} \sum_{k=1}^{\ell-1} \frac{1}{k^2} \;=\; O\Big((\log \ell)^{-1}\Big)\,,$$

from which the central limit theorem follows readily (see e.g. [2], page 17). □

Finally, let us close this section by pointing out an unexpected connection between RTCNs and a combinatorial structure known as river-crossings.

Remark 4.2.8. The number of RTCNs with ℓ labeled leaves is also the number of river-crossings using a two-person boat. It is recorded as sequence A167484 in the Online Encyclopedia of Integer Sequences [1], where it is described as follows:

> *For* ℓ *people on one side of a river, the number of ways they can all travel to the opposite side following the pattern of 2 sent, 1 returns, 2 sent, 1 returns, ..., 2 sent.*

However, there does not seem to be any natural bijection between river-crossings and RTCNs. Indeed,

1. RTCNs have a recursive structure that river-crossings lack.
2. For $\ell = 3$, the $C_3 = 6$ river-crossings are completely equivalent up to permutation of the labels, while the 6 RTCNs are not: 3 of them contain a reticulation while 3 of them don't. ◇

4.2.2 Forward-time construction of RTCNs

In this section, we give a forwards-in-time construction of RTCNs that will yield a second proof of Theorem 4.2.3. This proof is more combinatorial than the one we have already given, and will provide a different intuition as to why Stirling numbers arise in the enumeration of RTCNs.

The following notion of decorated RTCN will be useful.

Definition 4.2.9. A *decorated RTCN* is a pair (ν, θ), where

(i) ν is a RTCN with vertex set V and root ρ.

(ii) The *decoration* is a function $\theta : V \setminus \{\rho\} \to \{0, 1\}$ such that
- If v is a reticulation vertex or the child of a reticulation vertex, $\theta(v) = 0$.
- If u and v are siblings or step-siblings and none of them is a reticulation vertex, then $\theta(u) = 1 - \theta(v)$. ◇

Note that this formal definition is just a way to say that:
- For every tree vertex v, we distinguish one of the outgoing edges by assigning a "1" to one of the children of v.
- For every reticulation vertex v, we distinguish one of the incoming edges by assigning a "1" to one of the siblings of v (this unambiguously determines their common parent u and therefore the incoming edge $u\vec{v}$).

The notion of decorated RTCN is similar in spirit to that *ordered ranked tree*, where an ordering is specified for the children of each vertex, and indeed both notions are equivalent for trees. However, when working with RTCNs it is not useful to order the children of every vertex, in part because reticulated vertices already play a special role.

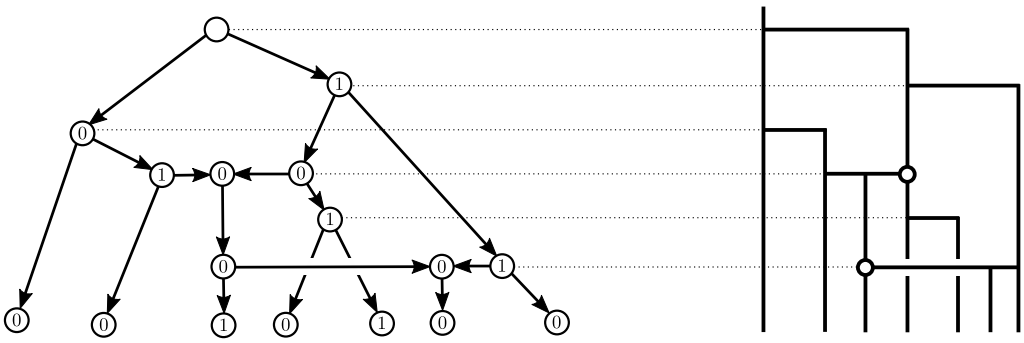


Figure 4.5: An example of decorated RTCN with its simplified representation on the right. For branchings, the decorated lineage is distinguished by drawing it to the right. For reticulations, it is distinguished by a white dot (but note that this white dot corresponds to which of the outgoing edges of the reticulation points to a 1 – not which vertex of the reticulation (if any) bears one.

An important characteristic of decorated RTCNs is that, unlike their undecorated counterparts, they are intrinsically labeled. Indeed, it is possible to assign a unique label to every vertex of a decorated RTCN (ν, θ) as follows: first, remove every reticulation edge distinguished by the decoration. This yields a tree $\tilde{\nu}$. Then, assign to each vertex v the label $\theta(u) \cdots \theta(v)$, where $(\rho, u, \ldots, v)$ is the unique path going from the root ρ to v in $\tilde{\nu}$. A consequence of this unique labeling of internal vertices is that there are $\ell!$ ways to label the leaves of a decorated RTCN with ℓ leaves.

Before describing the forward-time construction of decorated RTCNs, let us introduce one last combinatorial object.

Definition 4.2.10. A *subexcedant sequence of length n* is an integer-valued sequence $s = (s_1, \dots, s_n)$ such that, for all k, $1 \leqslant s_k \leqslant k$. For any two subexcedant sequences s and s', the *number of encounters of s and s'* is defined as

$$\mathrm{enc}(s, s') \;=\; \#\{k \geqslant 1 : s_k = s'_k\} \qquad \diamond$$

Lemma 4.2.11. *For any subsexcedant sequence s of length n, there are $\left[{n \atop k}\right]$ subexcedant sequences s' of length n such that* $\mathrm{enc}(s, s') = k$.

A combinatorial proof of this classic result is recalled in Section 4.A of the Appendix. However, it also follows immediately by considering a uniform subexcedant sequence s' and noting that its number of encounters with s is distributed as the sum for $k = 1$ to n of independent Bernoulli variables with parameters $1/k$.

Let us now describe how to encode decorated RTCNs using subexcedant sequences. Let s° and $s^\bullet$ be two subexcedant sequences of length $\ell - 1$. Start from a single lineage indexed "1" and, for $k = 1$ to $\ell - 1$:

- If $s^\circ_k = s^\bullet_k$, then let lineage s°_k branch to create lineage $k + 1$.
- If $s^\circ_k \neq s^\bullet_k$, then let lineages s°_k and $s^\bullet_k$ hybridize to form lineage $k + 1$.

At each step of this procedure, decorate the lineage s°_k with a white dot. This construction is illustrated in Figure 4.6.

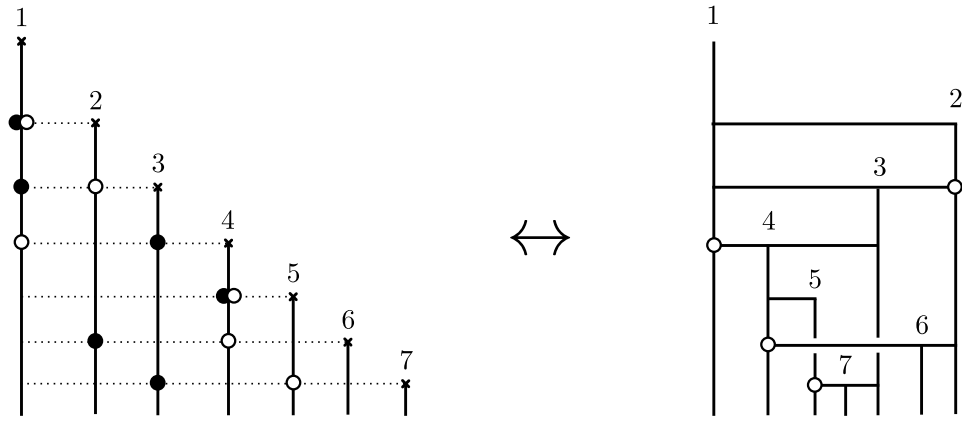


Figure 4.6: Graphical representation of the forward-time construction of RTCNs. On the left, the black and white dots represent the subexcedant sequences $s^\bullet$ and s°, respectively, and on the right the corresponding RTCN. At each step, the new lineage (marked by a cross) is linked to the black dot and the white dot. It these two dots fall on the same lineage, we get a branching.

Observe that in this construction every pair of subexcedant sequences $(s^\bullet, s^\circ)$ will yield a different decorated RTCN ν°, and that given a decorated RTCN ν° it is possible to unequivocally recover the subexcedant sequences $(s^\bullet, s^\circ)$ that produced it. Thus, letting $\mathscr{S}_{\ell-1}$ denote the set of subexcedant of length $\ell - 1$ and $\mathscr{C}^\circ_\ell$ that of decorated RTCNs with ℓ unlabeled leaves, this constructions gives a bijection between $\mathscr{S}_{\ell-1} \times \mathscr{S}_{\ell-1}$ and $\mathscr{C}^\circ_\ell$.

Also note that if at every step of the procedure we only link lineage $k + 1$ to lineage s°_k and decorate the latter with a white dot, we get a bijection between $\mathscr{S}_{\ell-1}$ and the set $\mathscr{T}^\circ_\ell$ of decorated ranked trees with ℓ unlabeled leaves.

From this forward-time encoding of decorated RTCNs and Lemma 4.2.11, we immediately get the following proposition.

Proposition 4.2.12. *There are* $\left[{\ell-1 \atop b}\right](\ell-1)!$ *decorated RTCNs with* ℓ *unlabeled leaves and* b *branchings.*

To recover Theorem 4.2.3 from Proposition 4.2.12, it suffices to recall that there are $\ell!$ ways to label the leaves of a decorated RTCN and to note that there are $2^{\ell-1}$ ways to decorate a RTCN.

Finally, let us close this section by pointing out that the following simple stochastic process generates uniform RTCNs.

Proposition 4.2.13. *Starting from a single lineage representing the root, let every lineage branch at rate 1 and every ordered pair of lineages hybridize at rate 1, decorating a lineage when it branches and decorating the first vertex of an ordered pair of lineages when it hybridizes. The RTCN obtained by stopping upon reaching* ℓ *lineages is uniform on the set of decorated RTCN with* ℓ *leaves. Relabeling its leaves uniformly at random and discarding its decoration yields a uniform RTCN with* ℓ *labeled leaves.*

4.3 RTCNs and ranked trees

The forward-time construction of the previous section gave us a way to encode a decorated RTCN using a decorated ranked tree and a subexcedant sequence. In fact, Theorem 4.2.3 shows that it is also possible to encode an *undecorated* leaf-labeled RTCN using a leaf-labeled ranked tree and, e.g., a subexcedant sequence or a permutation – even though we were unable to find any meaningful such encoding.

In this section, we specialize the results of the previous section in order to explain how to obtain RTCNs from ranked trees using simple graph operations, without having to explicitly manipulate subexcedant sequences. In particular, we give a simple way to sample a uniform RTCN conditional on displaying a given ranked tree. Readers who are more interested in the structural properties or RTCNs than in how to sample them can jump to the next section.

4.3.1 Reticulation, unreticulation and base trees

Let us recall that the classic graph operation known as the *vertex identification* of u and v consists in replacing u and v by a new vertex w whose neighbors (in-neighbors and out-neighbors, respectively) are exactly the neighbors of u and those of v (without introducing a self-loop in the case where u and v were neighbors). Conversely, the *cleaving* of $\vec{uv}$ consists in introducing an intermediate vertex w between u and v, that is: adding w to the graph, removing $\vec{uv}$ and adding $\vec{uw}$ and $\vec{wv}$.

Definition 4.3.1. The *unreticulation of a reticulation edge* $\vec{uv}$ is the graph operation consisting in

1. Removing $\vec{uv}$.
2. Identifying u and its (now only) child.
3. Identifying v and its child.

Conversely, the *reticulation of a tree edge e_1 into a tree edge e_2* consists in

1. Cleaving e_1 and e_2.
2. Adding an edge from the vertex introduced in the middle of e_1 to that introduced in the middle of e_2.

These operations are illustrated in Figure 4.7. ◇

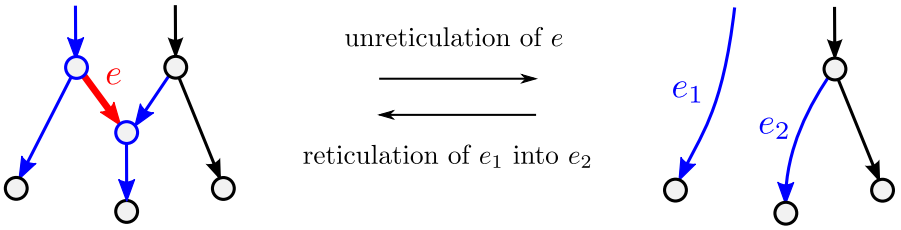


Figure 4.7: The operations of unreticulation and reticulation.

Unreticulating edges of a tree-child network N yields a tree N' with the same number of leaves as N as well as a partition into events and a genealogical order that are compatible with those of N. This justifies the following definition.

Definition 4.3.2. A ranked tree τ is a called a *base tree* of a RTCN ν if is possible to obtain it by unreticulating edges of ν. In that case, we write $\tau \sqsubset \nu$ and say that ν *displays* τ. ◇

Note that since each reticulation vertex has two incoming reticulating edges and that unreticulating each of these yields different RTCNs, every leaf-labeled RTCN with r reticulations displays exactly 2^r ranked trees.

4.3.2 Uniform base trees of a uniform RTCN

Proposition 4.3.3. *The ranked tree obtained by unreticulating one the of two incoming edges of each reticulation vertex of a uniform RTCN with ℓ labeled leaves, each with probability $1/2$, has the uniform distribution on the set of ranked trees with ℓ labeled leaves.*

Proof. Let ν be a uniform leaf-labeled RTCN. The procedure described in the proposition amounts to:

1. Decorating ν uniformly at random to obtain a decorated leaf-labeled RTCN ν°.
2. Unreticulating each of the undecorated reticulation edges of ν° to produce a decorated leaf-labeled tree τ°.
3. Discarding the decoration of τ°.

In the forward-time encoding, ν° corresponds to a unique triplet $\nu^\circ \simeq (s^\bullet, s^\circ, \sigma)$, where the permutation σ represents the leaf labeling. Now, since ν° is uniform on the set of decorated leaf-labeled RTCNs, (s°, σ) is also uniform and as a result so is $\tau^\circ \simeq (s^\circ, \sigma)$. Finally, the tree obtained by forgetting the decoration of a uniform decorated leaf-labeled ranked tree is uniform on the set of leaf-labeled ranked trees, concluding the proof. □

4.3.3 Sampling RTCNs conditional on displaying a tree

To obtain a RTCN from a ranked tree, we need to pay attention to the constraints imposed by the chronological order. Indeed, it is not possible to reticulate any tree edge of a RTCN into any other tree edge and obtain a RTCN. To formulate this restriction, we need to introduce the notion of contemporary edges.

Definition 4.3.4. Let us write $\bar{u} \preccurlyeq \bar{v}$ to indicate that $\bar{u} \prec \bar{v}$ or $\bar{u} = \bar{v}$, and convene that if u is a leaf then $\bar{u} = \partial V$ and $\bar{v} \prec \bar{u}$ for any internal vertex v.

The edge $\vec{uv}$ is said to be *alive between two events* $U \prec U'$ if

(i) It is a tree edge.

(ii) $\bar{u} \preccurlyeq U$ and $U' \preccurlyeq \bar{v}$.

Two edges are said to be *contemporary* if there exist two events such that they are both alive between these events. ◇

As for events, these definitions become very intuitive when using the graphical representation of RTCNs. Recall that, in this representation, we think of the vertical axis as time, and that tree edges correspond to vertical lines while events correspond to horizontal ones. An edge is alive between two events if a portion of it is located in the horizontal strip of the plane delimited by the two events. Two edges are contemporary if they overlap when projected on the vertical axis. This is illustrated in Figure 4.8.

Note that if we let $U_1 \prec \ldots \prec U_{\ell-1}$ denote the events of a RTCN with ℓ leaves, with the convention that $U_\ell = \partial V$, then there are exactly $k+1$ edges alive between U_k and U_{k+1}.

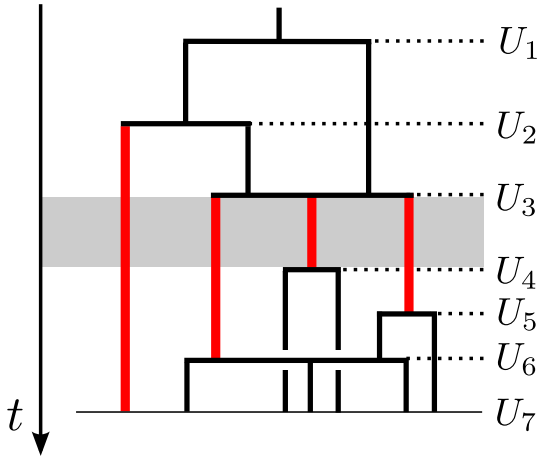


Figure 4.8: A RTCN with its events numbered in increasing order (and with the convention that $U_7 = \partial V$). The red edges are the four edges that are alive between events U_3 and U_4.

Recall the Definition 4.2.2 of a profile.

Proposition 4.3.5. *Let τ be a ranked tree with ℓ labeled leaves and let $\mathbf{q}$ be a profile. Letting $U_1 \prec \ldots \prec U_{\ell-1}$ denote the branchings of τ, for $k = 1$ to $\ell - 1$:*

- *If $q_k = 1$, do nothing.*
- *If $q_k = 0$, letting u denote the vertex such that $U_k = \{u\}$:*
  1. *Pick an edge e uniformly at random among the two outgoing edges of u.*
  2. *Pick an edge e' uniformly at random among the $k - 1$ edges that are contemporary with e and do not contain u.*
  3. *Reticulate e' into e.*

Then, resulting RTCN ν has the uniform distribution on the set of RTCNs with profile $\mathbf{q}$ displaying τ. Moreover, if τ is uniform then ν is uniform on the set of RTCNs with profile $\mathbf{q}$. If in addition $\mathbf{q}$ is distributed as the profile of a uniform RTCN (see Corollary 4.2.7), then the resulting RTCN is uniform on the set of RTCNs with ℓ labeled leaves.

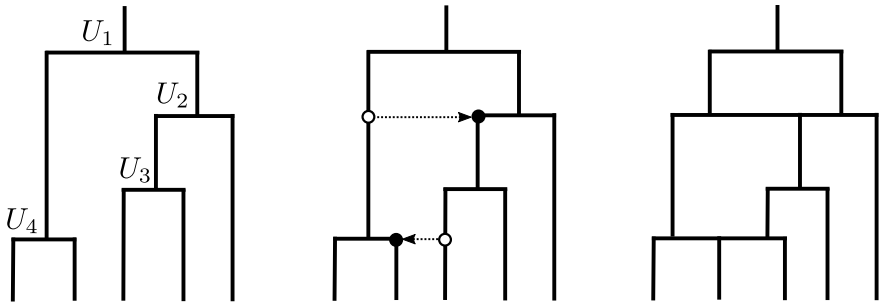


Figure 4.9: Example of the random construction of a RTCN from a ranked tree described in Proposition 4.3.5. Here, nothing happens for U_1 and U_3. For U_2 and U_4, the black dot represents the edge e and the white one the edge e'. Note that the modifications of branching events can be performed in any order, but that they have to be performed sequentially, so that contemporary edges remain well-defined at every step of the procedure.

Proof. The first part of the proposition follows from the fact (1) that for every fixed ranked tree θ and for every RTCN μ displaying θ there is exactly one sequence of modifications of θ that yields μ, and (2) that each possible realization of the procedure described in the proposition has the same probability, namely

$$\alpha(\mathbf{q}) = \prod_{k=1}^{\ell-1} \left(q_k + 2(k-1)(1-q_k)\right)^{-1},$$

although the exact value of this probability does not matter here. Thus, letting $Q(\mu)$ denote the profile of μ,

$$\mathbb{P}(\nu = \mu \,|\, \tau = \theta) \;=\; \alpha(\mathbf{q})\, \mathbb{1}_{\{Q(\mu)=\mathbf{q},\, \theta \sqsubset \mu\}}\,.$$

Now let us assume that τ is uniform. We have to show that for any RTCN μ,

$$\mathbb{P}(\nu = \mu) \;=\; \tfrac{1}{F(\mathbf{q})} \mathbb{1}_{\{Q(\mu)=\mathbf{q}\}}\,,$$

where $F(\mathbf{q})$ is the number of RTCNs with profile $\mathbf{q}$, whose exact value does not matter here. Since $\mathbb{P}(\tau = \theta) = 1/T_\ell$ for all fixed ranked tree θ,

$$\mathbb{P}(\nu = \mu) \;=\; \sum_{\theta \sqsubset \mu} \alpha(\mathbf{q}) \cdot \frac{1}{T_\ell} \cdot \mathbb{1}_{\{Q(\mu)=\mathbf{q}\}} \,.$$

Now let $r = \sum_{k=1}^{\ell-1}(1 - q_k)$ denote the number of reticulations of the profile $\mathbf{q}$. Since every RTCN with profile $\mathbf{q}$ displays exactly 2^r ranked trees, we have

$$\mathbb{P}(\nu = \mu) \;=\; \frac{2^r \alpha(\mathbf{q})}{T_\ell} \cdot \mathbb{1}_{\{Q(\mu)=\mathbf{q}\}} \,,$$

which concludes the proof since the factor $2^r\alpha(\mathbf{q})/T_\ell$ does not depend on μ. □

To close this section, note that to generate all RTCNs displaying a ranked tree it suffices to apply the procedure above without the restrictions of the profile, and that there are then $2(k-1)+1$ possible actions at step k of the procedure. This gives us the following proposition counting the number of RTCNs displaying a ranked tree.

Proposition 4.3.6. *The number of RTCNs displaying any ranked tree τ is*

$$\#\{\nu : \tau \sqsubset \nu\} \;=\; \prod_{k=1}^{\ell-1}(2k-1) \;=\; (2\ell-3)!!$$

Remark 4.3.7. Surprisingly, this is also the number of rooted unranked binary trees with ℓ labeled leaves – see e.g. [9] or [13]. While it is possible to give a bijective proof of this, we have not found a simple and visual bijection that would make this intuitive. ◇

4.4 Cherries and reticulated cherries

Cherries and reticulated cherries are among the most basic statistics of tree-child networks. In biological terms, a cherry is a pair of non-hybrid sibling species and a reticulated cherry is a group of three extant species such that one of these species was produced by the hybridization of the two others. These notions can be formalized as follows.

Definition 4.4.1. An event is said to be *external* when

$$\bigcup_{u \in U} \Gamma_{\text{out}}(u) \subset \partial V \,,$$

where $\Gamma_{\text{out}}(u)$ denotes the set of children of vertex u and ∂V is the leaf set of N.

- A *cherry* is an external branching event.
- A *reticulated cherry* is an external reticulation event.

These notions are illustrated in Figure 4.10 ◇

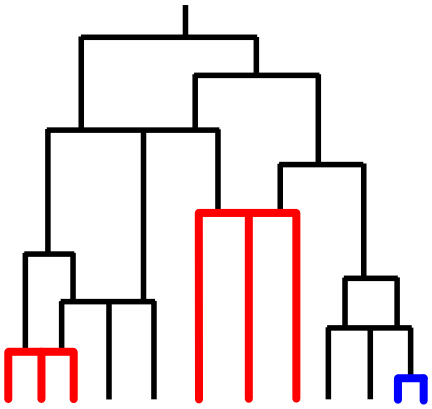


Figure 4.10: A RTCN with its only cherry (in blue) and its two reticulated cherries (in red).

4.4.1 Number of cherries

In this section, we prove the first part of what was announced in the introduction as Theorem 4.4.2.

Theorem 4.4.2 (Point (i))**.** *Let κ_ℓ be the number of cherries of a uniform RTCN with ℓ labeled leaves. Then,*

$$\kappa_\ell \xrightarrow[\ell\to\infty]{d} \text{Poisson}\left(\tfrac{1}{4}\right).$$

The proof consists in coupling κ_ℓ with a Markov chain to show that the factorial moments of κ_ℓ, that is, $m_\ell^n := \mathbb{E}(\kappa_\ell(\kappa_\ell-1)\cdots(\kappa_\ell-n+1))$, satisfy the recursion

$$m_{\ell+1}^n = \left(\frac{\ell-2n}{\ell}\right)^2 m_\ell^n + \frac{n(\ell-2n+2)}{\ell^2}\, m_\ell^{n-1},$$

and then using this recursion to prove that $m_\ell^n \to 1/4$ as $\ell \to \infty$ for all n. We separate each of these steps into different propositions.

Proposition 4.4.3. *Let $(X_\ell)_{\ell\geqslant 2}$ be the Markov chain defined by*

(i) $X_2 = 1$

(ii) *For $\ell \geqslant 2$, conditional on $X_\ell = k$,*

$$X_{\ell+1} = \begin{cases} k-2 & \text{with probability } \dfrac{2k(2k-2)}{\ell^2} \\ k-1 & \text{// } \dfrac{2k+4k(\ell-2k)}{\ell^2} \\ k & \text{// } \dfrac{(\ell-2k)(\ell-2k-1)+2k}{\ell^2} \\ k+1 & \text{// } \dfrac{\ell-2k}{\ell^2} \end{cases}$$

Then, for all $\ell \geqslant 2$, X_ℓ is distributed as the number κ_ℓ of cherries of a uniform RTCN with ℓ labeled leaves.

Proof. Let us use the forward-time construction to build a nested sequence $(\nu_\ell^\circ)_{\ell\geqslant 2}$ of uniform decorated RTCNs. Recall from Section 4.2.2 that to go from ν_ℓ° to $\nu_{\ell+1}^\circ$, we choose an ordered pair (u, v) of lineages ν_ℓ° uniformly at random. If $u = v$, we let the lineage branch; otherwise we let u and v hybridize.

Now, assume that ν_ℓ° has k cherries, so that there are $2k$ lineages that belong to cherries (which we refer to as C-leaves) and $\ell - 2k$ lineages that do not (F-leaves). Let us list all possible choices of (u, v) and see their effect on the number of cherries.

1. If the next event is a branching:
   (i) If a C-leaf is chosen, the number cherries does not change. This happens with probability $2k/\ell^2$.
   (ii) If a F-leaf is chosen, one cherry is created (probability: $(\ell - 2k)/\ell^2$).
2. If the next event is a hybridization:
   (i) If the two leaves of a same cherry are chosen, that cherry is destroyed (probability: $2k/\ell^2$).
   (ii) If two leaves of two different cherries are chosen, these two cherries are destroyed (probability: $2k(2k-2)/\ell^2$).
   (iii) If a C-leaf and a F-leaf are chosen, one cherry is destroyed (probability: $4k(\ell - 2k)/\ell^2$).
   (iv) If two F-leaves are chosen, the number of cherries does not change (probability: $(\ell - 2k)(\ell - 2k - 1)/\ell^2$).

Doing the book-keeping and observing that a RTCN with 2 leaves has one cherry concludes the proof. □

Notation 4.4.4. We denote by $x^{\underline{n}} = \prod_{k=0}^{n-1}(x-k)$ the n-th falling factorial of x. ⋄

Proposition 4.4.5. *Let $(X_\ell)_{\ell\geqslant 2}$ be the Markov chain defined in Proposition 4.4.3 and let $m_\ell^n = \mathbb{E}(X_\ell^{\underline{n}})$ denote the n-th factorial moment of X_ℓ. Then,*

$$m_{\ell+1}^n \;=\; \left(\frac{\ell-2n}{\ell}\right)^2 m_\ell^n \;+\; \frac{n(\ell-2n+2)}{\ell^2}\, m_\ell^{n-1}\,.$$

Proof. Let $p_{(-2)}, p_{(-1)}, p_{(+0)}$ and $p_{(+1)}$ denote the transition probabilities of X_ℓ, conditional on $X_\ell = k$. Then,

$$\begin{aligned}
&\mathbb{E}\Big(X_{\ell+1}^{\underline{n}} \,\Big|\, X_\ell = k\Big) \\
&\quad = (k-2)^{\underline{n}} p_{(-2)} \;+\; (k-1)^{\underline{n}} p_{(-1)} \;+\; k^{\underline{n}} p_{(+0)} \;+\; (k+1)^{\underline{n}} p_{(+1)} \\
&\quad = k^{\underline{n}} \left(\frac{(k-n)(k-n-1)}{k(k-1)} p_{(-2)} \;+\; \frac{k-n}{k} p_{(-1)} \;+\; p_{(+0)} \;+\; \frac{k+1}{k-n+1} p_{(+1)} \right) \\
&\quad = \frac{k^{\underline{n}}}{\ell^2} \Bigg(4(k-n)(k-n-1) \;+\; 2(k-n)(1+2(\ell-2k)) \\
&\qquad\qquad + \;(\ell-2k)(\ell-2k-1) + 2k + \frac{(k+1)(\ell-2k)}{k-n+1} \Bigg)
\end{aligned}$$

After a little algebra to rearrange this last expression, we get

$$\mathbb{E}\Big(X^n_{\ell+1} \,\Big|\, X_\ell = k\Big) \;=\; k^{\underline{n}} \left(\frac{\ell-2n}{\ell}\right)^2 \;+\; k^{\underline{n-1}}\, \frac{n(\ell+2-2n)}{\ell^2}$$

and the proposition follows by integrating in k. □

Proposition 4.4.6.

(i) *For all* $\ell \geqslant 2$, $\mathbb{E}(\kappa_\ell) = \frac{(3\ell-5)\ell}{12(\ell-1)(\ell-2)}$.

(ii) *For all* $n \geqslant 1$, $\mathbb{E}(\kappa^n_\ell) \to 1/4$ *as* $\ell \to \infty$. *As a result,* $\kappa_\ell \xrightarrow[\ell\to\infty]{d} \text{Poisson}(1/4)$.

Proof. The expression for $\mathbb{E}(\kappa_\ell)$ given in (i) follows from Lemma 4.B.1 and routine calculations. To prove (ii), we proceed by induction on n. First, we see from the expression in (i) that $m^1_\ell \to 1/4$ as $\ell \to \infty$. Now pick $n \geqslant 2$ and assume that $m^{n-1}_\ell \to 1/4^n$ as $\ell \to \infty$. By Proposition 4.4.5,

$$m^n_{\ell+1} \;=\; a_\ell\, m^n_\ell \;+\; b_\ell\,,$$

where

$$a_\ell = \left(\frac{\ell-2n}{\ell}\right)^2 \quad \text{and} \quad b_\ell = \frac{n(\ell-2n+2)}{\ell^2}\, m^{n-1}_\ell\,.$$

Since $a_\ell \neq 0$ for all $\ell \geqslant 2n+1$, and that

$$\prod_{j=2n+1}^{k} \left(\frac{j-2n}{j}\right)^2 \;=\; \frac{1}{\binom{k}{2n}^2}\,,$$

using Lemma 4.B.1 we get

$$m^n_\ell \;=\; \left(m^n_{2n+1} + \sum_{k=2n+1}^{\ell-1} \frac{n(k-2n+2)}{k^2} \binom{k}{2n}^2 m^{n-1}_k\right) \Big/ \binom{\ell-1}{2n}^2$$

Now, as $k \to \infty$, $\binom{k}{2n} \sim k^{2n}/(2n)!$ and, by the induction hypothesis, $m^{n-1}_k \sim 1/4^{n-1}$. As a result,

$$\frac{n(k-2n+2)}{k^2} \binom{k}{2n}^2 m^{n-1}_k \;\sim\; \frac{n/4^n}{(2n)!^2}\, k^{4n-1}\,.$$

Using Lemma 4.B.2 to get an asymptotic equivalent of the sum of these terms,

$$m^n_\ell \;\sim\; \frac{n/4^{n-1}}{(2n)!^2} \cdot \frac{\ell^{4n}}{4n} \cdot \left(\frac{(2n)!}{\ell^{2n}}\right)^2 \;=\; 1/4^n\,.$$

The convergence in distribution of X_ℓ to a Poisson distribution with mean $1/4$ is then a classic result (see e.g. Theorem 2.4 in [15]). □

4.4.2 Number of reticulated cherries

In this section, we prove the second part of Theorem 4.4.2.

Theorem 4.4.2 (Point (ii))**.** *Let* χ_ℓ *be the number of reticulated cherries of a uniform RTCN with* ℓ *labeled leaves. Then,*

$$\frac{\chi_\ell}{\ell} \;\xrightarrow[\ell\to\infty]{\mathbb{P}}\; \frac{1}{7}\,.$$

The proof is quite similar to that used in the previous section to study the number κ_ℓ of cherries – namely, we couple χ_ℓ with a Markov chain in order to compute its moments. The difference is that the moments of χ_ℓ are not as tractable as those of κ_ℓ. As a result, we only compute the first two moments explicitly and then use Chebyshev's inequality to prove the convergence in probability.

Proposition 4.4.7. *Let $(X_\ell)_{\ell\geqslant 2}$ be the Markov chain defined by*

(i) $X_2 = 0$

(ii) *For $\ell \geqslant 2$, conditional on $X_\ell = k$,*

$$X_{\ell+1} = \begin{cases} k-1 & \text{with probability} \quad \dfrac{3k\,(3k-2)}{\ell^2} \\ k & /\!/ \qquad\quad 1 - \dfrac{3k\,(3k-2) + (\ell-3k)(\ell-3k-1)}{\ell^2} \\ k+1 & /\!/ \qquad\quad \dfrac{(\ell-3k)(\ell-3k-1)}{\ell^2} \end{cases}$$

Then, for all $\ell \geqslant 1$, X_ℓ has the distribution of the number of reticulated cherries of a uniform RTCN with ℓ leaves.

Proof. As in the proof of Proposition 4.4.3, let us consider a nested sequence of uniform decorated RTCNs $(\nu_\ell^\circ)_{\ell\geqslant 2}$ produced by the forward-time construction, and see how the number of reticulated cherries is affected when we go from ν_ℓ° to $\nu_{\ell+1}^\circ$.

Assuming that there are k reticulated cherries in χ_ℓ, there are $3k$ leaves associated with reticulated cherries (RC-leaves) and $\ell - 3k$ other leaves (F-leaves). Now

1. If the next event is a branching:
   (i) If a F-leaf is chosen, the number of reticulated cherries does not change. This happens with probability $(\ell-3k)/\ell^2$.
   (ii) If a RC-leaf is chosen, the corresponding reticulated cherry is destroyed (probability: $3k/\ell^2$).
2. If the next event is a reticulation:
   (i) If two F-leaves are chosen, one reticulated cherry is created (probability: $(\ell-3k)(\ell-3k-1)/\ell^2$).
   (ii) If a F-leaf and a C-leaf are chosen, one reticulated cherry is destroyed and one is created (probability: $6k(\ell-3k)$).
   (iii) If two RC-leaves are chosen:
      a. If they do not belong to two different reticulated cherries, these are destroyed and a new reticulated cherry is created (probability: $3k(3k-3)/\ell^2$).
      b. If they belong to the same reticulated cherry, this reticulated cherry is destroyed and another one is created (probability: $6k/\ell^2$).

Doing the book-keeping and observing that a RTCN with 2 leaves has 0 reticulated cherries yields the proposition. □

Proposition 4.4.8. *The expected number $\mu_\ell = \mathbb{E}(\chi_\ell)$ of reticulated cherries of a uniform RTCN with ℓ leaves μ_ℓ satisfies the recursion*

$$\mu_{\ell+1} = \left(\frac{\ell-3}{\ell}\right)^2 \mu_\ell \; + \; \frac{\ell-1}{\ell}\,.$$

As a result, we have $\mu_2 = 0$, $\mu_3 = 1/2$, $\mu_4 = 2/3$ and, for $\ell \geqslant 4$,

$$\mu_\ell \;=\; \frac{(15\,\ell^3 - 85\,\ell^2 + 144\,\ell - 71)\,\ell}{105\,(\ell-1)(\ell-2)(\ell-3)}\,.$$

Proof. Using the Markov chain of Proposition 4.4.7, we have

$$\begin{aligned}\mathbb{E}(X_{\ell+1} \mid X_\ell = k) &= \; k \; + \; \mathbb{P}(X_{\ell+1} = k+1 \mid X_\ell = k) - \mathbb{P}(X_{\ell+1} = k-1 \mid X_\ell = k) \\ &= \; k\left(\frac{\ell-3}{\ell}\right)^2 \; + \; \frac{\ell-1}{\ell}\,,\end{aligned}$$

and the recursion follows by integrating against the law of X_ℓ.

The expression of μ_ℓ then follows from Lemma 4.B.1 and calculations that are better performed by a symbolic computation software such as [14]. □

Proposition 4.4.9. *The variance of the number of reticulated cherries of a uniform RTCN with ℓ leaves is*

$$\mathrm{Var}(\chi_\ell) \;=\; \frac{24}{637}\,\ell \; + \; \frac{1}{21} \; + \; o(1)\,.$$

The proof of this proposition is exactly the same as that of Proposition 4.4.8 but involves more complex expressions that can be found in Section 4.C of the Appendix.

Finally, the convergence in probability of χ_ℓ/ℓ to $1/7$ follows readily from Chebyshev's inequality and the fact that $\mathbb{E}(\chi_\ell) \sim \ell/7$ and $\mathrm{Var}(\chi_\ell) = O(\ell)$.

4.5 Random paths between the roots and the leaves

In this section, we study the length of two random paths going from the root to the leaf set:

1. A path obtained by starting from the root and going "down" towards the leaves, choosing each outgoing edge with equal probability whenever we reach a tree vertex.

2. A path obtained by starting from a uniformly chosen leaf and going "up" towards the root, choosing each incoming edge with equal probability whenever we reach a reticulation vertex.

Definition 4.5.1. The *length* of a path γ is its number of tree edges. ◇

The reason why we do not count reticulation edges when calculating the length of a path is that, from a biological point of view, reticulation edges are supposed to correspond to "instantaneous" hybridization events.

Before starting with the proofs, let us introduce some notation.

Notation 4.5.2. We denote by

$$H_n^{(m)} = \sum_{k=1}^{n} \frac{1}{k^m}$$

the n-th generalized harmonic number of order m. We also use the notation $\gamma = \lim_n H_n^{(1)} - \log n$ for the Euler-Mascheroni constant. ◇

Finally, let us recall a classic bound on the total variation distance between a sum of independent Bernoulli variables and the corresponding Poisson distribution.

Theorem B. *Let $X_1, \dots, X_n$ be independent Bernoulli variables with parameters $\mathbb{P}(X_i = 1) = p_i$, and let $\lambda_n = \sum_{i=1}^n p_i$. Then,*

$$d_{\mathrm{TV}}\left(\sum_{i=1}^{n} X_i,\ \mathrm{Poisson}(\lambda_n)\right) \leqslant \min(1, 1/\lambda_n) \sum_{i=1}^{n} p_i^2 \,,$$

where d_{TV} denotes the total variation distance.

This inequality is a consequence of the Stein-Chen method and can be found, e.g. as Theorem 4.6 in [12].

4.5.1 Length of a random walk from the root to a leaf

In this section, we prove the first part of what was announced as Theorem 4.5.3 in the introduction.

Theorem 4.5.3 (Point (i))**.** *Let ν be a uniform RTCN with ℓ leaves, and let $\gamma^{\downarrow}$ be a random path obtained by starting from the root and following the edges of ν, choosing each of the two out-going edges of a tree vertex with equal probability and stopping when we reach a leaf. Then,*

$$\mathrm{length}(\gamma^{\downarrow}) \ = \ \sum_{k=1}^{\ell-1} I_k \,,$$

where $I_1, \dots, I_{\ell-1}$ are independent Bernoulli variables with parameter

$$\mathbb{P}(I_k = 1) = \frac{2k-1}{k^2} \,.$$

In particular, letting $c^{\downarrow} = 2\gamma - \pi^2/6$, where γ is the Euler-Mascheroni constant,

(i) $\mathbb{E}\Big(\mathrm{length}(\gamma^{\downarrow})\Big) = 2\log \ell + c^{\downarrow} + o(1).$

(ii) $\mathrm{Var}\Big(\mathrm{length}(\gamma^{\downarrow})\Big) = 2\log \ell + O(1).$

(iii) $d_{\mathrm{TV}}(\mathrm{length}(\gamma^{\downarrow}), \mathrm{Poisson}(2\log \ell + c^{\downarrow})) \to 0.$

Proof. The idea of the proof is to use the forward-time construction to build jointly a nested sequence $(\nu_\ell^\circ)_{\ell \geqslant 2}$ of uniform decorated RTCNs and the random path $\gamma^{\downarrow}$. With the convention that ν_1° consists of a single lineage, for $k \geqslant 2$ let (u_k, v_k) denote the pair of lineages that was chosen to turn ν_{k-1}° into ν_k° (recall that if $u_k = v_k$ then

the next event is a branching) and let x_k record the position of the random walk among the leaves of ν°_{k-1}. With this notation, the length of $\gamma^\downarrow$ in ν°_ℓ is

$$\text{length}(\gamma^\downarrow) \;=\; \sum_{k=1}^{\ell-1} \mathbb{1}_{\{x_k \in \{u_k, v_k\}\}}\,,$$

where the variables $\mathbb{1}_{\{x_k \in \{u_k, v_k\}\}}$ are independent because (x_{k-1}, x_k) is independent of (u_k, v_k). Moreover, since (u_k, v_k) is chosen uniformly among the pairs of lineages of ν°_{k-1} and independently of x_k,

$$\mathbb{P}(x_k \in \{u_k, v_k\}) = \frac{2k-1}{k^2}\,,$$

which proves the first part of the proposition.

The rest of the proposition follows immediately from Theorem B since, letting $p_k = (2k-1)/k^2$,

- $\mathbb{E}\Big(\text{length}(\gamma^\downarrow)\Big) = \sum_{k=1}^{\ell-1} p_k = 2H^{(1)}_{\ell-1} - H^{(2)}_{\ell-1}$.
- $\text{Var}\Big(\text{length}(\gamma^\downarrow)\Big) = \sum_{k=1}^{\ell-1} p_k(1-p_k) = 2H^{(1)}_{\ell-1} - 5H^{(2)}_{\ell-1} + 4H^{(3)}_{\ell-1} - H^{(4)}_{\ell-1}$.
- $\sum_{k=1}^{\ell-1} p_k^2 = 4H^{(2)}_{\ell-1} - 4H^{(3)}_{\ell-1} + H^{(4)}_{\ell-1}$. □

4.5.2 Length of a random walk from a leaf to the root

In this section, we prove the second part of Theorem 4.5.3.

Theorem 4.5.3 (Point (ii)). *Let ν be a uniform RTCN with ℓ leaves, and let $\gamma^\uparrow$ be a random path obtained by starting from a uniformly chosen leaf and following the edges of ν in reverse direction, choosing each of the two incoming edges of a reticulation vertex with equal probability and stopping when we reach the root. Then,*

$$\text{length}(\gamma^\uparrow) \;=\; \sum_{k=2}^{\ell} J_k\,,$$

where $J_2, \dots, J_\ell$ are independent Bernoulli variables with parameter

$$\mathbb{P}(J_k = 1) = \frac{3k-4}{k(k-1)}\,.$$

In particular, letting $c^\uparrow = 3\gamma - 4$, where γ is the Euler-Mascheroni constant,

(i) $\mathbb{E}\Big(\text{length}(\gamma^\uparrow)\Big) = 3\log\ell + c^\uparrow + o(1)$.

(ii) $\text{Var}\Big(\text{length}(\gamma^\uparrow)\Big) = 3\log\ell + O(1)$.

(iii) $d_{\text{TV}}\big(\text{length}(\gamma^\uparrow), \text{Poisson}(3\log\ell + c^\uparrow)\big) \to 0$.

Remark 4.5.4. Note that the random path $\gamma^\uparrow$ is not uniformly chosen among all the paths going from the focal leaf to the root. ◇

Proof. The proof is similar to that of the previous section, but this time the idea is to use the backward-time construction to jointly build the RTCN ν and the random path $\gamma^{\uparrow}$. Recall that, in the backward-time construction, for $k = \ell$ down to 2, we go from k to $k-1$ lineages by choosing an event uniformly at random among the $k(k-1)/2$ possible coalescences and $k(k-1)(k-2)/2$ possible reticulations. Out of these, $k-1$ coalescences and $(k-1)(k-2)(k-3)/2$ reticulations involve the lineage through which $\gamma^{\uparrow}$ goes, and the choice is independent of the position of $\gamma^{\uparrow}$. As a result, the probability that the lineage containing $\gamma^{\uparrow}$ is involved in the event that is chosen is

$$\frac{k-1+3(k-1)(k-2)/2}{k(k-1)/2+k(k-1)(k-2)/2} = \frac{3k-4}{k(k-1)},$$

proving the first part of the proposition. The rest of the proposition then follows Theorem B and from the fact that, letting $p_k = \frac{3k-4}{k(k-1)}$,

- $\mathbb{E}\Big(\text{length}(\gamma^{\uparrow})\Big) = \sum_{k=2}^{\ell} p_k = 3H_{\ell}^{(1)} - 4 + 3/\ell.$
- $\text{Var}\Big(\text{length}(\gamma^{\uparrow})\Big) = \sum_{k=2}^{\ell} p_k(1-p_k) = 3H_{\ell}^{(1)} + 20 - 17H_{\ell}^{(2)} - 7/\ell + 1/\ell^2.$
- $\sum_{k=2}^{\ell} p_k^2 = -17H_{\ell}^{(2)} + 24 - 10/\ell + 1/\ell^2.$ □

4.5.3 An alternative proof of Theorem 4.5.3

In this section, we give another proof of Theorem 4.5.3. This proof is less direct than the previous one, but it will give another intuition as to where the Poisson distribution, the $\log \ell$ order of magnitude and the factors 2 and 3 come from.

Because writing down this proof formally would require introducing additional notation, and because we already have a formal proof, we allow ourselves to present it as a heuristic.

Let us start with $\gamma^{\downarrow}$. Slowing-down time in Proposition 4.2.13, consider the uniform decorated RTCN with ℓ leaves ν° obtained by:

1. Starting from one lineage.
2. Conditional on there being k lineages, letting:
   - each lineage branch at rate $1/k$;
   - each ordered pair of lineages hybridize at rate $1/k$.
3. Stopping upon reaching ℓ lineages.

Note that in this construction a branching event is viewed as the production of a new particle by another, rather than as the splitting of a particle into two new particles. Thus, we can consider the path $\widetilde{\gamma}^{\downarrow}$ obtained by always following the edge corresponding to the lineage of the first particle, as illustrated in Figure 4.11.

From the forward-time joint construction of ν° and $\gamma^{\downarrow}$, we see that the distribution of $\gamma^{\downarrow}$ does not depend on which lineage it chooses to follow, as long as this choice only depends on the past of the process. As a result,

$$\text{length}(\gamma^{\downarrow}) \overset{d}{=} \text{length}(\widetilde{\gamma}^{\downarrow}).$$

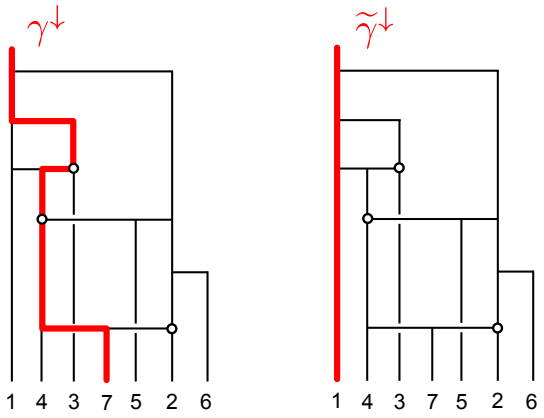


Figure 4.11: On the left, the path $\gamma^{\downarrow}$ and on the right the path $\tilde{\gamma}^{\downarrow}$ corresponding to the lineage of the first particle.

Now, length($\widetilde{\gamma}^{\downarrow}$) is simply the number of events affecting the lineage of the first particle, and the rate at which each given lineage is affected by events is $(k-1)\frac{2}{k}+\frac{1}{k} \approx 2$. Therefore, conditional on the time T it takes for the process to reach ℓ lineages,

$$\text{length}(\widetilde{\gamma}^{\downarrow}) \approx \text{Poisson}(2T)\,.$$

Finally, the total number of lineages increases by 1 at rate $k \cdot \frac{1}{k} + k(k-1) \cdot \frac{1}{k} = k$ and therefore follows a Yule process $(Y(t)\, t \geqslant 0)$. Since as $t \to \infty$, $Y(t)e^{-t} \to W$ almost surely, where W is a random variable (namely, an exponential variable with parameter 1), we see that the random time T it takes for the process to reach ℓ lineages is asymptotically

$$T \approx \log \ell - \log W\,.$$

Putting the pieces together, we see that length($\gamma^{\downarrow}$) $\approx$ Poisson($2 \log \ell$).

Let us now give a similar, forward-in-time construction of $\gamma^{\uparrow}$ where it can be identified with the lineage of the first particle. For this, we need to "straighten" $\gamma^{\uparrow}$ thanks to a set of deterministic rules telling us how to fix each bend, as illustrated in Figure 4.12. This yields a $\widetilde{\nu}^{\circ} = f(\nu^{\circ}, \gamma^{\uparrow})$ in which $\widetilde{\gamma}^{\uparrow}$, the image of $\gamma^{\uparrow}$, is the lineage of the first particle.

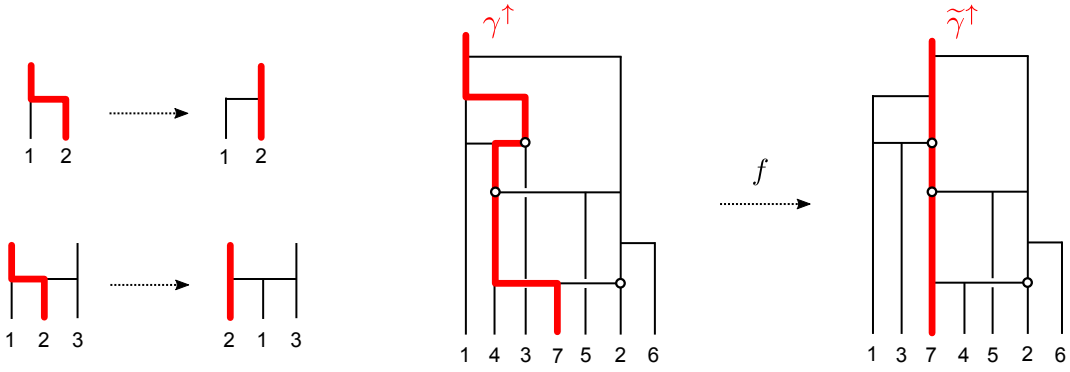


Figure 4.12: Illustration of the deterministic procedure used to "straighten" the path $\gamma^{\uparrow}$ in order to make it coincide with the lineage of the first particle. Left, the local modifications that are made each time a branching or a reticulation is encountered : these essentially consist in swapping lineages. Right, an example of application of the procedure to a RTCN.

Having done this, length($\gamma^{\uparrow}$) = length($\widetilde{\gamma}^{\uparrow}$) and, conditional on $\widetilde{\nu}^{\circ}$, the length of $\widetilde{\gamma}^{\uparrow}$ is the number of events affecting the lineage of the first particle.

Now, $\widetilde{\nu}^\circ$ has the same distribution as the decorated RTCN $\hat{\nu}^\circ$ generated by the following continuous-time Markov chain.

- Each lineage branches at rate $1/k$, except for that of the first particle, which branches at rate $2/k$.
- Each ordered pair of lineages hybridizes at rate $1/k$, except for pairs involving the lineage of the first particle, which hybridize at rate $3/(2k)$.

This is proved by writing down explicitly the laws $\widetilde{\nu}^\circ$ and $\hat{\nu}^\circ$, exactly as we did in Section 2.4 of Chapter 2 – something that requires introducing notation to give a formal description of $\widetilde{\nu}^\circ$. However, to see where the factors 2 and $3/2$ for the lineage of the first particle come from in the construction of $\hat{\nu}^\circ$, it suffices to note that when going from $k-1$ to k lineages the probability that the next event involves the lineage of the first particle is

$$\frac{3(k-2)+2}{3(k-2)+2+(k-2)(k-3)+(k-2)} = \frac{3k-4}{k(k-1)},$$

and so is indeed the same as the probability the lineage containing $\gamma^\uparrow$ is involved when going from k to $k-1$ lineages in the joint backward-time construction of ν° and $\gamma^\uparrow$ given in Section 4.5.2.

In the construction of $\hat{\nu}^\circ$, conditional on there being k lineages, events affect the lineage of the first particle at rate $2(k-1)\cdot\frac{3}{2k}+\frac{2}{k}\approx 3$ so that letting $\hat{T}$ denote the random time it takes for the process to reach ℓ lineages, $\mathrm{length}(\gamma^\uparrow)\approx\mathrm{Poisson}(3\hat{T})$. Finally, since the total number of lineages increases by 1 at rate $k+1$ when there are k lineages, the total number of lineages is distributed as $\hat{Y}(t)-1$, where $(\hat{Y}(t),\, t\geqslant 0)$ is a Yule process started from 2, so that $\hat{Y}(t)e^{-t}\to W+W'$, where W and W' are independent exponential variables with parameter 1. Therefore, $\hat{T}\approx\log\ell$ and we recover $\mathrm{length}(\gamma^\uparrow)\approx\mathrm{Poisson}(3\log\ell)$.

4.6 Number of lineages in the ancestry of a leaf

Let us start by giving a formal definition of the process counting the number of lineages in the ancestry of a leaf that was described in Section 4.1.4.

4.6.1 Definition and characterization of $X^{(\ell)}$

Definition 4.6.1. The *ancestry* of a vertex u of a RTCN ν is the subgraph $\nu_{|u}$ consisting of all paths going from the root of ν to u. ◇

Definition 4.6.2. The *number of lineages* of a subgraph μ of a RTCN ν is the process $(X_k,\, 0\leqslant k\leqslant \ell-2)$ defined by

$$X_k \;=\; \#\big\{e\in\mu : \text{the edge } e \text{ is alive between } U_k \text{ and } U_{k+1}\big\},$$

where $U_1\succ\cdots\succ U_{\ell-1}$ are the events of ν, taken in inverse chronological order, with the same convention as in Definition 4.3.4 that $U_0=\partial V$. ◇

In the rest of this section, we study the number of lineages in the ancestry of a uniformly chosen leaf of a uniform RTCN with ℓ labeled leaves, and denote it by $X^{(\ell)}$. See Figure 4.2 in Section 4.1.4 for an illustration. We also study the embedded

process $\tilde{X}^{(\ell)}$ defined by $\tilde{X}_i = X_{k_i}$, where $k_0 = 0$ and, for $i \geqslant 1$, $k_i = \inf\{k > k_{i-1} : X_k \neq X_{k_{i-1}}\}$.

Let us start by characterizing the law of $X^{(\ell)}$.

Proposition 4.6.3. *The process $X_k^{(\ell)}$ is the Markov chain characterized by $X_0^{(\ell)} = 1$ and the transition probabilities:*

- $\mathbb{P}\left(X_{k+1}^{(\ell)} = x+1 \,\middle|\, X_k^{(\ell)} = x\right) = \dfrac{x(\ell-k-x)(\ell-k-x-1)}{(\ell-k)(\ell-k-1)^2}$
- $\mathbb{P}\left(X_{k+1}^{(\ell)} = x-1 \,\middle|\, X_k^{(\ell)} = x\right) = \dfrac{x(x-1)^2}{(\ell-k)(\ell-k-1)^2}$
- $\mathbb{P}\left(X_{k+1}^{(\ell)} = x \,\middle|\, X_k^{(\ell)} = x\right) = 1 - \mathbb{P}\left(X_{k+1} = x \pm 1 \,\middle|\, X_k^{(\ell)} = x\right)$

Proof. The proof relies on the backward construction of a uniform RTCN and a bit of book-keeping to see how the $(k+1)$-th event, which takes us from $\ell-k$ lineages to $\ell-k-1$ lineages, affects the number of lineages in the ancestry of a leaf. Recall that in the backward construction there are $(\ell-k)(\ell-k-1)^2/2$ possibilities for the $(k+1)$-th event. Let us refer to the lineages in the ancestry of the focal leaf as marked lineages. Conditional on $X_k^{(\ell)} = x$, there are x marked lineages and $\ell-k-x$ unmarked lineages and thus there are:

- $x(x-1)/2$ possible coalescences between marked lineages. These decrease the number of lineages by 1.
- $x(x-1)(x-2)/2$ reticulations involving only marked lineages. These also decrease the number of lineages by 1.
- $x(\ell-k-x)(\ell-k-x-1)/2$ possible reticulations where the hybrid is a marked lineage and the other two lineages are unmarked. These increase the number of marked lineages by 1.

Other types of events (coalescences between two unmarked lineages, coalescences between a marked and an unmarked lineage, etc...) leave the number of marked lineages unchanged, as illustrated in Figure 4.13.

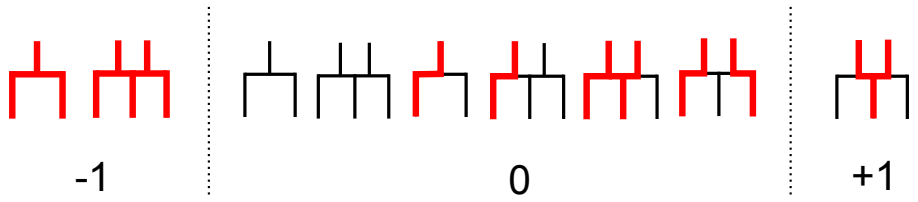


Figure 4.13: List of all the possible types of events and their effect on the number of marked lineages.

Since the event is chosen uniformly among all possibilities, this concludes the proof. □

4.6.2 Simulations

In this section, we present simulations supporting the conjectures about $X^{(\ell)}$ and $\tilde{X}^{(\ell)}$ made in Section 4.1.4 and outline some ideas to approach these conjectures. Let us start by looking at some individual trajectories of these processes, for increasing values of ℓ. As can be seen in Figure 4.14.A, most of the interesting behavior of $X^{(\ell)}$ seems to happen very close to the root so that to obtain a non-degenerate scaling-limit, we need to focus on a small window of time, for instance by considering $X^{(\ell)}_k$ for $k = \lfloor \ell - M\sqrt{\ell}(1-t) \rfloor$ and $t \in [0,1]$.

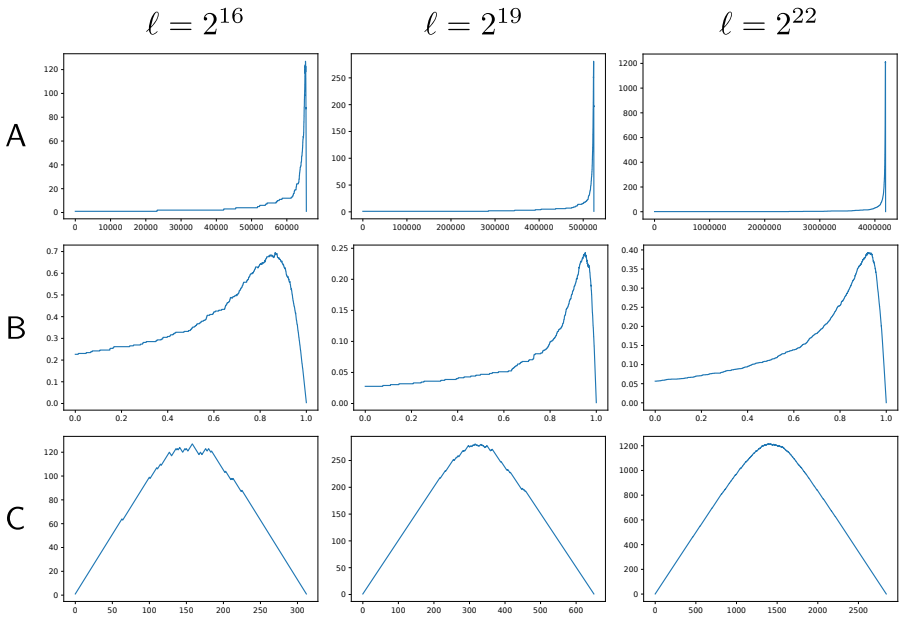


Figure 4.14: Individual trajectories of the processes described in the main text, for different values of ℓ. A, the process $X^{(\ell)}$, B the process $t \mapsto \frac{1}{\sqrt{\ell}} X^{(\ell)}_{\lfloor \ell M\sqrt{\ell}(1-t) \rfloor}$ and C, the process $\tilde{X}^{(\ell)}$.

Even though the trajectories represented in Figure 4.14 seem to become smooth as $\ell \to \infty$, they do not become deterministic, as made apparent by Figure 4.15, where the distributions of some statistics of $X^{(\ell)}$ are given. In particular, these simulations suggest that the relevant scaling limit for X^{ℓ} is indeed $\frac{1}{\sqrt{\ell}} X^{(\ell)}_{\lfloor \ell - M\sqrt{\ell}(1-t) \rfloor}$, and support Conjecture 4.6.6.

Our current idea to approach the study of the process $X^{(\ell)}$ is to separate it into two phases:

1. A slow, stochastic phase, where up to time $k = \lfloor \ell - M\sqrt{\ell} \rfloor$ the process $X^{(\ell)}_k$ remains relatively small and highly stochastic

2. A fast, deterministic phase, during which the internal dynamics of the rescaled process become deterministic, but retain a trace of the stochasticity of the first phase in the form of random initial conditions.

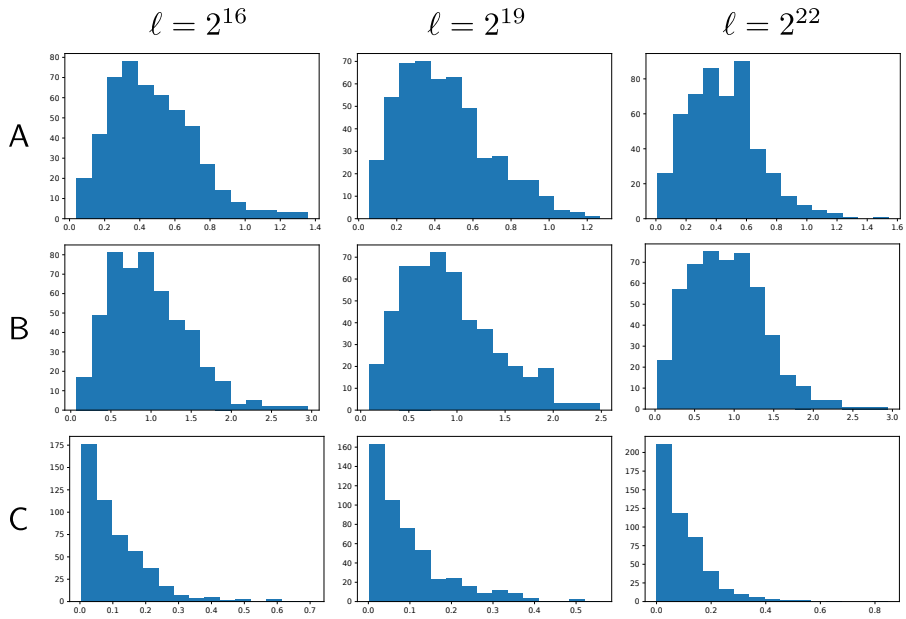


Figure 4.15: Distribution of some statistics of $X^{(\ell)}$ for $M = 10$ and 500 trajectories. A, $\frac{1}{\sqrt{\ell}} \max\{X_k^{(\ell)} : 0 \leqslant k \leqslant \ell - 2\}$; B, $\frac{1}{\sqrt{\ell}}\Big(\ell - \operatorname{argmax}\{X_k^{(\ell)} : 0 \leqslant k \leqslant \ell - 2\}\Big)$ and C, $\frac{1}{\sqrt{\ell}} X^{(\ell)}_{\lfloor \ell - M\sqrt{\ell} \rfloor}$

4.6.3 The stochastic phase

In this section, to avoid clutter we will sometimes drop the superscript in $X^{(\ell)}$.

Let us start by considering the process $Z = (Z_k, 0 \leqslant k \leqslant \ell - 2)$ characterized by $Z_0 = 1$, $Z_{k+1} - Z_k \in \{0, 1\}$ and

$$\mathbb{P}(Z_{k+1} = z + 1 \,|\, Z_k = z) \;=\; \frac{z}{\ell - k - 1}\,.$$

Comparing the transition probabilities of Z with those of X, we see that we can couple these two processes in such a way that $X_k \leqslant Z_k$ for all k. Let us now compute the first moments of Z.

Proposition 4.6.4.

(i) $\mathbb{E}(Z_k) = \dfrac{\ell}{\ell - k}$

(ii) $\mathbb{E}\Big(Z_k^2\Big) = \dfrac{\ell(\ell + k + 1)}{(\ell - k)(\ell - k + 1)}$

Proof. We have

$$\mathbb{E}(Z_{k+1}) = \left(1 + \frac{1}{\ell - k - 1}\right) \mathbb{E}(Z_k)\,.$$

Using that $\mathbb{E}(Z_0) = 1$, we get

$$\mathbb{E}(Z_k) \;=\; \prod_{i=0}^{k-1} \left(1 + \frac{1}{\ell - i - 1}\right) \;=\; \frac{\ell}{\ell - k}\,,$$

proving (i). For (ii), we note that

$$\mathbb{E}\Big(Z_{k+1}^2 \,\Big|\, Z_k\Big) \;=\; \frac{Z_k}{\ell - k - 1} + \left(\frac{2}{\ell - k - 1} + 1\right) Z_k^2\,.$$

Taking expectation and substituting $\mathbb{E}(Z_k)$ by $\ell/(\ell-k)$, we get

$$\mathbb{E}\Big(Z_{k+1}^2\Big) \;=\; \frac{\ell}{(\ell-k)(\ell-k-1)} \;+\; \frac{\ell-k+1}{\ell-k-1}\,\mathbb{E}\Big(Z_k^2\Big)\,.$$

Multiplying both sides by $(\ell-k)(\ell-k-1)$,

$$(\ell-k)(\ell-(k+1))\,\mathbb{E}\Big(Z_{k+1}^2\Big) \;=\; \ell \;+\; (\ell-(k-1))(\ell-k)\,\mathbb{E}\Big(Z_k^2\Big)$$

and therefore

$$\mathbb{E}\Big(Z_k^2\Big) \;=\; \frac{\ell(\ell+k+1)}{(\ell-k)(\ell-k+1)}\,,$$

concluding the proof. □

Proposition 4.6.5. *For all $\ell \geqslant 2$ and k such that $0 \leqslant k \leqslant \ell-2$,*

$$\frac{\ell}{\ell-k+1}\left(1-\frac{2k}{(\ell-k)(\ell-k-1)}\right) \;\leqslant\; \mathbb{E}(X_k) \;\leqslant\; \frac{\ell}{\ell-k}\;.$$

As a result, for $M>2$ and all ℓ large enough, for all $\varepsilon>0$,

$$(1-\varepsilon)\frac{1}{M}\left(1-\frac{2}{M^2}\right) \;\leqslant\; \mathbb{E}\left(\tfrac{1}{\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell-M\sqrt{\ell}\rfloor}\right) \;\leqslant\; \frac{1}{M}\,,$$

so that the sequence of random variables $\frac{1}{\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell-M\sqrt{\ell}\rfloor}$ is tight and bounded away from 0 in L^1.

Proof. The upper bound on $\mathbb{E}(X_k)$ follows immediately from $X_k \geqslant Z_k$ and Proposition 4.6.4. For the lower bound, let us denote by $\mathcal{F}_k = \sigma(X_0,\dots,X_k;Z_0,\dots,Z_k)$ the filtration generated by Z and X. Then,

$$\begin{aligned}
\mathbb{E}(X_{k+1}\,|\,\mathcal{F}_k) \;&= X_k + \frac{X_k(\ell-k-X_k)(\ell-k-X_k-1)}{(\ell-k)(\ell-k-1)^2} - \frac{X_k(X_k-1)^2}{(\ell-k)(\ell-k-1)^2}\\
&=\; X_k\left(1+\frac{(\ell-k)(\ell-k-1)-1}{(\ell-k)(\ell-k-1)^2} - \frac{2(\ell-k)-3}{(\ell-k)(\ell-k-1)^2}\,X_k\right)\\
&\geqslant\; X_k\left(1+\frac{1}{\ell-k}\right) - \frac{2\,Z_k^2}{(\ell-k-1)^2}
\end{aligned}$$

Taking expectations,

$$\mathbb{E}(X_{k+1}) \;\geqslant\; \left(1+\frac{1}{\ell-k}\right)\mathbb{E}(X_k) - \frac{2}{(\ell-k-1)^2}\mathbb{E}\Big(Z_k^2\Big)\,,$$

and using Proposition 4.6.4 we get

$$\mathbb{E}(X_{k+1}) \;\geqslant\; \left(1+\frac{1}{\ell-k}\right)\mathbb{E}(X_k) - \frac{2\,\ell(\ell+k+1)}{(\ell-k+1)(\ell-k)(\ell-k-1)^2}\,.$$

Multiplying both side of the inequality by $(\ell-k)$, we get

$$(\ell-k)\,\mathbb{E}(X_{k+1}) \;\geqslant\; (\ell-(k-1))\,\mathbb{E}(X_k) - \frac{2\,\ell(\ell+k+1)}{(\ell-k+1)(\ell-k-1)^2}\,.$$

and as a result,

$$(\ell-(k-1))\,\mathbb{E}(X_k) \;\geqslant\; \ell+1-2\ell\sum_{i=0}^{k-1}\frac{\ell+i+1}{(\ell-i+1)(\ell-i-1)^2}\,.$$

Using the fact that

$$\frac{\ell+i+1}{(\ell-i+1)(\ell-i-1)^2} \leqslant \frac{\ell+i}{(\ell-i)(\ell-i-1)(\ell-i-2)}$$

and that

$$\sum_{i=0}^{k-1} \frac{\ell+i}{(\ell-i)(\ell-i-1)(\ell-i-2)} = \frac{k}{(\ell-k)(\ell-k-1)},$$

we get

$$\mathbb{E}(X_k) \geqslant \frac{\ell}{\ell-k+1}\left(1-\frac{2k}{(\ell-k)(\ell-k-1)}\right).$$

Finally, to taking $k = \lfloor \ell - M\sqrt{\ell} \rfloor$ in these inequalities we get

$$\frac{\ell}{\sqrt{\ell}(M\sqrt{\ell}+2)}\left(1-\frac{2(\ell-M\sqrt{\ell})}{M\sqrt{\ell}(M\sqrt{\ell}-1)}\right) \leqslant \frac{1}{\sqrt{\ell}}\mathbb{E}\left(X_{\lfloor \ell-M\sqrt{\ell}\rfloor}\right) \leqslant \frac{1}{M}$$

where the term on the left-hand side of the inequality goes to $\frac{1}{M}(1-\frac{2}{M^2})$ as $\ell \to \infty$. Finally, the tightness follows from Markov's inequality. □

Proposition 4.6.5 makes the following conjecture seem likely to hold.

Conjecture 4.6.6. *The sequence of random variables* $\frac{1}{\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell-M\sqrt{\ell}\rfloor}$ *converges in distribution to a positive random variable* W_M. ◇

However, we have so far been unable to prove it. The problem is that as soon as we get in the regime $k = \lfloor \ell - M\sqrt{\ell}(1-t) \rfloor$, $Z^{(\ell)}$ and $X^{(\ell)}$ start to differ significantly and therefore the coupling is not so useful.

A very natural idea would be to use the backward-time construction to couple $X^{(\ell)}$ and $X^{(\ell+1)}_k$, but a difficulty with this approach is that this coupling lacks continuity in the sense that, with probability $\Theta(\ell)$, $X^{(\ell)}_k$ and $X^{(\ell+1)}$ will differ by a factor $\Theta(\ell)$ for $k = \lfloor \ell - M\sqrt{\ell} \rfloor$.

The approach we are currently investigating consists in finding another tractable process than $Z^{(\ell)}$ with which to couple $X^{(\ell)}$.

4.6.4 The deterministic phase

Proposition 4.6.7. *If Conjecture 4.6.6 holds, that is, if*

$$\tfrac{1}{\sqrt{M\ell}}X^{(\ell)}_{\lfloor \ell-M\sqrt{\ell}\rfloor} \xrightarrow[\ell\to\infty]{d} W_M > 0$$

then, for all ε *such that* $0 < \varepsilon < 1$, *as* $\ell \to \infty$,

$$\left(\tfrac{1}{M\sqrt{\ell}}X^{(\ell)}_{\lfloor \ell-M\sqrt{\ell}(1-t)\rfloor},\ t \in [0,1-\varepsilon]\right) \implies \left(y(t,W_M),\ t \in [0,1-\varepsilon]\right)$$

where $\implies$ *denotes convergence in distribution in the Skorokhod space and*

$$y(t,W_M) = \frac{(1-t)M}{C_M \cdot (1-t)^2 + 1}$$

where $C_M = W_M^{-1} - 1$.

Proof. Let us write for convenience $M_\ell \coloneqq M\sqrt{\ell}$ and

$$\mathscr{T}_\ell = \tfrac{1}{M_\ell}\{0, \dots, \ell - 2 - \lfloor \ell - M_\ell \rfloor\}\,.$$

Define the Markov chain $(Y_t^{(\ell)},\, t \in \mathscr{T}_\ell)$ taking values in $\frac{1}{M_\ell}\mathbb{N}$ by

$$Y_t^{(\ell)} \;=\; \tfrac{1}{M_\ell} X^{(\ell)}_{\lfloor \ell - M_\ell \rfloor + tM_\ell}\,.$$

The Markov chain $Y^{(\ell)}$ has infinitesimal mean

$$\begin{aligned} b^{(\ell)}(y,t) \;&=\; M_\ell\, \mathbb{E}\Big(Y^{(\ell)}_{t+1/M_\ell} - y \;\Big|\; Y_t^{(\ell)} = y\Big) \\ &=\; \mathbb{E}\Big(X^{(\ell)}_{\lfloor \ell - M_\ell \rfloor + tM_\ell + 1} - yM_\ell \;\Big|\; X^{(\ell)}_{\lfloor \ell - M_\ell \rfloor + tM_\ell} = yM_\ell\Big) \\ &=\; \frac{yM_\ell(\ell - k - yM_\ell)(\ell - k - yM_\ell - 1)}{(\ell - k)(\ell - k - 1)^2} - \frac{yM_\ell(yM_\ell - 1)^2}{(\ell - k)(\ell - k - 1)^2} \end{aligned}$$

where $k = \lfloor \ell - M_\ell \rfloor + tM_\ell$. Let us show that, for any $R > 0$ and any $\varepsilon > 0$,

$$b^{(\ell)}(y,t) \xrightarrow[\ell \to \infty]{} \frac{y(1-t-y)^2}{(1-t)^3} - \frac{y^3}{(1-t)^3}\,,$$

uniformly in $(y,t) \in [0,R] \times [0, 1-\varepsilon]$. Let us write

$$\begin{cases} b_+^{(\ell)}(y,t) = \dfrac{yM_\ell(\ell - k - yM_\ell)(\ell - k - yM_\ell - 1)}{(\ell - k)(\ell - k - 1)^2} \\[2ex] b_-^{(\ell)}(y,t) = \dfrac{yM_\ell(yM_\ell - 1)^2}{(\ell - k)(\ell - k - 1)^2} \end{cases}$$

Using that $\ell - M_\ell \leqslant k \leqslant \ell - M_\ell + 1$, we get

$$\frac{y(1-t-y-2/M_\ell)^2}{(1-t)^3} \;\leqslant\; b_+^{(\ell)}(y,t) \;\leqslant\; \frac{y(1-t-y)^2}{(1-t-2/M_\ell)^3}$$

As a result,

$$\begin{aligned} b_+^{(\ell)}(y,t) - \frac{y(1-t-y)^2}{(1-t)^3} \;&\geqslant\; \frac{y}{(1-t)^3}\left(-\frac{4}{M_\ell}(1-t-y) + \frac{4}{M_\ell^2}\right) \\ &\geqslant\; -\frac{4R}{\varepsilon^3 M_\ell} + O(1/M_\ell^2) \end{aligned}$$

Similarly,

$$\begin{aligned} b_+^{(\ell)}(y,t) - \frac{y(1-t-y)^2}{(1-t)^3} \;&\leqslant\; \frac{y(1-t-y)^2}{(1-t)^3(1-t-2/M_\ell)^3}\Big((1-t)^3 - (1-t-2/M_\ell)^3\Big) \\ &\leqslant\; \frac{R(1+R)^2}{\varepsilon^3(\varepsilon - 2/M_\ell)^3}\left(\frac{6}{M_\ell} + O(1/M_\ell^2)\right). \end{aligned}$$

This proves the uniform convergence of $b_+^{(\ell)}(y,t)$. The uniform convergence of $b_-^{(\ell)}(y,t)$ is treated similarly.

Now, $Y^{(\ell)}$ has infinitesimal variance

$$\begin{aligned} a^{(\ell)}(y,t) \;&=\; M_\ell\, \mathbb{E}\left(\Big(Y^{(\ell)}_{t+1/M_\ell} - y\Big)^2 \;\middle|\; Y_t^{(\ell)} = y\right) \\ &=\; \frac{1}{M_\ell}\Big(b_+^{(\ell)}(y,t) + b_-^{(\ell)}(y,t)\Big)\,, \end{aligned}$$

which goes to zero uniformly in $(y,t) \in [0,R] \times [0,1-\varepsilon]$. Assuming that

$$\tfrac{1}{M_\ell} X^{(\ell)}_{\lfloor \ell - M_\ell \rfloor} \xrightarrow[\ell\to\infty]{d} W_M \,,$$

the convergence of the piecewise constant interpolation of $Y^{(\ell)}$ to the solution of the Cauchy problem

$$\begin{cases} \dfrac{dy}{dt} = \dfrac{y(1-t-y)^2}{(1-t)^3} - \dfrac{y^3}{(1-t)^3} \\ y(0) = W_M \end{cases}$$

follows from Corollary 4.2 of [6] (see for instance Chapter 8.7 of [5] for a more practical introduction). Note that in these references, the results are stated for time-homogeneous Markov chains. However, they are easily adapted to time-inhomogeneous ones by extending the state space with time in order to obtain a time-homogeneous process.

Finally, the function given in the Proposition is then readily checked to be the unique solution of that Cauchy problem, concluding the proof. □

To close this section, let us mention briefly that an idea to study the embedded process $\tilde{X}^{(\ell)}$ is to introduce the process $S^{(\ell)}$ that counts the jumps of $X^{(\ell)}$. Indeed, with this process,

$$\tilde{X}^{(\ell)}_i = X^{(\ell)}_{(S^{(\ell)})^{-1}(i)} \,.$$

As a result, if we can study the convergence

$$\tfrac{1}{M_\ell}\left(X^{(\ell)}_{\lfloor \ell - M_\ell \rfloor + tM_\ell},\ S^{(\ell)}_{\lfloor \ell - M_\ell \rfloor + tM_\ell}\right) \implies (y(t), s(t)) \,,$$

this would show that

$$\tfrac{1}{M_\ell} \tilde{X}^{(\ell)}_{\lfloor \ell - M_\ell(1-t) \rfloor} \implies x(s^{-1}(t)) \,,$$

and we might be able to take $M \to \infty$ to study the convergence of $\frac{1}{\sqrt{\ell}} \tilde{X}^{(\ell)}_{\lfloor \ell t \rfloor}$. For the moment however these are just ideas.

Literature cited in this chapter

[1] The On-Line Encyclopedia of Integer Sequences, published electronically at http://oeis.org, 2019.

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

[3] G. Cardona, J. C. Pons, and S. Céline. Generation of tree-child phylogenetic networks. *arXiv preprint arXiv:1902.09015*, 2019.

[4] G. Cardona, F. Rossello, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):552–569, 2009.

[5] R. Durrett. *Stochastic calculus: a practical introduction*. CRC Press, 1996.

[6] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2nd edition, 2005.

[7] M. Fuchs, B. Gittenberger, and M. Mansouri. Counting phylogenetic networks with few reticulation vertices: Tree-child and normal networks. 2018.

[8] D. E. Knuth. *The art of computer programming: sorting and searching*, volume 3. Addison-Wesley, 1997.

[9] A. Lambert. Probabilistic models for the (sub)tree(s) of life. *Brazilian Journal of Probability and Statistics*, 31(3):415–475, 2017.

[10] C. McDiarmid, C. Semple, and D. Welsh. Counting phylogenetic networks. *Annals of Combinatorics*, 19(1):205–224, 2015.

[11] J. Pitman. *Combinatorial Stochastic Processes: École d'été de probabilités de Saint-Flour XXXII-2002*. Springer-Verlag Berlin Heidelberg, 2006.

[12] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:201–293, 2011.

[13] M. Steel. *Phylogeny: discrete and random processes in evolution*. SIAM, 2016.

[14] W. A. Stein et al. *Sage Mathematics Software (Version 8.2)*. The Sage Development Team, 2019. http://www.sagemath.org.

[15] R. Van Der Hofstad. *Random graphs and complex networks*. Cambridge University Press, 2016.

[16] S. J. Willson. Unique determination of some homoplasies at hybridization events. *Bulletin of mathematical biology*, 69(5):1709–1725, 2007.

Appendices to Chapter 4

4.A Permutations and subexcedant sequences

In this section, we recall a few classic combinatorial results that can be used to obtain a bijective proof Lemma 4.2.11 concerning the cycles of permutations and encounters of subexcedant sequences. While these results are not essential to our study, they are interesting in their own right and help get a better intuition as to why Stirling numbers emerge in our results.

Let us start by recalling the definition of subexcedant sequences and introducing some notation.

Definition 4.2.10. A *subexcedant sequence of length* n is an integer-valued sequence $s = (s_1, \dots, s_n)$ such that, for all k, $1 \leqslant s_k \leqslant k$. We denote by

$$\mathscr{S}_n = \prod_{k=1}^{n} \{1, \dots, k\}$$

the set of subexcedant sequences of length n. For any two s and $s' \in \mathscr{S}_n$, the *number of encounters of* s *and* s' is defined as

$$\operatorname{enc}(s, s') \;=\; \#\{k \geqslant 1 : s_k = s'_k\}\,. \qquad \diamond$$

Remark 4.A.1. In the literature, subexcedant sequences frequently refer to integer sequences $(s_n)_{n\geqslant 1}$ such that $0 \leqslant s_n \leqslant n-1$. They are also known as *inversion sequences*. $\diamond$

Notation 4.A.2. We denote by $\mathfrak{S}_n$ the set of permutations of $\{1, \dots, n\}$. $\diamond$

Let us now describe a bijection $L\colon \mathfrak{S}_n \to \mathscr{S}_n$ known as the *Lehmer code*. It is based on a simple idea: to describe a permutation $\sigma = \sigma_1\sigma_2\cdots\sigma_n$, it suffices to know that:

- σ_1 is the i_1-th element of $\{1, \dots, n\}$
- σ_2 is the i_2-th element of $\{1, \dots, n\} \setminus \{\sigma_1\}$

  $\vdots$
- σ_k is the i_k-th element of $\{1, \dots, n\} \setminus \{\sigma_1, \sigma_2, \dots, \sigma_{k-1}\}$

  $\vdots$
- σ_n is the i_n-th element of $\{1, \dots, n\} \setminus \{\sigma_1, \sigma_2, \dots, \sigma_{n-1}\}$

The sequence $s = i_n i_{n-1} \cdots i_2 i_1$ constitutes the Lehmer code $L(\sigma) \in \mathscr{S}_n$ of σ. Let us give a slightly more compact and formal definition.

Definition 4.A.3. The *Lehmer code* $L(\sigma)$ of a permutation σ is the sequence $s_1 s_2 \cdots s_n$ given by

$$s_k = \#\{j \leqslant k : \sigma_j \leqslant \sigma_k\}\,. \qquad \diamond$$

Note that, by definition of the Lehmer code, if $s = L(\sigma)$ then

$$s_k = k \iff \forall j < k,\ \sigma_j < \sigma_k\,.$$

that is, $s_k = k$ if and only if σ has a left-to-right maximum at k. As a result, get the following lemma.

Lemma 4.A.4. *Let $s^\star = 1\,2 \cdots n \in \mathscr{S}_n$. For all i,*

$$\#\{s \in \mathscr{S}_n : \mathrm{enc}(s, s^\star) = i\} \;=\; \#\{\sigma \in \mathfrak{S}_n : \sigma \textit{ has } i \textit{ left-to-right maxima}\}\,.$$

Let us now recall briefly why the number of permutations of $\{1, \ldots, n\}$ with i left-to-right maxima is the same as the number of permutations of $\{1, \ldots, n\}$ with i cycles. Observe that by:

(1) Writing permutations in canonical cycle notation, (that is, making cycles start by their largest element and ordering them in increasing order of their largest element).

(2) Dropping the parentheses delimiting cycles.

we obtain a bijection $F : \mathfrak{S}_n \to \mathfrak{S}_n$ – a result known as Foata's transition lemma. Moreover, the largest element of each cycle of σ correspond to a left-to-right maximum of $F(\sigma)$ and vice-versa. This proves the following lemma.

Lemma 4.A.5. *For all i,*

$$\#\{\sigma \in \mathfrak{S}_n : \sigma \textit{ has } i \textit{ left-to-right maxima}\} \;=\; \#\{\sigma \in \mathfrak{S}_n : \sigma \textit{ has } i \textit{ cycles}\}.$$

Finally, to obtain a proof of Lemma 4.2.11 it suffices to note that the number of subexcedant sequences that have i encounters with a fixed subexcedant sequence $s^\star$ does not depend on $s^\star$.

Lemma 4.A.6. *For any two $s^\star$ and $s \in \mathscr{S}_n$, for all i,*

$$\#\{s' \in \mathscr{S}_n : \mathrm{enc}(s', s^\star) = i\} \;=\; \#\{s' \in \mathscr{S}_n : \mathrm{enc}(s', s) = i\}$$

Proof. For any fixed pair of subexcedant sequences $s^\star$ and s, let $\Phi(s') \coloneqq s''$, where for all k

$$s''_k \coloneqq s'_k - s^\star_k + s_k \bmod \{1, \ldots, k\}\,,$$

with $i \bmod \{1, \ldots, k\}$ denoting the unique $j \in \{1, \ldots, k\}$ such that $j = i + mk$ for some $m \in \mathbb{Z}$. This yields a bijection $\Phi : \mathscr{S}_n \to \mathscr{S}_n$ such if $\Phi(s') = s''$ then for all k, $s'_k = s^\star_k$ if and only if $s''_k = s_k$. □

Combining all of these results proves Lemma 4.2.11:

Lemma 4.2.11. *For all $s \in \mathscr{S}_n$, for all i,*

$$\#\{s' \in \mathscr{S}_n : \mathrm{enc}(s, s') = i\} \;=\; \#\{\sigma \in \mathfrak{S}_n : \sigma \textit{ has } i \textit{ cycles}\}$$

4.B Lemmas used in Section 4.4

In this section, we recall the proof of two elementary lemmas that were used to study the number of cherries and of reticulated cherries in Section 4.4.

Lemma 4.B.1. *Let (u_ℓ) be a sequence satisfying the recursion*

$$u_{\ell+1} = a_\ell\, u_\ell + b_\ell$$

and let i be such that $\forall \ell \geqslant i$, $a_\ell \neq 0$. Then,

$$\forall \ell \geqslant i, \quad u_\ell = \left(u_i + \sum_{k=i}^{\ell-1} \frac{b_k}{\prod_{j=i}^{k} a_j} \right) \prod_{k=i}^{\ell-1} a_k$$

Proof. For $k \geqslant i$, since $a_k \neq 0$ we have

$$\frac{u_{k+1}}{\prod_{j=i}^{k} a_j} - \frac{u_k}{\prod_{j=i}^{k-1} a_j} = \frac{b_k}{\prod_{j=i}^{k} a_j},$$

where the empty product is one. As a result, for all $\ell \geqslant i$,

$$\frac{u_\ell}{\prod_{j=i}^{\ell-1} a_j} - u_i = \sum_{k=i}^{\ell-1} \left(\frac{u_{k+1}}{\prod_{j=i}^{k} a_j} - \frac{u_k}{\prod_{j=i}^{k-1} a_j} \right) = \sum_{k=i}^{\ell-1} \frac{b_k}{\prod_{j=i}^{k} a_j},$$

and the proof is over. □

Lemma 4.B.2. *Let (v_ℓ) be such that $v_\ell \sim \alpha\, \ell^{\,p}$, where $p \geqslant 0$ and $\alpha \neq 0$. Then,*

$$\sum_{k=i}^{\ell-1} v_k \sim \frac{\alpha}{p+1} \ell^{\,p+1}.$$

Proof. Let $\varepsilon_\ell \to 0$ be such that $v_\ell = \alpha\, \ell^{\,p} + \varepsilon_\ell\, \ell^{\,p}$. Then,

$$\sum_{k=i}^{\ell-1} v_k = \alpha \sum_{k=i}^{\ell-1} k^{\,p} + \sum_{k=i}^{\ell-1} \varepsilon_k\, k^{\,p}$$

Comparison with an integral shows that $\sum_{k=i}^{\ell-1} k^{\,p} \sim \ell^{\,p+1}/(p+1)$, so to finish the proof we simply have to show that $\sum_{k=i}^{\ell-1} \varepsilon_k\, k^{\,p} = o(\ell^{\,p+1})$. Since $\varepsilon_\ell \to 0$, for all $\eta > 0$ there exists j_η such that $\forall k \geqslant j_\eta$, $|\varepsilon_k| < \eta$. Therefore,

$$\left| \sum_{k=i}^{\ell-1} \varepsilon_k\, k^{\,p} \right| < \sum_{k=i}^{j_\eta - 1} |\varepsilon_k\, k^{\,p}| + \eta \sum_{k=j_\eta}^{\ell-1} k^{\,p}$$

For fixed η, the first of these sums has a fixed number of terms and thus is bounded. The second one is asymptotically equivalent to $\eta\, \ell^{\,p+1}/(p+1)$. Thus, for all $\eta > 0$, for ℓ large enough,

$$\left| \frac{\sum_{k=i}^{\ell-1} \varepsilon_k\, k^{\,p}}{\ell^{\,p+1}} \right| < \eta$$

and the proof is over. □

4.C Variance of χ_ℓ

In this section, we prove Proposition 4.4.9 by obtaining an explicit expression for the variance of the number χ_ℓ or reticulated cherries of a uniform RTCN with ℓ labeled leaves.

Consider the Markov chain $(X_\ell)_{\ell \geqslant 2}$ defined in Proposition 4.4.7, and let $s_\ell = \mathbb{E}(X_\ell^2)$. From the transition probabilities of X_ℓ, we get

$$\mathbb{E}\Big(X_{\ell+1}^2 \,\Big|\, X_\ell = k\Big) = \left(\frac{\ell-6}{\ell}\right)^2 k^2 \;+\; \frac{2\ell^2-\ell-3}{\ell^2}\, k \;+\; \frac{\ell-1}{\ell}\,.$$

Integrating in k, this yields

$$s_{\ell+1} = \left(\frac{\ell-6}{\ell}\right)^2 s_\ell \;+\; \frac{2\ell^2-8\ell-3}{\ell^2}\,\mu_\ell \;+\; \frac{\ell-1}{\ell}$$

where, for $\ell \geqslant 4$, we can substitute the expression of μ_ℓ given in Proposition 4.4.8. Rearranging a bit get that for all $\ell \geqslant 4$,

$$s_{\ell+1} = \left(\frac{\ell-6}{\ell}\right)^2 s_\ell \;+\; \frac{30\,\ell^5 - 185\,\ell^4 + 188\,\ell^3 + 746\,\ell^2 - 1649\,\ell + 843}{105\,(\ell-1)(\ell-2)(\ell-3)\ell}$$

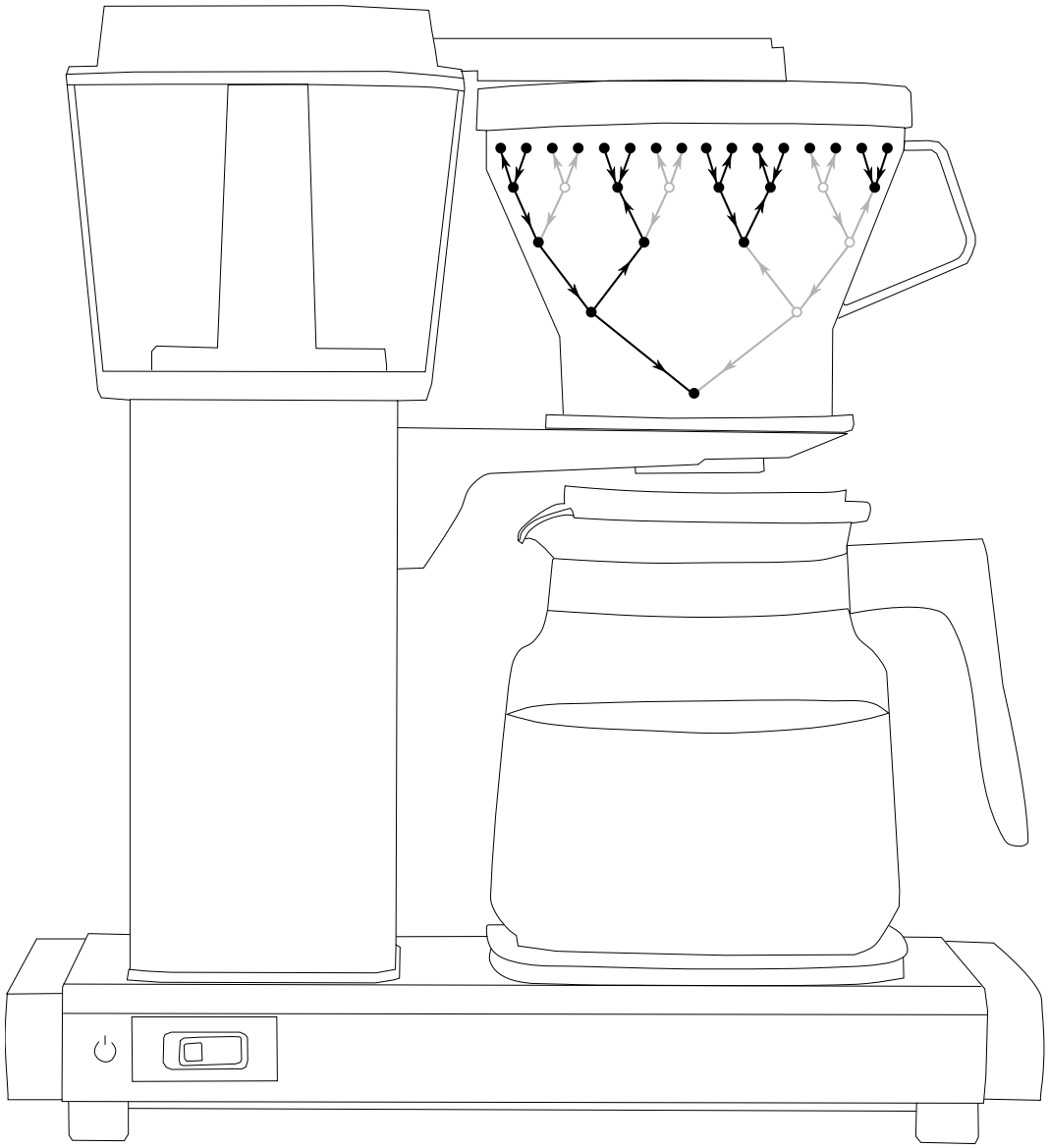
Using Lemma 4.B.1 and a symbolic computation software, we get an explicit expression for s_ℓ, and, from there,

$$\mathrm{Var}(\chi_\ell) = \frac{(59400\,\ell^9 - 1618650\,\ell^8 + O(\ell^7))\ell}{1576575\,(\ell-1)^2(\ell-2)^2(\ell-3)^2(\ell-4)(\ell-5)(\ell-6)}$$

from which Proposition 4.4.9 follows.

Chapter 5

Oriented percolation in randomly oriented graphs



This chapter was initially supposed to be about first passage oriented percolation on the Bernoulli site percolation cluster of the hypercube. This idea came from Amaury Lambert, who saw it as way to model the exploration by mutants of a "holey" genotypic landscape, where some genotypes are unviable. His idea was that we might be able to adapt and simplify classic results of Fill and Pemantle [5], and perhaps even a more recent results of Martinsson [8].

Unfortunately, I have so far been unable to get any significant results and this project is still in its early stages – which is why it has not been included in this thesis. Nevertheless, while working on these questions I started reading about oriented percolation. In particular, I found about an article recently published in *Probability, Combinatorics and Computing* by Narayanan [9], in which he proves that for any graph whoses edges have been randomly and independently oriented, for any set of vertices S, writing $\{S \rightsquigarrow i\}$ to indicate that there is an oriented path going from a vertex $s \in S$ to vertex i, the events $\{S \rightsquigarrow i\}$ and $\{S \rightsquigarrow j\}$ are positively correlated for any fixed pair of vertices i and j.

The surprising thing about Narayanan's paper, which he pointed out, was that his proof was unexpectedly complex. In particular, he had not been able to find a proof that did not use the Ahlswede-Daykin inequality (also known as the four functions theorem). I therefore tried to find a more elegant coupling proof, and even though I was unable to find one I eventually realized that it was possible to prove a much stronger result than the pairwise positive correlation of the events $\{S \rightsquigarrow i\}$, without resorting to the Ahlswede-Daykin inequality.

Having recently learned about the Stein-Chen method during my work on the split-and-drift random graph (see Chapter 2), I tried to find a simple example of randomly oriented graph in which I could use positive association to show that the number of vertices in the oriented percolation cluster was Poissonian. If anything, this made me experience first-hand the trade-off between dullness and intractability that was mentioned in Section 1.2: all the models that I could conceive proved either trivially simple, or too hard for me to study. As a testimony to my endeavours, I have included one of these models in Section 5.3 of this chapter. While this model clearly tends to fall on the "dull" side of the spectrum, it is not entirely trivial and has a few interesting properties (see in particular Proposition 5.3.9).

Publication: Sections 5.1 and 5.2 of this chapter have been accepted for publication in *Probability, Combinatorics and Computing* under the title "Positive association of the oriented percolation cluster in randomly oriented graphs".

Chapter contents

5.1 Introduction . . . 129
5.1.1 Positive association and related notions . . . 129
5.1.2 Notation . . . 130
5.2 Positive association of the percolation cluster . . . 130
5.2.1 Preliminary lemma . . . 130
5.2.2 Main result . . . 131
5.2.3 Corollaries . . . 132
5.3 Percolation from the leaves of a binary tree . . . 133
5.3.1 Setting and notation . . . 133
5.3.2 General results . . . 135
5.3.3 Badly-subcritical regime . . . 137
Chapter references . . . 144

5.1 Introduction

Oriented percolation is the study of connectivity in a random oriented graph. In most settings, one starts from a graph with a fixed orientation and then keeps each edge with a given probability. Classical such models include the north-east lattice [3] and the hypercube [5].

Another broad and natural class of random oriented graphs is obtained by starting from a fixed graph and then orienting each edge, independently of the orientations of other edges. Note that, in the general case, the orientations of the edges need not be unbiased: some edges can be allowed to have a higher probability to point towards one of their ends than towards the other. Percolation on such *randomly oriented graphs* has been studied, e.g. in [7], and more recently in [9], which motivated this chapter.

In [9], Narayanan showed that if the edges of any fixed graph are randomly and independently oriented, then writing $\{S \rightsquigarrow i\}$ to indicate that there is an oriented path going from a vertex $s \in S$ to vertex i, we have

$$\mathbb{P}(S \rightsquigarrow i, S \rightsquigarrow j) \;\geqslant\; \mathbb{P}(S \rightsquigarrow i)\,\mathbb{P}(S \rightsquigarrow j)\,.$$

The aim of this chapter is to strengthen and simplify the proof of this result. More specifically, let V be the vertex set of the graph. We prove that the events $\{S \rightsquigarrow i\}$, $i \in V$, are positively associated, without resorting to advanced results such as the Ahlswede–Daykin inequality [1].

5.1.1 Positive association and related notions

There are many ways to formalize the idea of a positive dependence between the random variables of a family $\mathbf{X} = (X_i)_{i\in I}$. A straightforward, weak one is to ask that these variables be pairwise positively correlated, i.e.

$$\forall i, j \in I, \quad \mathbb{E}(X_i X_j) \;\geqslant\; \mathbb{E}(X_i)\,\mathbb{E}(X_j)\,.$$

A much stronger condition, due to [4], is known as positive association. In the following definition and throughout the rest of this note, we use bold letters to denote vectors, as in $\mathbf{X} = (X_i)_{i\in I}$, and we write $\mathbf{X} \leqslant \mathbf{X}'$ to say that $X_i \leqslant X_i'$ for all i. Finally, a function $f\colon \mathbb{R}^I \to \mathbb{R}$ is said to be *increasing* when $\mathbf{X} \leqslant \mathbf{X}' \implies f(\mathbf{X}) \leqslant f(\mathbf{X}')$.

Definition 5.1.1. The random vector $\mathbf{X} = (X_i)_{i\in I}$ is said to be *positively associated* when, for all increasing functions f and g,

$$\mathbb{E}(f(\mathbf{X})g(\mathbf{X})) \;\geqslant\; \mathbb{E}(f(\mathbf{X}))\,\mathbb{E}(g(\mathbf{X}))$$

whenever these expectations exist. ◇

Without further mention, we only consider test functions f and g for which $\mathbb{E}(f(\mathbf{X}))$, $\mathbb{E}(g(\mathbf{X}))$ and $\mathbb{E}(f(\mathbf{X})g(\mathbf{X}))$ exist.

We say that the events A_i, $i \in I$, are positively associated when the corresponding vector of indicator variables $(\mathbb{1}_{A_i})_{i\in I}$ is positively associated. Similarly, a random subset R of the fixed set I can be seen as the vector

$$\mathbf{R} = \left(\mathbb{1}_{\{i\in R\}}\right)_{i\in I}\,,$$

so that R is said to be positively associated when the events $\{i \in R\}$, $i \in I$, are. This is equivalent to saying that for any increasing functions f and g from the power set of I to $\mathbb{R}$,

$$\mathbb{E}(f(R)g(R)) \geqslant \mathbb{E}(f(R))\,\mathbb{E}(g(R))\,,$$

where f being increasing is understood to mean that $r' \subset r \implies f(r') \leqslant f(r)$.

Positive association is famous for the FKG theorem, which states that it is implied by a lattice condition that can sometimes be very easy to check [6]. Another reason why it is so useful is that it implies weaker positive dependence notions that have to be checked in applications. One example of this is the existence of increasing couplings and the corresponding notion of *positive relation* used in the Stein–Chen method – see e.g, [2] and [10].

5.1.2 Notation

Let us fix some notation to be used throughout the rest of this chapter.

We study the simple graph $G = (V, E)$. Unless explicitly specified otherwise, V is assumed to be finite and we denote by $|V|$ its cardinality. The edges of G have a random orientation that is independent of the orientations of other edges and we write $\{i \to j\}$ to indicate that the edge $\{ij\}$ is oriented towards j. Formally, we are thus given a family of events $(\{i \to j\}, \{ij\} \in E)$ such that $\{i \to j\} = \{j \to i\}^{\mathsf{c}}$ and for all $\{ij\}$, $\{i \to j\} \perp\!\!\!\perp (\{k \to \ell\}, \{k\ell\} \neq \{ij\})$.

Finally, for every pair of vertices i and j, we write $\{i \rightsquigarrow j\}$ for the event that there exists an oriented path going from i to j. Similarly, for every source set S we let $\{S \rightsquigarrow i\} = \bigcup_{j \in S}\{j \rightsquigarrow i\}$ be the event that there is an oriented path from S to i, and for every target set T we let $\{i \rightsquigarrow T\} = \bigcup_{j \in T}\{i \rightsquigarrow j\}$ be the event that there is an oriented path from i to T. If there is an ambiguity regarding which graph is considered for these events, we will specify it with the notation $\{i \overset{G}{\rightsquigarrow} j\}$.

5.2 Positive association of the percolation cluster

5.2.1 Preliminary lemma

Lemma 5.2.1. *Let Γ be a finite set and let R be a positively associated random subset of Γ. Let X_i^r, $r \subset \Gamma$ and $i \in V$, be a family of events on the same probability space as R with the property that*

(i) $r' \subset r \implies X_i^{r'} \subset X_i^r$, $\forall i \in V$.

(ii) *For all $r \subset \Gamma$, $(X_i^r)_{i \in V}$ is positively associated and independent of R.*

For all $i \in V$, define X_i^R by

$$X_i^R := \bigcup_{r \subset \Gamma} \{R = r\} \cap X_i^r\,.$$

Then, the events X_i^R, $i \in V$, are positively associated.

Proof. Let f and g be two increasing functions. We have

$$\begin{aligned}\mathbb{E}\Big(f(\mathbf{X}^R)g(\mathbf{X}^R)\Big) &= \sum_{r\subset\Gamma}\mathbb{E}(f(\mathbf{X}^r)g(\mathbf{X}^r)\mathbb{1}_{\{R=r\}})\\ &= \sum_{r\subset\Gamma}\mathbb{E}(f(\mathbf{X}^r)g(\mathbf{X}^r))\,\mathbb{P}(R=r)\\ &\geqslant \sum_{r\subset\Gamma}\mathbb{E}(f(\mathbf{X}^r))\,\mathbb{E}(g(\mathbf{X}^r))\,\mathbb{P}(R=r)\,,\end{aligned}$$

because $\mathbf{X}^r \perp\!\!\!\perp R$ and $\mathbf{X}^r$ is positively associated. Now, let $u\colon r\mapsto \mathbb{E}(f(\mathbf{X}^r))$ and $v\colon r\mapsto\mathbb{E}(g(\mathbf{X}^r))$, so that the last sum is $\mathbb{E}(u(R)v(R))$. Note that u and v are increasing, since f and g are and, by hypothesis, $r'\subset r \implies \mathbf{X}^{r'}\leqslant\mathbf{X}^r$. Therefore, by the positive association of R,

$$\mathbb{E}(u(R)v(R)) \;\geqslant\; \mathbb{E}(u(R))\,\mathbb{E}(v(R))\,.$$

Finally, using again the independence of $\mathbf{X}^r$ and R, we have $\mathbb{E}(u(R))=\mathbb{E}(f(\mathbf{X}^R))$ and $\mathbb{E}(v(R))=\mathbb{E}(g(\mathbf{X}^R))$, which concludes the proof. □

5.2.2 Main result

Theorem 5.2.2. *Let G be a finite graph with vertex set V, whose edges have been randomly and independently oriented. Then, for any source set S, the events $\{S\rightsquigarrow i\}$, $i\in V$, are positively associated, i.e., for all increasing functions f and g and writing $\mathbf{X}=(\mathbb{1}_{\{S\rightsquigarrow i\}})_{i\in V}$,*

$$\mathbb{E}(f(\mathbf{X})g(\mathbf{X})) \;\geqslant\; \mathbb{E}(f(\mathbf{X}))\,\mathbb{E}(g(\mathbf{X}))\,.$$

Proof. Our proof uses the same induction on the number of vertices as Narayanan's. The difference is that we use Lemma 5.2.1 rather than the Ahlswede–Daykin inequality to propagate the positive dependence.

The theorem is trivial for the graph consisting of a single vertex (a family of a single variable being always positively associated) so let us assume that it holds for every graph with strictly less than $|V|$ vertices. Let Γ be the neighborhood of S, i.e.

$$\Gamma=\big\{v\in V\setminus S : \exists s\in S \text{ s.t. } \{vs\}\in E\big\}\,.$$

Then, let R be the random subset of Γ defined by

$$R=\big\{v\in\Gamma : \exists s\in S \text{ s.t. } s\to v\big\}\,.$$

Observe that the events $\{i\in R\}$, $i\in\Gamma$ are independent, so that the set R is positively associated.

Next, let H be the subgraph of G induced by $V\setminus S$. Note that, for all $i\in V\setminus S$,

$$\Big\{S\overset{G}{\rightsquigarrow} i\Big\}=\Big\{R\overset{H}{\rightsquigarrow} i\Big\}\,.$$

For every fixed $r\subset\Gamma$, the family $\{r\overset{H}{\rightsquigarrow} i\}$ for $i\in V\setminus S$ is independent of R because it depends only on the orientations of the edges of H, while R depends only on the orientations of the edges of G that go from S to Γ – and these two sets of edges are disjoint. Moreover, by the induction hypothesis, the events $\{r\overset{H}{\rightsquigarrow} i\}$, $i\in V\setminus S$, are positively associated. Since for fixed sets r and r' such that $r'\subset r$, $\{r'\rightsquigarrow i\}\implies\{r\rightsquigarrow i\}$ for all vertices, we can apply Lemma 5.2.1 to conclude that the events $\{R\rightsquigarrow i\}$, $i\in V\setminus S$, are positively associated.

To conclude the proof, note that the events $\{S\rightsquigarrow i\}$ are certain for $i\in S$ and that the union of a family of positively associated events and of a family of certain events is still positively related. □

5.2.3 Corollaries

Corollary 5.2.3. *Let G be a finite graph with independently oriented edges. For any target set T, the events $\{i \rightsquigarrow T\}$, $i \in V$, are positively associated.*

Proof. Consider the randomly oriented graph H obtained by reversing the orientation of the edges of G, i.e. such that $\{i \overset{H}{\to} j\} = \{j \overset{G}{\to} i\}$. Then for all $i \in V$,

$$\{i \overset{G}{\rightsquigarrow} T\} = \{T \overset{H}{\rightsquigarrow} i\},$$

and we already know from Theorem 5.2.2 that the events $\{T \overset{H}{\rightsquigarrow} i\}$, $i \in V$, are positively associated. □

Corollary 5.2.4. *Let G be an infinite graph with independently oriented edges. Let f and g be increasing, non-negative functions on $\mathbb{R}^V$ that depend only on a finite number of coordinates (i.e. such that there exists a finite set $U \subset V$ and $\tilde{f}\colon \mathbb{R}^U \to [0, +\infty[$ such that $f = \tilde{f} \circ \varphi$, where φ is the canonical surjection from $\mathbb{R}^V$ to $\mathbb{R}^U$). Then, for any source set S, letting $\mathbf{X} = (\mathbb{1}_{\{S \rightsquigarrow i\}})_{i \in V}$,*

$$\mathbb{E}(f(\mathbf{X})g(\mathbf{X})) \geqslant \mathbb{E}(f(\mathbf{X}))\,\mathbb{E}(g(\mathbf{X})).$$

Proof. Let G_n be an increasing sequence of finite graphs such that $G = \bigcup_n G_n$, and for all $i \in V$, let

$$X_i^{(n)} = \{S \overset{G_n}{\rightsquigarrow} i\},$$

so that $X_i^{(n)} \subset X_i^{(n+1)}$ and $X_i = \bigcup_n X_i^{(n)}$. Since the functions f and g are increasing, so are the sequences $f(\mathbf{X}^{(n)})$ and $g(\mathbf{X}^{(n)})$. Thus, using Theorem 5.2.2 and monotone convergence,

$$\mathbb{E}\Big(\lim_n f(\mathbf{X}^{(n)})g(\mathbf{X}^{(n)})\Big) \geqslant \mathbb{E}\Big(\lim_n f(\mathbf{X}^{(n)})\Big)\,\mathbb{E}\Big(\lim_n g(\mathbf{X}^{(n)})\Big).$$

Finally, if f and g depend on a finite number of events X_i, then for every realization of $\mathbf{X}$ we have $\lim_n f(\mathbf{X}^{(n)}) = f(\mathbf{X})$ and $\lim_n g(\mathbf{X}^{(n)}) = g(\mathbf{X})$. □

Corollary 5.2.5 (Narayanan, 2016)**.** *For any (possibly infinite) graph with independently oriented edges, for any source set S and for any two vertices i and j,*

$$\mathbb{P}(S \rightsquigarrow i, S \rightsquigarrow j) \geqslant \mathbb{P}(S \rightsquigarrow i)\,\mathbb{P}(S \rightsquigarrow j)$$

Proof. Take $f\colon (x_k)_{k \in V} \mapsto x_i$ and $g\colon (x_k)_{k \in V} \mapsto x_j$ in Corollary 5.2.4. □

Corollary 5.2.6. *Let G be a finite graph with independently oriented edges and vertex set V. For any source set S, let*

$$N = \sum_{i \in V \setminus S} \mathbb{1}_{\{S \rightsquigarrow i\}}$$

denote the size of the oriented percolation cluster of G, and set $\lambda = \mathbb{E}(N)$. Then,

$$d_{\mathrm{TV}}(N, \mathrm{Poisson}(\lambda)) \leqslant \min(1, \lambda^{-1}) \left(\mathrm{Var}(N) - \lambda + 2 \sum_{i \in V \setminus S} \mathbb{P}(S \rightsquigarrow i)^2 \right),$$

where d_{TV} denotes the total variation distance.

Proof. This is a direct application of the Stein–Chen method to the positively related variables $\mathbb{1}_{\{S \rightsquigarrow i\}}$, $i \in V \setminus S$ – see e.g. Theorem 4.20 in [10] (this theorem can be found as Theorem A in Chapter 2). □

The interest of Corollary 5.2.6 is that one only needs a suitable upper bound on $\mathrm{Cov}(\mathbb{1}_{\{S \rightsquigarrow i\}}, \mathbb{1}_{\{S \rightsquigarrow j\}})$ to show that the size of the oriented percolation cluster is Poissonian, as illustrated in the next section.

5.3 Percolation from the leaves of a binary tree

In this section, we study percolation on the randomly oriented complete binary tree of height n. We start by introducing this graph and some notation.

5.3.1 Setting and notation

The binary tree T_n

Let V_n be the set of words of length at most n on the alphabet $\{0,1\}$, i.e.

$$V_n = \bigcup_{k=0}^{n} \{0,1\}^k ,$$

where $\{0,1\}^0$ is understood to represent the empty word.

A word v is said to be a *successor* of u when $v = us$, with $s \in \{0,1\}$. Thus, every word of length less than n has two successors in V_n. Similarly, every non-empty word of V_n has exactly one predecessor. With this terminology, let

$$E_n = \big\{\{u,v\} : (u,v) \in V_n^2,\ v \text{ is a successor of } u\big\} .$$

What we call the complete binary tree of height n is the graph $T_n = (V_n, E_n)$. Let us fix some vocabulary and notation for working with T_n.

The *leaves* of T_n are the vertices of degree 1, and its *root* is the only vertex of degree 2. The root will always be denoted by r.

The *level* of a vertex is its distance from the leaf set. Thus, the leaves are the level-0 vertices, and the root is the only vertex of level n. We will write $\ell(v)$ for the level of vertex v.

The unique path between two vertices u and v will be denoted by $[u,v]$. Sometimes, we will need to remove one of its ends from $[u,v]$, in which case we will write $]u,v]$ for $[u,v] \setminus \{u\}$ and $[u,v[$ for $[u,v] \setminus \{v\}$.

Finally, there is a natural order $\preccurlyeq$ on the vertices of T_n, defined, e.g, by

$$u \preccurlyeq v \iff v \in [u,r]$$

Thus ordered, $(V_n, \preccurlyeq)$ is a join-semilattice, i.e. we can define the *join* of any u and v, denoted by $u \vee v$, as

$$u \vee v = \inf\big([u,r] \cap [v,r]\big) = \sup [u,v]$$

These definitions are illustrated in Figure 1A.

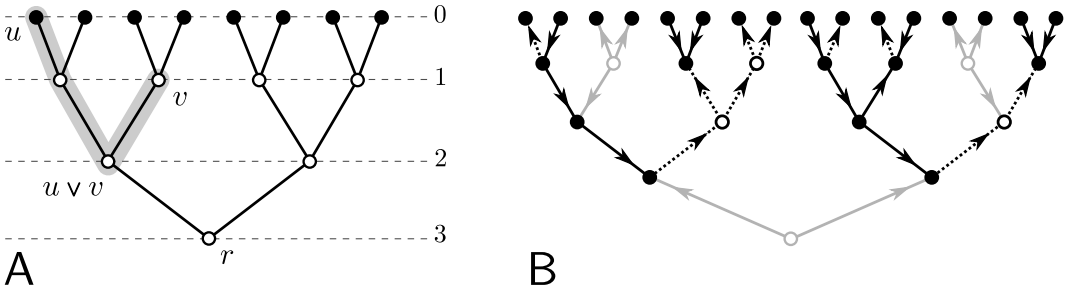


Figure 5.1: A, the complete binary tree T_3. The black vertices are the leaves of the tree, and r is the root. The numbers on the right indicate the levels of the vertices. The path $[u, v]$ between u and v has been highlighted and $u \vee v$, the join of u and v, can be seen to be the unique vertex of maximum level in $[u, v]$. B, percolation and downwards percolation on T_4. Water starts from the leaves and then flows downwards through black edges and upwards through dotted edges. It does not reach the grayed-out portions of the tree. The percolation cluster C_n consists of both black vertices and white vertices, while the downwards percolation cluster $C_n^{\downarrow}$ consists of black vertices only. Note that the leaves are excluded from both percolation clusters.

Percolation and downwards percolation on T_n

Let every edge of T_n be oriented towards the root with probability p and towards the leaf set with probability $1 - p$, independently of the other edges.

In this application, the source set L will be the leaf set of T_n. In other words, we pump water into the leaves of T_n and let it flow through those edges whose orientation matches that of the flow, as depicted in Figure 1B. For any vertex v, write

$$X_v = \{L \rightsquigarrow v\},$$

for the event that the water reaches v, and

$$\pi_k^{(n)} = \mathbb{P}(X_v), \quad \text{where } k = \ell(v)$$

for the probability of this event. In the special case where $v = r$ is the root, we use the notation

$$\rho_n = \pi_n^{(n)} = \mathbb{P}(X_r).$$

Finally, let

$$C_n = \{v \in V_n \setminus L : X_v\}$$

denote the percolation cluster.

As will become clear, this percolation model is closely related to a simpler one where, in addition to respecting the orientation of edges, water is constrained to flow towards increasing levels of the tree. If we think of the root of T_n as its bottom and of the leaves as its top, then water runs down from the leaves, traveling through downwards-oriented edges; hence we refer to this second model as *downwards percolation*. Again, this is represented in Figure 1B.

Let us write Y_v for the event that that vertex v gets wet in downwards percolation, and let

$$C_n^{\downarrow} = \{v \in V_n \setminus L : Y_v\}$$

be the downwards-percolation cluster.

How are percolation and downwards percolation related? First, it follows directly from the definition that $Y_v \subset X_v$. Second, note that

$$Y_r = X_r$$

because every path from the leaf set to r is downwards-oriented. Furthermore, letting $T_n^{(\preccurlyeq v)}$ denote the subtree of T_n induced by v and the vertices that are above it, then the randomly oriented trees $T_n^{(\preccurlyeq v)}$ and $T_{\ell(v)}$ have the same law. As a result, for all $v \in V_n$,

$$\mathbb{P}(Y_v) = \rho_{\ell(v)}\,,$$

from which the next proposition follows.

Proposition 5.3.1. *Let $|C_n^{\downarrow}|$ be the number of wet vertices (not counting the leaves) in the downwards-percolation model on T_n. We have*

$$\mathbb{E}\Big(|C_n^{\downarrow}|\Big) = \sum_{k=1}^{n} 2^{n-k}\rho_k\,.$$

5.3.2 General results

Percolation threshold

If the probability p that an edge is oriented towards the root is sufficiently small, the probability ρ_n of the root getting wet will go to zero as n goes to infinity. Define the percolation threshold as

$$\theta_c = \sup\{p : \rho_n \to 0\}\,.$$

Proposition 5.3.2. *The probability ρ_n of the root of T_n getting wet in either percolation model satisfies the following recurrence:*

$$\rho_{n+1} = 2p\rho_n - (p\rho_n)^2\,, \quad \text{with } \rho_0 = 1.$$

The percolation threshold is therefore $\theta_c = 1/2$ and

(i) $p \leqslant \theta_c \implies \rho_n \to 0.$

(ii) $p > \theta_c \implies \rho_n \to (2p-1)/p^2 > 0.$

Proof. First, note that

$$Y_r = (Y_0 \cap \{0 \to r\}) \cup (Y_1 \cap \{1 \to r\})\,,$$

where 1 and 0 are the two successors of the root r. These four events are independent and we have $\mathbb{P}(0 \to r) = \mathbb{P}(1 \to r) = p$ and $\mathbb{P}(Y_0) = \mathbb{P}(Y_1) = \rho_{n-1}$, whence the recurrence relation.

Now, let $f_p\colon x \mapsto 2px - (px)^2$, so that $\rho_{n+1} = f_p(\rho_n)$. For $p \leqslant 1/2$, the only solution to the equation $f_p(x) = x$ in $[0,1]$ is $x = 0$, and $f_p(x) < x$ for all $0 < x \leqslant 1$. This proves (i). For $p > 1/2$, the equation $f_p(x) = x$ has a non-zero solution $\alpha = (2p-1)/p^2$ in $[0,1]$. Finally, $f_p(x) > x$ for $0 < x < \alpha$ and $f_p(x) < x$ for $\alpha < x < 1$, proving (ii). □

Remark 5.3.3. Another way to obtain Proposition 5.3.2 is to note that the existence of an open path from the leaf set of T_n to its root is equivalent to the existence of a path of length n starting from the root of a Galton–Watson tree with Binomial$(2,p)$ offspring distribution, i.e. to its non-extinction after n generations. In the limit as $n \to \infty$, the probability of non-extinction is strictly positive if and only if the expected number of offspring is greater than 1 – i.e., in our case, $2p > 1$. ⋄

Expected size of the percolation cluster

Let us clarify the relation between percolation and downwards percolation by expressing the probability $\pi_k^{(n)}$ that a vertex gets wet in (bidirectional) percolation as a function of $\rho_k, \dots, \rho_n$.

Proposition 5.3.4. *Let $\pi_k^{(n)} = \mathbb{P}(X_v)$, where $\ell(v) = k$. We have*

$$\pi_k^{(n)} = \rho_k + (1-\rho_k)\,\alpha_k^{(n)}\,,$$

where

$$\alpha_k^{(n)} = (1-p)\,p\sum_{i=0}^{n-1-k}(1-p)^i\rho_{k+i}\prod_{j=0}^{i-1}(1-p\rho_{k+j})$$

is the probability that water reaches v "from below" and ρ_k is the probability that it reaches it "from above".

Remark 5.3.5. To make sense of the expression of $\alpha_k^{(n)}$, it can also be written as

$$\alpha_k^{(n)} = \sum_{i=1}^{n-k}\mathbb{P}(M=k+i)\,,\quad \text{with } \mathbb{P}(M=k+i) = (1-p)^i p\rho_{k+i-1}\prod_{j=0}^{i-2}(1-p\rho_{k+j}).$$

In this expression, M is the level of the highest (that is, minimal with respect to $\preccurlyeq$) vertex $u \in]v,r]$ such that $Y_u \cap \{u \rightsquigarrow v\}$ (with $M = +\infty$ if there is no such vertex). ⋄

Proof. Water can reach v from above (i.e. coming from one of its successors) or from below (coming from its predecessor). These two events are independent, because they depend on what happens in disjoint regions of T_n.

Water reaches v from above if and only if v gets wet in downwards percolation. To reach v from below, water had to travel through a portion of the path $[v,r]$ from v to the root. To enter this portion of the path, it had to reach at least one vertex, say u, from above. Let $\varphi(u)$ be the successor of u that does not belong to $[v,r]$. The water had to get to $\varphi(u)$ from above, flow to u, and from here to v.

This reasoning, which is illustrated in Figure 2A, leads us to rewrite X_v as

$$X_v \;=\; Y_v \;\cup \bigcup_{u\in]v,r]}\Big(Y_{\varphi(u)}\cap\{\varphi(u)\to u\}\cap\{u\rightsquigarrow v\}\Big)$$

In order to compute the probability of this event, we rewrite it as the disjoint union

$$X_v = Y_v \;\cup \bigcup_{u\in]v,r]}\left(Y_v^{\mathsf{c}}\cap\Big(\bigcap_{w\in]v,u[}\widetilde{Y}_w^{\mathsf{c}}\Big)\cap\widetilde{Y}_u\cap\{u\rightsquigarrow v\}\right),$$

where

$$\widetilde{Y}_x = Y_{\varphi(x)}\cap\{\varphi(x)\to x\}\,.$$

Next, we note that the factors of each term of the union over $u \in]v,r]$ are independent, because they are determined by the orientations of disjoint sets of edges: Y_v depends only on the orientations of the edges of $T_n^{(\preccurlyeq v)}$; each $\widetilde{Y}_x$ of those of the edges of $T_n^{(\preccurlyeq\varphi(x))}$ and of $\{x,\varphi(x)\}$; and $\{u\rightsquigarrow v\}$ of the edges of $[u,v]$. Using that $\mathbb{P}(Y_v) = \rho_{\ell(v)}$, $\mathbb{P}\left(\widetilde{Y}_x\right) = p\rho_{\ell(x)-1}$ and $\mathbb{P}(u\rightsquigarrow v) = (1-p)^{d(u,v)}$ and replacing the sum on the vertices of $]v,r]$ by a sum on their levels, we get the desired expression. □

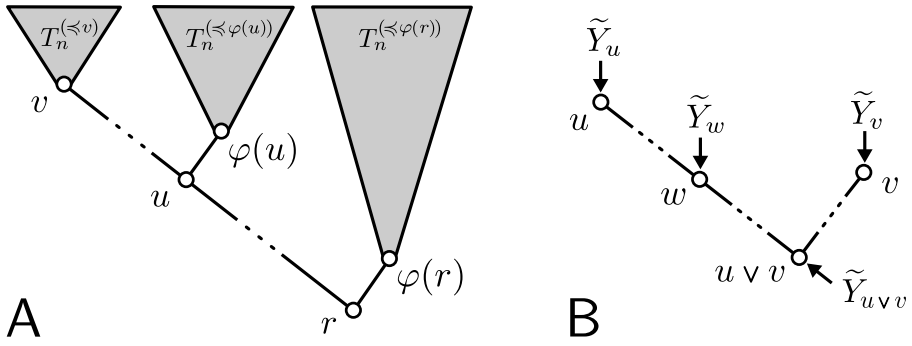


Figure 5.2: A, the notations used in the proof of Proposition 5.3.4. Water can reach v from above, i.e. traveling through $T_n^{(\preccurlyeq v)}$, or from below, coming from some vertex $u \in]v, r]$. B, the notations used in the proof of Proposition 5.3.8. The arrows represent possible entry points for the water, and the $\widetilde{Y}_x$ the associated events, i.e., $\widetilde{Y}_x$ is the event that x receives water from the corresponding arrow.

From Proposition 5.3.4, we get the following expression for the expected size of the percolation cluster:

Proposition 5.3.6. *Let $|C_n|$ be the number of wet vertices, not counting the leaves, in the (bidirectional) percolation model on T_n. Then,*

$$\mathbb{E}\big(|C_n|\big) = \sum_{k=1}^{n} 2^{n-k}\Big(\rho_k + (1-\rho_k)\,\alpha_k^{(n)}\Big)\,,$$

where $\alpha_k^{(n)}$ is defined in Proposition 5.3.4.

Using a similar reasoning, it is also possible to express $\mathbb{P}(X_u X_v)$ – and from there $\mathrm{Var}(|C_n|)$ – as a function of p and $\rho_1, \dots, \rho_n$ only. However, the resulting expression is rather complicated, and thus of little interest. We will therefore only give the asymptotic estimates that are needed to apply Theorem 5.2.2.

5.3.3 Badly-subcritical regime

In this section, we focus on what happens when $p = p_n$ is allowed to depend on n and made to go to zero as n goes to infinity. We are therefore in a "badly-subcritical" regime, where only a negligible fraction of the vertices are going to get wet.

Note that the results of the previous sections still hold, provided that ρ_k is understood to depend on n as the solution of

$$\rho_{k+1} = 2p_n\rho_k - (p_n\rho_k)^2\,, \quad \rho_0 = 1.$$

To avoid clutter, the dependence in n will remain implicit and we will keep the notation ρ_k.

Asymptotic cluster size and maximum depth

Proposition 5.3.7. *When $p_n \to 0$, then as $n \to \infty$,*

$$\rho_k \sim (2p_n)^k\,,$$

where the convergence is uniform in k.

Proof. Clearly,

$$\rho_k \leqslant (2p_n)^k .$$

Plugging this first inequality into the recurrence relation for ρ_k, we get

$$\rho_{k+1} \geqslant (2p_n - p_n^2(2p_n)^k)\rho_k ,$$

from which it follows that

$$\begin{aligned}\rho_k &\geqslant (2p_n)^k \prod_{i=0}^{k-1}\Big(1 - \frac{p_n}{2}(2p_n)^i\Big) \\ &\geqslant (2p_n)^k \prod_{i=1}^{k}\Big(1 - (2p_n)^i\Big) .\end{aligned}$$

Let us show that

$$P_n^{(k)} = \prod_{i=1}^{k}\Big(1 - (2p_n)^i\Big)$$

has a lower bound that goes to 1 uniformly in k as $n \to \infty$. For all $k \geqslant 1$,

$$\log\Big(P_n^{(k)}\Big) \geqslant \sum_{i=1}^{\infty}\log\Big(1 - (2p_n)^i\Big) .$$

Now,

$$\sum_{i=1}^{\infty}\log\Big(1 - (2p_n)^i\Big) = -\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\tfrac{1}{j}(2p_n)^{ij} \geqslant -\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(2p_n)^{ij}$$

and

$$-\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(2p_n)^{ij} = -\sum_{i=1}^{\infty}\frac{(2p_n)^i}{1-(2p_n)^i} \geqslant -\sum_{i=1}^{\infty}\frac{(2p_n)^i}{1-2p_n} = -\frac{2p_n}{(1-2p_n)^2} ,$$

so that $P_n^{(k)} \geqslant \exp(-2p_n/(1-2p_n)^2)$. Putting the pieces together,

$$e^{-\frac{2p_n}{(1-2p_n)^2}}(2p_n)^k \leqslant \rho_k \leqslant (2p_n)^k ,$$

which terminates the proof. □

Proposition 5.3.8. *When $p_n \to 0$, then as $n \to \infty$,*

$$\mathbb{E}(|C_n|) \sim \mathbb{E}\Big(|C_n^{\downarrow}|\Big) \sim 2^n p_n .$$

Proof. From the expression of $\alpha_k^{(n)}$ given in Proposition 5.3.4,

$$\alpha_k^{(n)} \leqslant p_n \sum_{i=0}^{n-1-k}\rho_{k+i} .$$

But since $\rho_{k+i} \leqslant (2p_n)^i\rho_k$,

$$\alpha_k^{(n)} \leqslant \frac{p_n\,\rho_k}{1-2p_n} .$$

Using this in Propositions 5.3.1 and 5.3.6, we see that for n large enough,

$$\mathbb{E}(|C_n|) \leqslant \Big(1 + \frac{p_n}{1-2p_n}\Big)\,\mathbb{E}\Big(|C_n^{\downarrow}|\Big) ,$$

Next, using again that $\rho_k \leqslant (2p_n)^{k-1}\rho_1$,

$$2^{n-1}\rho_1 \;\leqslant\; \sum_{k=1}^{n} 2^{n-k}\rho_k \;\leqslant\; 2^{n-1}\rho_1 \sum_{k=1}^{n} p_n^{k-1}\,.$$

Since the sum in right-hand side is bounded above by $1/(1-p_n)$ and since $\rho_1 \sim 2p_n$, this finishes the proof. □

Proposition 5.3.8 shows that, in the badly-subcritical regime, the overwhelming majority of wet vertices are level-1 vertices. It is therefore natural to wonder: how deep does water go?

Proposition 5.3.9. *Let $\ell_{\max}^{(n)}$ be the maximum level reached by water, and let*

$$\kappa_n = \frac{\log(2)n}{\log(1/p_n)}\,.$$

If $p_n \to 0$, then letting $\lfloor x \rceil = \lfloor x + 1/2 \rfloor$ denote the nearest integer to x,

$$\mathbb{P}\Big(\ell_{\max}^{(n)} = \lfloor \kappa_n \rceil - 1 \text{ or } \lfloor \kappa_n \rceil\Big) \to 1$$

as $n \to \infty$. In particular,

- *If $p_n = n^{-\alpha}$, then $\kappa_n = cn/\log(n)$, with $c = \log(2)/\alpha$.*
- *If $p_n = \gamma^{-n}$, $1 < \gamma \leqslant 2$, then $\kappa_n = \frac{\log(2)}{\log(\gamma)}$.*

Proposition 5.3.9 shows that the maximum level reached by water is remarkably deterministic in the limit as n goes to infinity, independently of the speed of convergence of p_n to zero. It also shows that, even in the badly-subcritical regime, water can go infinitely deep – even though these depths will always represent a negligible fraction of the total height of T_n.

Before jumping to the proof, let us give a simple heuristic. Let

$$B_k^{(n)} = \mathrm{Card}\Big\{v \in C_n^{\downarrow} : \ell(v) = k\Big\}$$

be the number of level-k vertices that get wet in downwards-percolation. Using Proposition 5.3.7, we see that

$$\mathbb{E}\Big(B_k^{(n)}\Big) = \rho_k\, 2^{n-k} \;\sim\; (p_n)^k\, 2^n\,.$$

If k is such that this expectation goes to zero, then the probability that this level will be reached by water will go to zero and k will be a lower bound on $\ell_{\max}^{(n)}$. Conversely, if this expectation goes to infinity then it seems reasonable to expect that $B_k^{(n)} \geqslant 1$ with high probability, in which case we would have $\ell_{\max}^{(n)} \geqslant k$.

Proof. Let L_k be the set of level-k vertices. The event that water does not reach level k is

$$\Big\{\ell_{\max}^{(n)} < k\Big\} \;=\; \bigcap_{v \in L_k} Y_v{}^{\mathsf{c}}\,.$$

Since each Y_v depends only on $T_n^{(\preccurlyeq v)}$, these events are independent and we have

$$\mathbb{P}\Big(\ell_{\max}^{(n)} < k\Big) \;= (1-\rho_k)^{2^{n-k}}\,.$$

Whether this expression goes to 0 or to 1 is determined by whether $\rho_k 2^{n-k}$ goes to $+\infty$ or to 0, respectively. Now let $k = k_n$ depend on n. By Proposition 5.3.7, we have

$$\rho_{k_n} 2^{n-k_n} \sim (p_n)^{k_n} 2^n .$$

Again, whether this quantity goes to $+\infty$ or to 0 depends on whether

$$w_n = \log(p_n)\, k_n + \log(2)\, n$$

goes to $+\infty$ or to $-\infty$, respectively. Setting $\kappa_n = \frac{\log(2)n}{\log(1/p_n)}$, we see that:

(i) If there exists $\eta > 0$ such that $k_n < \kappa_n - \eta$ for all n, then $w_n \to +\infty$ and as a result $\mathbb{P}\left(\ell_{\max}^{(n)} \geqslant k\right) \to 1$.

(ii) If there exists $\eta > 0$ such that $k_n > \kappa_n + \eta$ for all n, then $w_n \to -\infty$ and as a result $\mathbb{P}\left(\ell_{\max}^{(n)} < k\right) \to 1$.

Finally, we note that

- $\lfloor \kappa_n \rceil - 1 \leqslant \kappa_n - 1/2$. By (i), this shows that $\mathbb{P}\left(\ell_{\max}^{(n)} \geqslant \lfloor \kappa_n \rceil - 1\right) \to 1$.
- $\lfloor \kappa_n \rceil + 1 > \kappa_n + 1/2$. By (ii), this shows that $\mathbb{P}\left(\ell_{\max}^{(n)} < \lfloor \kappa_n \rceil + 1\right) \to 1$.

As a result,

$$\mathbb{P}\left(\lfloor \kappa_n \rceil - 1 \leqslant \ell_{\max}^{(n)} \leqslant \lfloor \kappa_n \rceil\right) \to 1 ,$$

and the proof is complete. □

Second moments and main result

We have seen in the previous section that level-1 vertices account for a fraction 1 of the expected size of both percolation clusters. But do they also account for a fraction 1 of the variances?

For downwards percolation, it is not hard to convince oneself that it is. Indeed, the number $B_1^{(n)}$ of wet vertices of level 1 is a binomial variable with parameters ρ_1 and 2^{n-1}. From here, if we neglect "collision" events, where a vertex receives water from both vertices immediately above it, then the downwards percolation cluster resembles $B_1^{(n)}$ independent paths with geometric lengths, that is,

$$|C_n^{\downarrow}| \;\approx\; \sum_{i=1}^{B_1^{(n)}} \tau_i, \quad \text{where } \tau_i \sim \text{Geometric}(1 - p_n) .$$

Since $\text{Var}(\tau_i) \sim p_n$, by a simple application of the law of total variance we find that

$$\text{Var}\left(|C_n^{\downarrow}|\right) \approx \text{Var}\left(B_1^{(n)}\right) \sim p_n\, 2^n .$$

For bidirectional percolation however, things are not so obvious because there is a very strong feedback from higher-level vertices to lower-level ones: if vertex v gets wet, water will flow up from it to most vertices of $T_n^{(\preccurlyeq v)}$ that are not already wet. Thus, every rare event where water reaches a vertex of level k will results in approximately 2^k additional vertices getting wet – which it seems could increase the variance of $|C_n|$. However we will see that this is not the case.

Proposition 5.3.10. *In the regime $p_n \to 0$,*

(i) *If $u \preccurlyeq v$ then* $\mathrm{Cov}(X_u, X_v) \leqslant \pi^{(n)}_{\ell(v)}$.

(ii) *Otherwise,* $\mathrm{Cov}(X_u, X_v) \leqslant d(u,v)\rho_{\ell(u\vee v)}$.

As a result, $\mathrm{Var}(|C_n|) \sim 2^n p_n$ *and* $\mathrm{Var}(|C_n|) - \mathbb{E}(|C_n|) = O(2^n p_n^2)$.

Proof. Point (i) is clear, since

$$\mathrm{Cov}(X_u, X_v) \leqslant \mathbb{P}(X_u \cap X_v) \leqslant \min\{\mathbb{P}(X_u), \mathbb{P}(X_v)\} = \pi^{(n)}_{\ell(v)}\,.$$

As a side note, this upper bound on $\mathbb{P}(X_u \cap X_v)$ is not as crude as it may seem. Indeed, $X_v \cap \{v \rightsquigarrow u\} \subset X_u \cap X_v$ and is it not hard to check that $\mathbb{P}(X_v \cap \{v \rightsquigarrow u\})$ is greater than $(1-p_n)^{d(u,v)}\rho_{\ell(v)}/2 \;\sim\; \pi^{(n)}_{\ell(v)}/2$.

To prove (ii), let us show that

$$\mathbb{P}(X_u \cap X_v) \;\leqslant\; \pi^{(n)}_u \pi^{(n)}_v + d(u,v)\rho_{\ell(u\vee v)}\,.$$

As in the proof of Proposition 5.3.4, we start by re-expressing X_u. For every $w \in]u,v[$, let $\varphi(w)$ be the successor of w that does not belong to $[u,v]$, and for every $z \in]u\vee v, r]$, let $\psi(z)$ be the successor of z that does not belong to $[u\vee v, r]$. Then, for every $w \in [u,v]$, define $\widetilde{Y}_w$ by

$$\widetilde{Y}_w = \begin{cases} Y_w & \text{if } w = u \text{ or } w = v. \\ \displaystyle\bigcup_{z\in]u\vee v,\, r]} Y_{\psi(z)} \cap \{\psi(z) \to z\} \cap \{z \rightsquigarrow u\vee v\} & \text{if } w = u\vee v. \\ Y_{\varphi(w)} \cap \{\varphi(w) \to w\} & \text{otherwise.} \end{cases}$$

These definitions are illustrated in Figure 2B. Note that $\widetilde{Y}_{u\vee v}$ is simply the event that $u\vee v$ receives water "from below". Thus, using the notation of Proposition 5.3.4, we have

$$\mathbb{P}\left(\widetilde{Y}_{u\vee v}\right) \;=\; \alpha^{(n)}_{\ell(u\vee v)}.$$

For both $s = u$ and $s = v$ we have

$$X_s \;=\; \bigcup_{w\in[u,v]} \widetilde{Y}_w \cap \{w \rightsquigarrow s\}\,.$$

Again, we rewrite this as the disjoint union

$$X_s = \bigcup_{w\in[u,v]} Z^s_w\,,$$

where

$$Z^s_w = \Big(\bigcap_{z\in[s,w[} \widetilde{Y}^c_z\Big) \cap \widetilde{Y}_w \cap \{w \rightsquigarrow s\}\,.$$

Next, we note that for any vertices x and y in $[u,v]$,

- If $[u,x] \cap [y,v] = \emptyset$, then $Z^u_x \perp\!\!\!\perp Z^v_y$.
- If $[u,x] \cap [y,v] = \{w\}$, then $Z^u_x \cap Z^v_y = \widetilde{Y}_w \cap \{w \rightsquigarrow u\} \cap \{w \rightsquigarrow v\}$.
- Otherwise, $Z^u_x \cap Z^v_y = \emptyset$.

As a result,

$$X_u \cap X_v = \bigcup_{x\in[u,v]} \bigcup_{y\in[u,v]} Z_x^u \cap Z_y^v$$
$$= \bigcup_{x\in[u,v]} \left((\bigcup_{y\in]x,v]} Z_x^u \cap Z_y^v) \ \cup \ (Z_x^u \cap Z_x^v) \right)$$

It follows that

$$\mathbb{P}(X_u, X_v) \ = \ \sum_{x\in[u,v]} \sum_{y\in]x,v]} \mathbb{P}(Z_x^u)\mathbb{P}(Z_y^v) + \mathbb{P}\left(\widetilde{Y}_x, x \rightsquigarrow u, x \rightsquigarrow v\right).$$

To bound this sum, first note that

$$\sum_{x\in[u,v]} \sum_{y\in]x,v]} \mathbb{P}(Z_x^u)\mathbb{P}(Z_y^v) \ \leqslant \ \sum_{x\in[u,v]} \sum_{y\in[u,v]} \mathbb{P}(Z_x^u)\mathbb{P}(Z_y^v) \ = \ \pi_u^{(n)}\pi_v^{(n)}.$$

Next, $\widetilde{Y}_x$ and $\{x \rightsquigarrow u, x \rightsquigarrow v\}$ are independent and, writing $m(x) = d(x, u\vee v)$ for the number of downwards-oriented edges in the unique configuration of the edges of $[u, v]$ such that $\{x \rightsquigarrow u, x \rightsquigarrow v\}$,

$$\mathbb{P}(x \rightsquigarrow u, x \rightsquigarrow v) = p_n^{m(x)}(1-p_n)^{d(u,v)-m(x)}$$

while

$$\mathbb{P}\left(\widetilde{Y}_x\right) = \begin{cases} \alpha^{(n)}_{\ell(u\vee v)} & \text{if } x = u\vee v \\ \rho_{\ell(x)} & \text{if } x = u \text{ or } x = v \\ p_n\, \rho_{\ell(x)-1} & \text{otherwise.} \end{cases}$$

Since for $x \in [u, v]$, $\ell(x) = \ell(u\vee v) - m(x)$, and that

$$\rho_{\ell(u\vee v)-k} \ \sim \ (2p_n)^{-k}\rho_{\ell(u\vee v)} \ \leqslant \ p_n^{-k}\rho_{\ell(u\vee v)},$$

we see that for every $x \in [u, v]$, $x \neq u\vee v$,

$$\mathbb{P}\left(\widetilde{Y}_x, x \rightsquigarrow u, x \rightsquigarrow v\right) \ \leqslant \ \rho_{\ell(u\vee v)},$$

while for $x = u\vee v$ we already know from Proposition 5.3.8 and its proof that

$$\alpha^{(n)}_{\ell(u\vee v)} \leqslant p_n\, \rho_{\ell(u\vee v)}.$$

Discarding this negligible last contribution and summing these inequalities over the $d(u, v)$ vertices of $[u, v] \setminus \{u\vee v\}$, we find that

$$\sum_{x\in[u,v]} \mathbb{P}\left(\widetilde{Y}_x, x \rightsquigarrow u, x \rightsquigarrow v\right) \ \leqslant \ d(u,v)\, \rho_{\ell(u\vee v)},$$

which complete the proof of (ii).

Now let us show that $\mathrm{Var}(|C_n|) \leqslant \mathbb{E}(|C_n|) + O(2^n p_n^2)$. For $w \in]v, r]$, let $\varphi(w)$ denote the successor of w that does not belong to $[v, r]$. We decompose $\mathrm{Var}(|C_n|)$ into

$$\mathrm{Var}(|C_n|) \ = \ \sum_{v\in T_n} \left(\sum_{u\in T_n^{(\preccurlyeq v)}} \mathrm{Cov}(X_v, X_u) \ + \sum_{w\in]v,r]} \left(\mathrm{Cov}(X_v, X_w) + \sum_{x\in T_n^{(\preccurlyeq \varphi(w))}} \mathrm{Cov}(X_v, X_x) \right) \right)$$

where it is understood that the sums exclude leaves. Using (i), we see that

$$\sum_{u \in T_n^{(\preccurlyeq v)}} \mathrm{Cov}(X_v, X_u) \;\leqslant\; (2^{\ell(v)} - 1)\, \pi^{(n)}_{\ell(v)}\,.$$

Similarly, using (ii) we have

$$\begin{aligned}\sum_{x \in T_n^{(\preccurlyeq \varphi(w))}} \mathrm{Cov}(X_v, X_x) \;&\leqslant\; \sum_{x \in T_n^{(\preccurlyeq \varphi(w))}} \big(d(v,w) + d(w,x)\big)\rho_{\ell(w)} \\ &\leqslant\; (2^{\ell(w)-1} - 1)\,\big(d(v,w) + \ell(w) - 1\big)\,\rho_{\ell(w)}\,.\end{aligned}$$

Since $\mathrm{Cov}(X_v, X_w) \leqslant \pi^{(n)}_{\ell(w)}$, which is asymptotically equivalent to $\rho_{\ell(w)}$, we have

$$\mathrm{Cov}(X_v, X_w) + \sum_{x \in T_n^{(\preccurlyeq \varphi(w))}} \mathrm{Cov}(X_v, X_x) \;\leqslant\; \big(\ell(w) + d(v,w)\big)\, 2^{\ell(w)} \rho_{\ell(w)}$$

Replacing the sum on w by a sum on its level and letting k denote the level of v, we get, for every $\varepsilon > 0$,

$$\begin{aligned}\sum_{w \in]v,r]} \Big(\mathrm{Cov}(X_v, X_w) + \sum_{x \in T_n^{(\preccurlyeq \varphi(w))}} \mathrm{Cov}(X_v, X_x)\Big) \;&\leqslant\; \sum_{i=1}^{n-k} (k+2i)\, 2^{k+i} \rho_{k+i} \\ &\leqslant\; 2^k \rho_k \left(k \sum_{i=1}^{n-k} (4p_n)^i + 2 \sum_{i=1}^{n-k} i (4p_n)^i \right) \\ &\leqslant\; (1+\varepsilon)\, p_n\, (k+2)\, 2^{k+2} \rho_k\end{aligned}$$

Putting the pieces together, we find that

$$\mathrm{Var}(|C_n|) \;\leqslant\; \sum_{k=1}^{n} 2^{n-k}(2^k - 1)\, \pi_k^{(n)} \;+\; (1+\varepsilon)\, p_n \sum_{k=1}^{n} 2^{n+2}(k+2)\, \rho_k\,.$$

The first sum is

$$\sum_{k=1}^{n} 2^{n-k}(2^k - 1)\, \pi_k^{(n)} \;=\; \mathbb{E}(|C_n|) \;+\; \sum_{k=1}^{n} 2^{n-(k-1)}(2^{k-1} - 1)\, \pi_k^{(n)}$$

where

$$\sum_{k=1}^{n} 2^{n-(k-1)}(2^{k-1} - 1)\, \pi_k^{(n)} \;\leqslant\; 2^n \sum_{k=2}^{n} \pi_k^{(n)} \;=\; O(2^n p_n^2)\,,$$

since $\pi_k^{(n)} \leqslant (1+\varepsilon)\rho_k$ and $\rho_k \leqslant (2p_n)^k$. Finally, the second sum is also clearly $O(2^n\, p_n^2)$, and the proof is complete. □

With Proposition 5.3.10, Theorem 5.2.2 makes the following result immediate.

Proposition 5.3.11. *In the regime $p_n \to 0$, we have*

$$d_{\mathrm{TV}}\big(|C_n|, \mathrm{Poisson}(2^n p_n)\big) \;=\; O(p_n)$$

where d_{TV} denotes the total variation distance.

Proof. The proposition is a direct application of the Stein–Chen method to the positively related variables X_v, $v \in T_n$. □

Literature cited in this chapter

[1] R. Ahlswede and D. E. Daykin. An inequality for the weights of two families of sets, their unions and intersections. *Probability Theory and Related Fields*, 43(3):183–185, 1978.

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

[3] R. Durrett. Oriented percolation in two dimensions. *The Annals of Probability*, 12(4):999–1040, 1984.

[4] J. D. Esary, F. Proschan, and D. W. Walkup. Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5):1466–1474, 1967.

[5] J. A. Fill and R. Pemantle. Percolation, first-passage percolation and covering times for Richardson's model on the n-cube. *The Annals of Applied Probability*, 3(2):593–629, 1993.

[6] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[7] S. Linusson. A note on correlations in randomly oriented graphs. *arXiv preprint arXiv:0905.2881*, 2009.

[8] A. Martinsson. Unoriented first-passage percolation on the n-cube. *The Annals of Applied Probability*, 26(5):2597–2625, 2016.

[9] B. Narayanan. Connections in randomly oriented graphs. *Combinatorics, Probability and Computing*, pages 1–5, 2016.

[10] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:201–293, 2011.

Chapter 6

The equivocal "mean age of parents at birth"

Photo credit: Marc Manceau.

As mentioned in Section 1.3, this Chapter is not related to random graphs, and was not originally meant to be part of this thesis. It began after I met Mauricio González-Forero to talk about his work on the definition of species [13] and talk about the split-and-drift random graph presented in Chapter 2.

Because we both shared an interest in structured populations, we also discussed Mauricio's current work on the subject and he showed me some of his calculations. Intriguingly, the "mean age at reproduction" he had computed also corresponded to the mean age at death in the population. However, the problem did not come from Mauricio's calculations, but from the formula he had used. This was very surprising, because this well-known formula has been the standard for decades. So even though the problem with the formula itself was simple enough – it approximates the expectation of a ratio by the ratio of expectations – the fact that it could lead to incoherencies in practice was somewhat worrying.

I therefore derived formulas for a measure of the mean age at reproduction that was more conform to the intuition and compared this measure with the classic one, both on theoretical examples and in real-world models. The results were quite surprising, as the margin by which the classic measure was off was more often significant than not.

From a mathematical point of view, this work consists mostly of straightforward calculations involving Poisson point processes and is admittedly not very exciting. Its interest comes from the fact that it sheds a new light on one of the most basic measures used by biologists in a variety of concrete applications, ranging from conservation studies to the theoretical study of evolution.

Publication: This chapter has been published, in a slightly different format, in *The American Naturalist* under the title "The equivocal mean age of parents in a cohort" [1].

Chapter contents

6.1 Introduction . . . 147
6.2 Derivation and interpretation of the expressions of μ_1 and τ . . . 148
6.2.1 An explicit model for the population . . . 148
6.2.2 The mean age of parents of offspring produced by a cohort 149
6.2.3 The mean age at reproduction . . . 150
6.2.4 Computing τ in matrix population models . . . 152
6.3 Examples . . . 153
6.3.1 Theoretical examples . . . 153
6.3.2 Real-world examples . . . 155
6.4 Discussion . . . 157
Chapter references . . . 158
6.A Basic facts about Poisson point processes . . . 160
6.B Expression of τ for discrete age structures . . . 161
6.C Proof of $\mathbb{E}(\tilde{T}) \to \mathbb{E}(T^2)/\mathbb{E}(T)$ as $m \to 0$ . . . 163
6.D Proof of $\tau \leqslant \mu_1$ . . . 163
6.E Computing μ_1 and τ for Compadre/Comadre . . . 166
6.F Projection matrices for *A. mexicanum* . . . 167

6.1 Introduction

The mean age at reproduction is a central notion in the study of the evolution of reproductive timing and of the slow-fast continuum. It also plays an important role in demography. However, as with many descriptors of populations, it is not clear how it should be defined – let alone quantified in practice. A standard measure of it is the *mean age of parents of offspring produced by a cohort*, also frequently referred to as the *cohort generation time*. To obtain it, consider all offspring produced by a cohort of newborns over its lifetime; for each of these offspring, record the age that their parents (mother, in the case of a female-based model) had when the offspring was born; finally, take the average of these ages.

It is straightforward to compute this quantity from complete census data. In practice however, it is usually estimated from life-tables using the following formula:

$$\mu_1 \;=\; \frac{\int_0^{+\infty} t\, m(t)\, \ell(t)\, dt}{\int_0^{+\infty} m(t)\, \ell(t)\, dt}\,. \tag{6.1}$$

In this expression, the *survivorship function* ℓ gives the probability that an individual of the chosen cohort reaches age t, and the *age-specific fertility* m represents its rate of offspring production in such a way that, assuming the individual remains alive between ages a and b, the expected number of offspring it will produce in that interval of time is $\int_a^b m(t)\, dt$. There is also a discrete-time version of formula (6.1):

$$\mu_1 \;=\; \frac{\sum_{t=1}^{+\infty} t\, \ell_t\, m_t}{\sum_{t=1}^{+\infty} \ell_t\, m_t}\,, \tag{6.2}$$

where ℓ_t is the probability that an individual survives to age t and m_t is the expected number of offspring produced at age t by individuals who reach that age.

Formulas (6.1) and (6.2) go back a long way and are ubiquitous in the literature. They have been popularized by classic references such as [16] and [4] in demography, and [3] and [2] in biology. They can also be found in more recent works of reference, including [15], [21] and [18].

A consensus interpretation of μ_1 is that it represents the mean age at which a typical parent produces offspring. The aim of this chapter is to show that this interpretation is inaccurate and can be problematic in practice. To do so, I introduce a more direct measure of the mean age at reproduction of a typical parent. Consider a typical parent, and compute the average of the ages at which it gives birth to its offspring. The expected value of this average is what we term the *mean age at reproduction*. Under standard assumptions, it is given by

$$\tau \;=\; \frac{1}{c} \int_0^{+\infty} \frac{\int_0^t s\, m(s)\, ds}{\int_0^t m(s)\, ds} \left(1 - e^{-\int_0^t m(s)\, ds}\right) f(t)\, dt\,, \tag{6.3}$$

where f denotes the probability density function of the lifespan of an individual and the constant

$$c \;=\; \int_0^{+\infty} \left(1 - e^{-\int_0^t m(s)\, ds}\right) f(t)\, dt \tag{6.4}$$

is the fraction of individuals that produce offspring during their lifetime. As with μ_1, there is a discrete-time formula for τ:

$$\tau \;=\; \frac{1}{c} \sum_{t \geqslant 1} \frac{\sum_{s=1}^t s\, m_s}{\sum_{s=1}^t m_s} \left(1 - \prod_{s=1}^t e^{-m_s}\right) p_t\,, \tag{6.5}$$

where $p_t = \ell_t - \ell_{t+1}$ is the probability mass function of the lifespans of individuals and

$$c \;=\; \sum_{t\geqslant 1}\left(1 - \prod_{s=1}^{t} e^{-m_s}\right) p_t\,. \tag{6.6}$$

Using the expressions of μ_1 and of τ, we show that these two quantities can differ greatly, even in the most simple models. We also prove that μ_1 is always greater than τ, and that the difference between the two can be arbitrarily large. Finally, comparing the two measures numerically for 3871 real-world models from the Compadre and Comadre databases [7, 6], we obtain an average discrepancy of 20.6% and find than in one model out of four they differ by more than 30%.

6.2 Derivation and interpretation of the expressions of μ_1 and τ

In this section, we give a rigorous interpretation of the quantities μ_1 and τ. Our first step will be to lay out our assumptions about the dynamics of the population. This is rarely done formally in the sources presenting μ_1, which might explain why there has been some confusion about its interpretation.

6.2.1 An explicit model for the population

The setting that we use is that of a Crump-Mode-Jagers process [8, 9, 14], where the population consists of a discrete set of individuals such that:

(i) Each individual i has a random lifespan T_i with distribution ν and which is independent of everything else.

(ii) Individual i produces a new offspring at age t for every point of P_i at t such that $t \leqslant T_i$, where P_i is a point process with intensity m on $[0, +\infty[$ that is independent of everything else.

Note that the point processes P_i are not homogeneous (m is a function of the age of individuals) and that they do not have to be simple (an individual can give birth to several offspring simultaneously). For mathematical tractability however, it is often convenient to work with Poisson point processes (a brief introduction to Poisson point process can be found in Section 6.A of the Appendix). While the assumption that P_i are Poisson point processes is not needed in the study of μ_1, it will be required to derive explicit formulas for τ.

In this setting, the definition and interpretation of the survivorship function and of the age-specific fertility are straightforward. The survivorship is defined by [1]

$$\ell(t) \;=\; \mathbb{P}(T_i \geqslant t) \;=\; \nu\big([t, +\infty[\big)\,.$$

Working with the measure ν is convenient because it makes it possible to treat the case where T_i is a continuous random variable and the case where it is a discrete random variable simultaneously. However, in many applications T_i will have a density f. Thus, we will do most of our calculations with ν but express our final results in terms of f or ℓ, as in formulas (6.1) and (6.3).

[1] In probability theory and statistics, the *survival function* almost invariably refers to the complementary cumulative distribution function of T_i, $t \mapsto \mathbb{P}(T_i > t)$. Here, however, we will stick to the convention used in biology.

The age-specific fertility is the function m. If we denote by $M_i(a,b)$ the integer-valued random variable corresponding to the number of offspring produced by i between ages a and b, then assuming that $b \leqslant T_i$ we have, as expected,

$$\mathbb{E}(M_i(a,b)) = \int_a^b m(t)\,dt\,,$$

Obviously, the framework of Crump-Mode-Jagers processes has some serious restrictions. For instance, it assumes that individuals are independent and thus excludes any kind of density dependence. Similarly, the (optional) assumption that individuals reproduce at rate m is constraining, and in particular implies that they cannot produce several offspring simultaneously. Nevertheless, this framework is close to the minimal setting containing all the ingredients needed to define most descriptors of populations, whilst being simple enough to remain tractable and make it possible to derive explicit formulas for these descriptors. Moreover, the hypotheses above correspond quite well to the assumptions that are made, typically implicitly, to obtain the classic expressions of many of descriptors of populations.

Finally, to obtain discrete-time equivalents of formulas (6.1) and (6.3) we will need to consider the following version of the model, which allows simultaneous births: we keep assumption (i) under the extra hypothesis that the lifespan T_i is an integer-valued random variable, and we replace (ii) by the assumption that at each age $t = 1, \dots, T_i$, individual i gives birth to $M_t^{(i)}$ new individuals. Again, this corresponds quite well to the usual hypotheses on which many classic formulas rely.

6.2.2 The mean age of parents of offspring produced by a cohort

We now give a rigorous interpretation of the quantity μ_1 given by formulas (6.1) and (6.2). As we will see, this interpretation is more subtle than what is usually assumed. This is because μ_1 does not correspond to the expected value of the average of the ages of the parents of the offspring produced by a cohort, but only to the limit of this average when the size of this cohort goes to infinity.

Let $\mathcal{C}$ denote a cohort, that is, a set of n individuals considered from the time of their birth to the time of their death. Let T_i be the lifespan of individual i, and P_i be the set of ages at which it produces offspring. Note that in our setting, conditional on T_i, P_i is a point process with intensity m on $[0, T_i]$.

The average of the ages of the parents of the offspring produced by the cohort over its lifetime is

$$Z_{\mathcal{C}} \;=\; \frac{\sum_{i\in\mathcal{C}} \sum_{t\in P_i} t}{\sum_{i\in\mathcal{C}} \sum_{t\in P_i} 1} \;=\; \frac{\sum_{i\in\mathcal{C}} S_i}{\sum_{i\in\mathcal{C}} N_i}\,,$$

where $N_i = \sum_{t\in P_i} 1$ is the number of offspring produced by individual i, and $S_i = \sum_{t\in P_i} t$ is the sum of the ages at which it produces them. Note that $Z_{\mathcal{C}}$ is well-defined only when $\sum_{i\in\mathcal{C}} N_i > 0$, but that this happens with probability arbitrarily close to one for a large enough cohort.

As we have already seen, the expected number of offspring produced by an individual i whose lifespan is $T_i = t$ is

$$\int_0^t m(s)\,ds\,.$$

This quantity can be thought of as "$\mathbb{E}(N_i \mid T_i = t)$", even though this interpretation is subject to some caution. At any rate, it follows that

$$\mathbb{E}(N_i) = \int_0^{+\infty} \left(\int_0^t m(s)\,ds \right) d\nu(t)\,.$$

Moreover, using Fubini's theorem,

$$\int_0^{+\infty} \left(\int_0^t m(s)\, ds \right) d\nu(t) \;=\; \int_0^{+\infty} m(s) \left(\int_s^{+\infty} d\nu(t) \right) ds\,.$$

Using that $\int_s^{+\infty} d\nu(t) = \ell(s)$, we get the well-known expression for R_0, the *mean number of offspring produced by an individual during its lifetime*:

$$R_0 \;=\; \mathbb{E}(N_i) \;=\; \int_0^{+\infty} m(t)\, \ell(t)\, dt$$

Using Campbell's formula (equation (6.11) in Appendix 6.A) and the exact same reasoning, we can express the *mean sum of the ages at which an individual produces offspring* as

$$\mathbb{E}(S_i) \;=\; \int_0^{+\infty} t\, m(t)\, \ell(t)\, dt$$

Now let N (resp. S) denote a random variable that has the common distribution of the variables N_i (resp. S_i). Then, as pointed out in most sources presenting the measure μ_1, we have

$$\mu_1 \;=\; \frac{\mathbb{E}(S)}{\mathbb{E}(N)}\,.$$

This however does not establish a link between μ_1 and $Z_{\mathcal{C}}$, the average age of the parents of offspring produced by the cohort. To see how these two quantities are related, observe that since the variables N_i (resp. S_i) are independent, if we denote by $n = \mathrm{Card}(\mathcal{C})$ the size of the cohort then by the law of large numbers, as $n \to +\infty$,

$$\frac{1}{n} \sum_{i \in \mathcal{C}} N_i \xrightarrow[n \to +\infty]{\text{a.s.}} \mathbb{E}(N) \quad \text{and} \quad \frac{1}{n} \sum_{i \in \mathcal{C}} S_i \xrightarrow[n \to +\infty]{\text{a.s.}} \mathbb{E}(S)\,.$$

As a result,

$$Z_{\mathcal{C}} \;=\; \frac{\frac{1}{n} \sum_{i \in \mathcal{C}} S_i}{\frac{1}{n} \sum_{i \in \mathcal{C}} N_i} \;\xrightarrow[n \to +\infty]{\text{a.s.}}\; \mu_1\,.$$

Importantly, note that μ_1 is *not* the expected value of S_i/N_i or of $Z_{\mathcal{C}}$. In fact, the expected value of S_i/N_i (conditional on this variable being well-defined) is precisely what we call the mean age at reproduction. We explain how to compute it in the next section.

6.2.3 The mean age at reproduction

Recall that we would like the mean age at reproduction τ to represent the mean age at which a typical parent produces offspring. Formally, assuming that individual i has some offspring, the average age at which it produces them is

$$\bar{X}_i = \frac{1}{N_i} \sum_{t \in P_i} t\,,$$

where, as before, N_i is the total number of offspring produced by i and P_i is the set of ages at which it produces them. We thus define the mean age at reproduction as

$$\tau = \mathbb{E}\Big(\bar{X}_i \;\Big|\; N_i > 0 \Big)\,,$$

which, given our assumptions, does not depend on i or on the composition of the population.

Note that if I is a "typical parent", i.e. is sampled uniformly among the individuals that produce offspring during their lifetime, we have

$$\mathbb{E}\Big(\bar{X}_i \,\Big|\, N_i > 0\Big) = \mathbb{E}\Big(\bar{X}_I\Big)\,.$$

Moreover, letting $\tilde{T}$ denote the lifespan of I, $\bar{X}_I$ is the average of a point process with intensity m on $[0, \tilde{T}]$. As explained in Section 6.A of the Appendix, in the case of a Poisson point process, the expected value of this average is simply the expected value of a random point of $[0, \tilde{T}]$ with density $t \mapsto m(t) / \int_0^{\tilde{T}} m(s)\,ds$. The remarkable fact that it does not depend on the value of N_i is a consequence of the absence of internal structure of Poisson point processes. From this, we get

$$\mathbb{E}\Big(\bar{X}_I \,\Big|\, \tilde{T}\Big) \;=\; \frac{\int_0^{\tilde{T}} s\, m(s)\,ds}{\int_0^{\tilde{T}} m(s)\,ds}\,.$$

As a result,

$$\tau \;=\; \int_0^{+\infty} \frac{\int_0^t s\, m(s)\,ds}{\int_0^t m(s)\,ds}\, d\tilde{\nu}(t)\,,$$

where $\tilde{\nu}$ is the law of the lifespan $\tilde{T}$ of I. Note that it is different from ν, the lifespan of a fixed individual, because conditioning on the fact that an individual produces offspring biases its lifespan; for instance, if – as frequently the case in real applications – there exists an age α such that $m(t) = 0$ for $t < \alpha$, then individuals that produce offspring all live longer than α, whereas it is not necessarily the case for other individuals.

The last thing that we need to do in order to get an explicit formula for τ is thus to determine $\tilde{\nu}$. For this, note that

$$\begin{aligned}\mathbb{P}\Big(\tilde{T} \leqslant t\Big) &= \mathbb{P}(T_i \leqslant t \mid N_i > 0)\\ &= \frac{\mathbb{P}(T_i \leqslant t, N_i > 0)}{\mathbb{P}(N_i > 0)}\,.\end{aligned}$$

Conditioning on T_i, using the void probabilities of Poisson point processes for the probability that an individual with lifetime s produces some offspring, and finally integrating against ν, we get

$$\mathbb{P}(T_i \leqslant t, N_i > 0) = \int_0^t \Big(1 - e^{-\int_0^s m(r)\,dr}\Big)\, d\nu(s)\,.$$

As a result,

$$d\tilde{\nu}(t) \;=\; \frac{1}{c}\Big(1 - e^{-\int_0^t m(s)\,ds}\Big)\, d\nu(t)\,,$$

where the constant $c = \mathbb{P}(N_i > 0)$ is given by

$$c \;=\; \int_0^{+\infty} \Big(1 - e^{-\int_0^t m(s)\,ds}\Big)\, d\nu(t)\,.$$

Note that, by integrating by parts and using that $\ell(t) \to 0$ as $t \to +\infty$, we can also express c directly in terms of ℓ and m as

$$c \;=\; \int_0^{+\infty} e^{-\int_0^t m(s)\,ds}\, m(t)\, \ell(t)\,dt\,.$$

Putting the pieces together in the case where T_i has a density f, we get formula (6.3):

$$\tau \;=\; \frac{1}{c}\int_0^{+\infty} \frac{\int_0^t s\, m(s)\, ds}{\int_0^t m(s)\, ds}\left(1 - e^{-\int_0^t m(s)\, ds}\right) f(t)\, dt\,.$$

Note that neither the biological interpretation of τ nor the derivation of its expression depend on the assumption that individuals are independent.

Sometimes, especially when studying evolution, one is interested in the average of a function z of the ages at which a parent produces offspring, rather than in the average of the ages themselves.[2] In that case, letting A be uniformly chosen among the ages at which a typical parent produces offspring, for every function z,

$$\mathbb{E}(z(A)) \;=\; \frac{1}{c}\int_0^{+\infty} \frac{\int_0^t z(s)\, m(s)\, ds}{\int_0^t m(s)\, ds}\left(1 - e^{-\int_0^t m(s)\, ds}\right) f(t)\, dt\,,$$

with the constant c given in equation (6.4). This is proved by working with

$$\bar{W}_i \;=\; \frac{1}{N_i}\sum_{t\in P_i} z(t)$$

instead of $\bar{X}_i$, and using equation (6.10) instead of equation (6.9) to get

$$\mathbb{E}\left(\bar{W}_I \,\middle|\, \tilde{T}\right) \;=\; \frac{\int_0^{\tilde{T}} z(s)\, m(s)\, ds}{\int_0^{\tilde{T}} m(s)\, ds}\,.$$

Finally, the justification of the expression of τ for discrete age structures can be found in Section 6.B of the Appendix. It essentially consists in approaching the discrete-time model with the continuous-time one by choosing appropriate age-specific fertilities, and relies on the assumption that the number of offspring produced each year by each individual follows a Poisson distribution. It should also be pointed out that, because in the discrete-time setting individuals can produce several offspring simultaneously, there are two possibilities to define the average age at offspring production: counting all births equally, or weighting them by the number of offspring produced. Formula (6.5) is obtained by weighting the ages by the number of offspring produced when averaging them.

6.2.4 Computing τ in matrix population models

Matrix population models are widely used quantitative models of structured population dynamics. Their mathematical tractability and ease of use in real-world applications makes them a tool of choice for many biologists and demographers. The framework of matrix population models is the following: consider a population structured in discrete classes $i = 1, \ldots, m$ and assume that the the per-capita contribution of individuals in class j at time t to the composition of class i at time $t+1$ is a_{ij}. Then, letting $n_i(t)$ denote the number of individuals in class i at time t, the dynamics of the population are governed by the equation

$$\mathbf{n}(t+1) = \mathbf{A}\mathbf{n}(t)\,,$$

where $\mathbf{n}(t) = (n_i(t))$ is called the *population vector* and $\mathbf{A} = (a_{ij})$ the *population projection matrix*.

[2] This was pointed out by Mauricio González-Forero.

What makes these models interesting from a mathematical point of view is their connection to the theory of nonnegative matrices (and especially the Perron-Frobenius theorem) and their connection to multitype branching processes. For a presentation of matrix population models, see e.g. [2].

An expression of μ_1 is available for matrix population models: if we let $\mathbf{S}$ be the survival matrix and $\mathbf{F}$ be the fertility matrix (i.e. if we decompose the projection matrix $\mathbf{A}$ into $\mathbf{A} = \mathbf{S}+\mathbf{F}$ to separate survival probabilities from fertilities) and denote by $\mathbf{w}$ the stable distribution of the population (the dominant right-eigenvector of $\mathbf{A}$) and $\mathbf{e} = (1, \dots, 1)$ the row vector consisting only of ones, then we have the following modern version of the classic formula of [5], which can be found in [11]:

$$\mu_1 \;=\; \frac{\mathbf{eF}(\mathbf{I}-\mathbf{S})^{-2}\mathbf{Fw}}{\mathbf{eF}(\mathbf{I}-\mathbf{S})^{-1}\mathbf{Fw}}\,. \tag{6.7}$$

Note that $(\mathbf{I}-\mathbf{S})^{-1} = \sum_{t\geqslant 1}\mathbf{S}^{t-1}$ and that $(\mathbf{I}-\mathbf{S})^{-2} = \sum_{t\geqslant 1} t\,\mathbf{S}^{t-1}$, so that this expression closely parallels (6.2). The entries of $\mathbf{e}$ represent the weight given to each type of offspring when computing the average age of the parents. Should we wish to give more importance to some offspring type, any vector with positive entries could be used in place of $\mathbf{e}$ – in fact [5] suggest using the reproductive values as weights. See [11] for more on this.

To obtain an equivalent of formula (6.5) giving τ in the contex of matrix population models, one would need to (1) find the law of the conditional trajectory of an individual in the life cycle given that it produces offspring and (2) integrate the average of the ages at which it produces offspring against this law. While the first of these steps is feasible [3], it is unclear whether the second is – and whether the resulting expression, if it could be obtained, would be simple enough to be useful.

Nevertheless, the definition of τ as the mean age at which a typical parent produces offspring can easily be transposed to the framework of matrix population models, making it straightforward to estimate it via individual-based simulations. This is detailed in Appendix 6.E

6.3 Examples

In order to get a sense about how the measures μ_1 and τ differ, let us compare them on some examples, using both theoretical and real-world models. More general mathematical results, such as the fact that $\tau \leqslant \mu_1$ that was announced in the introduction, will be proved in the Appendix.

6.3.1 Theoretical examples

Let us start with a simple but fundamental example, where individuals reproduce at constant rate m. In that case,

$$\mu_1 \;=\; \frac{\mathbb{E}\left(\int_0^T ms\,ds\right)}{\mathbb{E}\left(\int_0^T m\,ds\right)} \;=\; \frac{1}{2}\frac{\mathbb{E}(T^2)}{\mathbb{E}(T)}$$

and

$$\tau \;=\; \mathbb{E}\left(\frac{\int_0^{\tilde{T}} ms\,ds}{\int_0^{\tilde{T}} m\,ds}\right) \;=\; \frac{1}{2}\,\mathbb{E}\left(\tilde{T}\right).$$

[3] This was explained to me by Stephen Ellner – see e.g. Chapter 3 of [12].

The expression of τ is unsurprising: when birth events are uniformly distributed on the lifetime of individuals, on average they occur in the middle of their life. Also, since

$$\mathbb{E}\Big(\tilde{T}\Big) \;=\; \frac{\mathbb{E}\Big(T\Big(1-e^{-mT}\Big)\Big)}{\mathbb{E}(1-e^{-mT})}\,,$$

and that for all $t>0$, $1-e^{-mt}$ increases to 1 as m goes to infinity, it follows from the monotone convergence theorem that

$$\mathbb{E}\Big(\tilde{T}\Big) \to \mathbb{E}(T) \quad \text{as } m \to +\infty\,.$$

By a similar argument (see Appendix 6.C), we also have

$$\mathbb{E}\Big(\tilde{T}\Big) \to \frac{\mathbb{E}(T^2)}{\mathbb{E}(T)} \quad \text{as } m \to 0\,.$$

Furthermore, since $\mathbb{E}(\tilde{T})$ is a decreasing function of m, we conclude that when individuals reproduce at a constant rate,

$$\tfrac{1}{2}\mathbb{E}(T) \;\leqslant\; \tau \;\leqslant\; \mu_1\,.$$

In fact, the inequality $\tau \leqslant \mu_1$ holds for general age-specific fertility functions: see Proposition 6.D.2 in Appendix 6.D.

To make this example more concrete, let us further assume that individuals die at constant rate η, so that T is an exponential variable and that $\ell(t)=e^{-\eta t}$. In that case, we get

$$\mu_1 = \frac{1}{\eta} \quad \text{and} \quad \tau = \frac{1}{2\,\eta}\left(1+\frac{1}{1+m/\eta}\right). \tag{6.8}$$

Note that here μ_1 is also equal to the expected lifespan in the population. Interpreting it as the mean age at which parents reproduce would therefore lead to a contradiction, because – in the case where the fertility m is large enough, so that most individuals get to reproduce during their lifetime and that the lifespan of a typical parent is not very different from that of a typical individual – this would imply that, on average, the age at which an individual reproduces is the same as the age at which it dies. This is absurd, because unless individuals reproduce exactly when they die, the former has to be smaller than the latter.

From (6.8), we also see that for m/η large enough, $\mu_1 \approx 2\tau$. For $m=\eta$, which corresponds to the minimum ratio m/η for a viable population, the difference is already 25% of the value of μ_1. The relative difference between μ_1 and τ as a function of m/η is plotted in Figure 6.1.

Now consider the closely related discrete-time model where individuals survive from one year to the other with probability p and produce Poisson(m) offspring at each age $t \geqslant 1$, so that

$$p_t = (1-p)\,p^t \quad \text{and} \quad \ell_t = p^t\,,$$

After straightforward calculations, we find that the numerator in formula (6.2), which corresponds to the mean sum of the ages at childbirth, is $mp/(1-p)^2$ and that the denominator is $mp/(1-p)$. As a result,

$$\mu_1 = \frac{1}{1-p}\,.$$

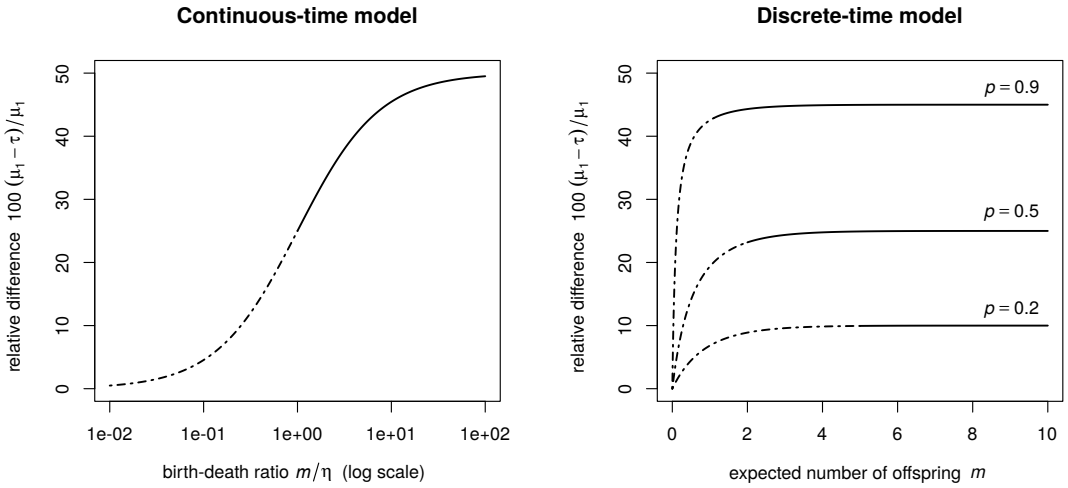


Figure 6.1: Relative difference between μ_1 and τ as a function of the parameters of the models considered. Left, the continuous-time model in which individuals give birth at constant rate m and die at constant rate η. Right, the discrete-time model in which they survive from one year to the other with probability p and give birth to Poisson(m) offspring each year. Dashed lines indicate values of the parameters for which the population is not viable in the long term.

Note that this model can also be seen as a 1×1 matrix population model with survival matrix $\mathbf{S} = (p)$ and fertility matrix $\mathbf{F} = (m)$, so formula (6.7) can also be used and gives the same result.

Because $\mathbb{E}(T) = p/(1-p)$, we see that

$$\mu_1 \;=\; \mathbb{E}(T) + 1\,,$$

which also corresponds to the expected lifespan of individuals that reach age 1. For the same reason as before, this implies that μ_1 is not credible as an estimate of the mean age at which a typical parent produces offspring.

Finally, after routine calculations we find that

$$\tau = \frac{1}{2}\left(\frac{1}{1-p} + \frac{1}{1-p\,e^{-m}}\right).$$

As previously, $\frac{1}{2}\mu_1 \leqslant \tau \leqslant \mu_1$, but the difference between μ_1 and τ can be quite high, even for very reasonable values of p and m: for instance, with $p = 0.5$ and $m = 2$ both measures differ by 23% of the value of μ_1; for $p = 0.9$ and $m = 2$, by 44%. Again, this is illustrated in Figure 6.1.

6.3.2 Real-world examples

The examples of the previous section show that μ_1 and τ can be very different, even in the most simple models. But do they differ significantly in practice? To answer this question, μ_1 and τ were calculated for every model of the COMPADRE Plant Matrix Database [7] and COMADRE Animal Matrix Database [6] for which this could be done. Because there is no formula for τ in matrix population models, it was estimated numerically in such a way that, for each estimated value, the width of the 95% confidence interval was less than 2% of the estimated value itself (see Section 6.E of the Appendix for details). Figure 6.2 gives the distribution of the relative difference between the two quantities, computed as $\Delta_{\%} = 100(\mu_1 - \tau)/\mu_1$,

and Table 6.1 lists some statistics of this distribution. These conclusively show that the measures μ_1 and τ differ significantly for most real-world models. In particular, the fact that the median of $(\mu_1 - \tau)/\mu_1$ is of order 20% means that, by using μ_1 to quantify the mean age at reproduction, one overestimates its actual value by more than 25% in half of the cases.

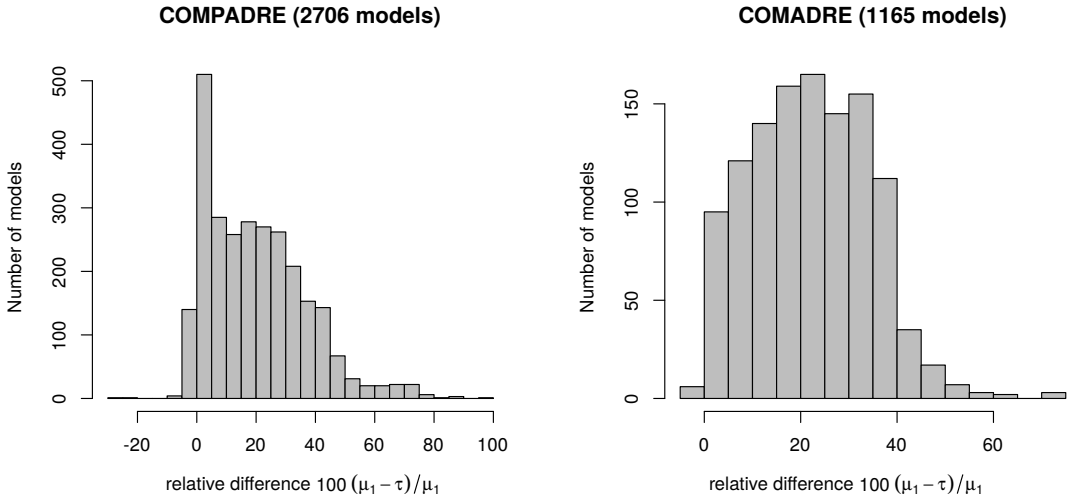


Figure 6.2: Distribution of the relative difference between μ_1 and τ for the Compadre and Comadre databases. The difference is given as a percentage of μ_1; for instance, a 30% difference means that $\tau = 0.7\mu_1$.

	Mean	1st quartile	Median	3rd quartile
Compadre	19.97	05.26	17.73	30.49
Comadre	22.16	12.54	22.60	31.14

Table 6.1: Statistics of the distribution of $(\mu_1 - \tau)/\mu_1$ for the Compadre and Comadre databases. All values are percentages.

For a detailed example of a real-world model in which μ_1 and τ differ greatly, see Appendix 6.F. This example is particularly interesting because it illustrates the fact that μ_1 can be greater than the expected lifespan conditional on reproduction, which decisively rules out its interpretation as the mean age at reproduction.

Before closing this section, let us comment on the fact that some models (152 out of 3871) appear to have $\tau < \mu_1$. These are in fact models for which τ is very close to μ_1, but because of the uncertainty in its estimation appears to be slightly smaller than it. Indeed, for most of these models $\mu_1 - \tau$ is very close to zero (only ten of them have a relative difference such that $|\Delta_{\%}| > 1\%$). All things considered, the fact that μ_1 lies below the 95% confidence interval of τ for only 0.46% of all models is consistent with the fact that $\tau \leqslant \mu_1$ (it would have to be more than 2.5% to constitute a contradiction).

Finally, the excess of models for which $\mu_1 \approx \tau$ in Compadre compared to Comadre is due to (mostly 2×2) models with very short generation times, presumably corresponding to annuals plants in which the lifespans of individuals exhibit little to no variation.

6.4 Discussion

The mean age of the parents of the offspring produced by a cohort μ_1 and the mean age at reproduction τ are two genuinely different notions. So why have they not been recognized as such before? Probably because precise definitions of these quantities are seldom given. For instance, in the references given above – which are or have been among the most influential in the field – μ_1 is variously described as the "mean age at childbearing in the stationary population" [4] by [16]; as the "mean age of childbearing in a cohort" by [4, eq. (2.10) p. 19]; as the "mean age at reproduction of a cohort of females" by [3, eq. (1.47a) p. 30]; and as the "mean age of the mothers at the time of their daughter's birth" by [21, eq. (4.12) p. 98]. Yet these four definitions fail to detail how this "mean" should be computed, and could thus be thought to refer to τ.

It is not obvious from the definitions of μ_1 and τ how these two quantities are related – or indeed why they should differ at all. One helpful way to think about it is the following: μ_1 can be seen as an *offspring-centric* measure of the mean age of parents, whereas τ is a *parent-centric* measure of it. Indeed, to compute μ_1 we ask each newborn produced by a cohort "how old is your parent?", while for τ we ask a parent "how old are you going to be when you have offspring?" These questions have distinct answers because they correspond to two different ways to sample a parent.

Among other things, this explains why μ_1 is greater than τ: indeed, parents that live longer tend to have more offspring, and thus have a higher probability of being sampled via their offspring than when the sampling is done uniformly at random. As a result, they contribute more to μ_1 than to τ. Since these parents with longer lifespans are also those that tend to have a higher mean age at reproduction, this biases μ_1 upward compared to τ.

This also explains why the difference $\mu_1 - \tau$ goes to zero as the fertility becomes vanishingly small (see Appendix 6.C): in that case, the proportion of parents that give birth to more than one offspring during their lifetime goes to zero, and as the result the two parent-sampling schemes become equivalent.

To close this series of remarks regarding the link between μ_1 and τ, observe that, from a purely mathematical point of view, the difference between the two can be made arbitrarily large. Indeed, recall that, when individuals reproduce at a constant rate m, $\mu_1 = \mathbb{E}(T^2)/\mathbb{E}(T)$ and $\tau \to \frac{1}{2}\mathbb{E}(T)$ as $m \to +\infty$. Thus, by choosing an appropriate distribution for the lifespan T and taking m large enough, we can make μ_1 arbitrarily large and τ arbitrarily small.

Now that we have seen that μ_1 and τ are two different concepts, that they differ significantly in practice, and that we better understand the link between them, one important question remains: which of μ_1 or τ should be favored in which context?

From a practical point of view, the expressions of τ are, admittedly, more complex than those of μ_1. This of course is not a problem for real-world applications, where they are going to be evaluated numerically; for theoretical applications however, this does make exact calculations harder, if possible at all.

Another important difference between both measures is their slightly different domain of validity. While the interpretation of μ_1 hinges on the assumption that there are no interactions between individuals, the expression of τ relies on that of

[4] What Keyfitz calls the *stationary population* is actually a cohort.

Poissonian births. One might cynically argue that this is hardly a problem, because both hypotheses are often used jointly in theoretical models, and never met in real-world applications. Nevertheless, there is a real difference here that should be taken into account when deciding which measure to choose.

Lastly, τ has the advantage of having a more direct interpretation than μ_1. Judging from the phrasing used by several authors, it seems that it is sometimes τ they have in mind, even when working with μ_1. Moreover, the interpretation of μ_1 might not be as intuitive as we usually assume; notably, the fact that it can be greater not only than the expected lifespan but also than the expected lifespan conditional on reproduction (as illustrated by the *Medium density* scenario for *Astrocaryum mexicanum* in Section 6.F of the Appendix) is likely to come as a surprise to many ecologists and demographers.

Literature cited in this chapter

[1] F. Bienvenu. The equivocal mean age of parents in a cohort. *The American Naturalist*, 194(2), 2019.

[2] H. Caswell. *Matrix Population Models: Construction, Analysis, and Interpretation*. Sinauer Associates, 2nd edition, 2001.

[3] B. Charlesworth. *Evolution in Age-Structured Populations*. Cambridge Studies in Mathematical Biology. Cambridge University Press, 2nd edition, 1994.

[4] A. J. Coale. *The Growth and Structure of Human Populations*. Princeton University Press, 1972.

[5] M. E. Cochran and S. P. Ellner. Simple methods for calculating age-based life history parameters for stage-structured populations. *Ecological Monographs*, 62(3):345–364, 1992.

[6] Comadre Animal Matrix Database. Max Planck Institute for Demographic Research (Germany). Available at www.compadre-db.org, Data downloaded on 02/02/2019, version 2.0.1.

[7] Compadre Plant Matrix Database. Max Planck Institute for Demographic Research (Germany). Available at www.compadre-db.org, Data downloaded on 02/02/2019, version 4.0.1.

[8] K. S. Crump and C. J. Mode. A general age-dependent branching process I. *Journal of Mathematical Analysis and Applications*, 24(3):494–508, 1968.

[9] K. S. Crump and C. J. Mode. A general age-dependent branching process II. *Journal of Mathematical Analysis and Applications*, 25(1):8–17, 1969.

[10] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Probability and its Applications. Springer-Verlag New York, 2nd edition, 2003.

[11] S. P. Ellner. Generation time in structured populations. *The American Naturalist,*, 192(1):105–110, 2018.

[12] S. P. Ellner, D. Z. Childs, and M. Rees. *Data-driven modelling of structured populations*. Springer International Publishing, 2016.

[13] M. González-Forero. Removing ambiguity from the biological species concept. *Journal of theoretical biology*, 256(1):76–80, 2009.

[14] P. Jagers. A general stochastic model for population development. *Scandinavian Actuarial Journal*, 1969(1-2):84–103, 1969.

[15] S. E. Jørgensen and B. Fath. *Encyclopedia of Ecology*. Elsevier Science, 1st edition, 2008.

[16] N. Keyfitz. *Introduction to the Mathematics of Population*. Addison-Wesley, 1968.

[17] J. F. C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992.

[18] R. M. Kliman. *Encyclopedia of Evolutionary Biology*. Academic Press, 2016.

[19] S. Legendre and J. Clobert. ULM, a software for conservation and evolutionary biologists. *Journal of Applied Statistics*, 22(5-6):817–834, 1995. The ULM software can be downloaded at http://www.biologie.ens.fr/~legendre/ulm/ulm.html.

[20] D. Pinero, M. Martinez-Ramos, and J. Sarukhan. A population model of astrocaryum mexicanum and a sensitivity analysis of its finite rate of increase. *The Journal of Ecology*, 72(3):977–991, 1984.

[21] L. L. Rockwood. *Introduction to Population Ecology*. Wiley–Blackwell, 2nd edition, 2015.

Appendices to Chapter 6

6.A Basic facts about Poisson point processes

This section was originally written as an online supplement to the article that forms the basis of this chapter. Its aim was to present the basics of Poisson point processes to biologists, we focusing on the properties on which our calculations rely. Thus no attempt is made at stating the results in full generality, and we do not preoccupy ourselves with technical conditions such as measurability. For a detailed presentation of Poisson point processes, see e.g. [17] or [10].

It is common in modelling to assume that an event *occurs at rate* $r(t)$ *at time* t. Loosely speaking, this means that the probability that the event happens between t and $t + dt$ is independent of its previous occurrences, and is approximately $r(t)\,dt$. The rigorous way to formalize this is to say that the events are distributed according to an (inhomogeneous) Poisson point process with intensity r. Such a process can be seen as a random set of points characterized by the following properties: writing $N(I)$ for the number of points that fall in a fixed set $I \subset \mathbb{R}$,

(i) $N(I)$ is a Poisson random variable with mean $\int_I r(t)\,dt$.

(ii) $N(I)$ and $N(J)$ are independent whenever I and J are disjoint.

Note that the following useful fact is an immediate consequence of (i):

$$\mathbb{P}(N(I) = 0) = \exp\left(-\int_I r(t)\,dt\right).$$

Property (ii), often known as the *independent scattering* property, essentially says that Poisson point processes have a "completely random" structure.

From now on, we consider a fixed set $I \subset \mathbb{R}$ such that $\int_I r(t)\,dt < +\infty$. We let P be a Poisson point process with intensity r on I and denote by $N = \mathrm{Card}(P)$ its number of points. Let X be a random point of I with density $t \mapsto r(t)/\int_I r(t)\,dt$, i.e. whose distribution is characterized by

$$\forall A \subset I, \quad \mathbb{P}(X \in A) = \frac{\int_A r(t)\,dt}{\int_I r(t)\,dt},$$

and note in passing that

$$\mathbb{E}(X) = \frac{\int_I t\, r(t)\,dt}{\int_I r(t)\,dt}.$$

An important property of Poisson point processes is that, conditional on the event $\{N = n\}$, P consists of n independent copies of X – that is, for every function φ,

$$\mathbb{E}(\varphi(P) \mid N = n) = \mathbb{E}\big(\varphi(\{X_i : i = 1, \dots, n\})\big)\,,$$

where X_i, $i = 1, \dots, n$, are independent copies of X. A consequence of this is that the expected value of the average of the points in P is $\mathbb{E}(X)$. Formally, if $N > 0$ then we can define a random variable $\bar{X}$ by

$$\bar{X} = \frac{1}{N} \sum_{t \in P} t\,.$$

We then have

$$\mathbb{E}\Big(\bar{X} \,\Big|\, N > 0\Big) = \mathbb{E}(X)\,. \tag{6.9}$$

Indeed,

$$\mathbb{E}\Big(\bar{X} \,\Big|\, N > 0\Big) \;=\; \frac{1}{\mathbb{P}(N > 0)} \sum_{n \geqslant 1} \mathbb{E}\Big(\bar{X} \,\Big|\, N = n\Big)\mathbb{P}(N = n)\,,$$

and, for every $n \geqslant 1$,

$$\mathbb{E}\Big(\bar{X} \,\Big|\, N = n\Big) \;=\; \mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) \;=\; \mathbb{E}(X)\,.$$

In fact, given a function f, the exact same reasoning can be applied to

$$\bar{W} \;=\; \frac{1}{N} \sum_{t \in P} f(t)$$

to show that

$$\mathbb{E}\Big(\bar{W} \,\Big|\, N > 0\Big) = \mathbb{E}(f(X))\,. \tag{6.10}$$

We close this short overview with a fundamental result known as Campbell's formula. This formula states that, for every function f,

$$\mathbb{E}\left(\sum_{t \in P} f(t)\right) \;=\; \int_I f(t)\, r(t)\, dt\,. \tag{6.11}$$

In contrast to (6.9) and (6.10), which are consequences of the independent scattering property, Campbell's formula is not specific to Poisson point processes.

6.B Expression of τ for discrete age structures

In discrete time, individual i has an integer-valued lifespan T_i and, at each age $t = 1, \dots, T_i$, produces $M_t^{(i)}$ new individuals, where the variables $M_t^{(i)}$ are integer-valued and independent of everything else. Here we will also need to assume that each variable $M_t^{(i)}$ is a Poisson random variable with mean m_t.

In that setting, the average $\bar{X}_i$ of the ages at which individual i produces offspring can be defined as

$$\bar{X}_i = \frac{1}{N_i} \sum_{t=1}^{T_i} t\, M_t^{(i)}\,,$$

this definition being valid only when $N_i = \sum_{t=1}^{T_i} M_t^{(i)} > 0$. Note that, in this expression, each age at which i produces offspring is weighted by the number of offspring produced. This is similar to what is done for μ_1, where each offspring

contributes to the average age of the parents. But another possibility would be to weight all ages equally, that is, use the variable

$$\bar{Y}_i = \frac{\sum_{t=1}^{T_i} t\, I_t^{(i)}}{\sum_{t=1}^{T_i} I_t^{(i)}}\,,$$

where $I_t^{(i)} = 1$ if $M_t^{(i)} > 0$ and 0 otherwise.

Since $\bar{X}_i = \bar{Y}_i$ when individuals cannot give birth to several offspring simultaneously (or, more generally, when the number of offspring produced is either 0 or some constant m), the two definitions were equivalent in the continuous-time setting. But now, $\bar{X}_i$ and $\bar{Y}_i$ are two different and legitimate candidates for the "average age at which i produces offspring". However, $\bar{Y}_i$ does not lend itself to analysis as easily as $\bar{X}_i$ and to obtain formula (5) – which is arguably the natural discrete-time equivalent of formula (3) – it is $\bar{X}_i$ that should be used. Therefore, we define τ to be $\mathbb{E}(\bar{X}_i \mid N_i > 0)$.

The reasoning that lead to (3) could be adapted to obtain an expression for τ. However, it is also possible to deduce this expression directly from our results in continuous time. Indeed, the calculations of Section C of the Appendix are valid for general lifespans, including discrete ones: when ν is discrete, we simply have for any function φ

$$\int_0^{+\infty} \varphi(t)\, d\nu(t) \;=\; \sum_{t\geqslant 1} \varphi(t)\, p_t\,,$$

where $p_t = \mathbb{P}(T_i = t)$.

Moreover, observe that if we let the age-specific fertility m be the piecewise constant function defined by

$$m(t) = \sum_{s\geqslant 1} m_s \mathbb{1}_{]s-1,s]}(t)\,,$$

where $\mathbb{1}_{]s-1,s]}$ is the function that evaluates to 1 if $t \in]s-1, s]$ and 0 otherwise, then the number of offspring produced by an individual between ages $(t-1)$ and t is a Poisson variable with parameter m_t. Thus, the only difference with the discrete setting is that the ages at which these offspring are produced are uniformly distributed in $]t-1, t]$ instead of all equal to t.

Now, if we take the age-specific fertility to be the function $m^{(\varepsilon)}$ defined by

$$m^{(\varepsilon)}(t) = \sum_{s\geqslant 1} \frac{m_s}{\varepsilon} \mathbb{1}_{]s-\varepsilon,s]}(t)\,,$$

then the number of offspring produced between ages $(t-1)$ and t is still a Poisson variable with parameter m_t, but this time the ages at which these offspring are produced are uniformly distributed in $]t-\varepsilon, t]$. Taking ε to zero, the mean age at childbirth will therefore tend to that of the discrete-time model. We spare the reader the straightforward but somewhat technical argument by which this can be made rigorous. Noting that, for continuous functions g,

$$\int_0^t g(s)\, m^{(\varepsilon)}(s)\, ds \;\xrightarrow[\varepsilon\to 0]{}\; \sum_{s=1}^{t} g(s)\, m_s\,,$$

we obtain the following discrete-time equivalent of (3):

$$\mu_1 \;=\; \frac{1}{c} \sum_{t\geqslant 1} \frac{\sum_{s=1}^{t} s m_s}{\sum_{s=1}^{t} m_s} \left(1 - \prod_{s=1}^{t} e^{-m_s}\right) p_t\,,$$

where $p_t = \mathbb{P}(T_i = t) = \ell_t - \ell_{t+1}$, and

$$c = \sum_{t \geqslant 1} \left(1 - \prod_{s=1}^{t} e^{-m_s} \right) p_t \, .$$

6.C Proof of $\mathbb{E}(\tilde{T}) \to \mathbb{E}(T^2)/\mathbb{E}(T)$ as $m \to 0$

In this section we prove that, when the lifespan T of a fixed individual has a second moment – a condition that is always met in practice – and the age-specific fertility is constant and equal to m, then the expected lifespan of individuals that produce offspring during their lifetime converges to $\mathbb{E}(T^2)/\mathbb{E}(T)$ as $m \to 0$. As seen in the main text, it follows immediately that $\tau \to \mu_1$, both in the continuous setting where offspring production occurs at a constant rate m during the lifetime of individuals and in the discrete setting where individuals produce a Poisson(m) number of offspring at each integer-valued age $t \geqslant 1$.

Proposition 6.C.1. *Let T denote the lifespan of a fixed individual, and let $\tilde{T}$ have the distribution of T conditional on reproduction in the model where reproduction happens at constant rate (or in the model where individuals produce* Poisson(m) *offspring at each integer age $t \geqslant 1$ at which they are alive), i.e.*

$$\mathbb{E}\left(\tilde{T}\right) \;=\; \frac{\mathbb{E}\left(T(1 - e^{-mT})\right)}{\mathbb{E}(1 - e^{-mT})} \, .$$

Then, if $\mathbb{E}(T^2) < +\infty$,

$$\mathbb{E}\left(\tilde{T}\right) \to \frac{\mathbb{E}(T^2)}{\mathbb{E}(T)} \quad \text{as } m \to 0 \, .$$

Proof. The following proof is due to Stephen P. Ellner and is a welcome simplification of my original proof.

Let $g(m,t) = (1 - e^{-mt})/m$, so that

$$\mathbb{E}\left(\tilde{T}\right) \;=\; \frac{\mathbb{E}(Tg(m,T))}{\mathbb{E}(g(m,T))} \, .$$

Since

$$\frac{\partial}{\partial m} g(m,t) \;=\; \left(1 + mt - e^{mt}\right) \frac{e^{-mt}}{m^2}$$

and that $1 + x \leqslant e^x$ for all x, we see that $g(m,t)$ increases to t as m decreases to 0. By the monotone convergence theorem, it follows that $\mathbb{E}(g(m,T)) \uparrow \mathbb{E}(T)$ and $\mathbb{E}(Tg(m,T)) \uparrow \mathbb{E}(T^2)$ as $m \downarrow 0$. This terminates the proof. □

6.D Proof of $\tau \leqslant \mu_1$

In this section we prove that μ_1, as defined by formulas (1) and (2), is always greater than or equal to τ, as defined by formulas (3) and (5). This will be a simple consequence of the following lemma.

Lemma 6.D.1. *Let X be a positive random variable, and let g and h be positive functions such that $x \mapsto g(x)/x$ is nondecreasing and $x \mapsto h(x)/x$ is nonincreasing. Then,*

$$\mathbb{E}\left(\frac{g(X)h(X)}{X}\right)\mathbb{E}(X) \;\leqslant\; \mathbb{E}(g(X))\,\mathbb{E}(h(X))$$

Proof. Let Y be a random variable with the same distribution as X and that is independent of X. We have to show

$$\begin{aligned} & \mathbb{E}(g(X))\,\mathbb{E}(h(Y)) - \mathbb{E}\Big(\frac{g(Y)h(Y)}{Y}\Big)\mathbb{E}(X) \;\geqslant\; 0 \\ \iff \quad & \mathbb{E}\Big(Xh(Y)\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\Big) \;\geqslant\; 0 \\ \iff \quad & \mathbb{E}\Big(Xh(Y)\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\mathbb{1}_{\{X>Y\}}\Big) \\ & \quad + \;\; \mathbb{E}\Big(Xh(Y)\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\mathbb{1}_{\{X<Y\}}\Big) \;\geqslant\; 0 \end{aligned} \tag{6.12}$$

Since $x \mapsto g(x)/x$ is nondecreasing,

$$\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\mathbb{1}_{\{X<Y\}} \;\leqslant\; 0\,,$$

and since $x \mapsto h(x)/x$ is nonincreasing,

$$0 \;\leqslant\; Xh(Y)\mathbb{1}_{\{X<Y\}} \;\leqslant\; Yh(X)\mathbb{1}_{\{X<Y\}}\,.$$

As a result,

$$\begin{aligned} & \mathbb{E}\Big(Xh(Y)\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\mathbb{1}_{\{X<Y\}}\Big) \\ \geqslant\; & \mathbb{E}\Big(Yh(X)\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\mathbb{1}_{\{X<Y\}}\Big) \\ =\; & -\mathbb{E}\Big(Xh(Y)\Big(\frac{g(X)}{X} - \frac{g(Y)}{Y}\Big)\mathbb{1}_{\{X>Y\}}\Big)\,. \end{aligned}$$

Plugging this into (6.12) finishes the proof. □

Proposition 6.D.2. *Let T denote the lifespan of a fixed individual. Define the random variables M and M^* by*

$$M = \int_0^T m(s)\,ds \quad \textit{and} \quad M^* = \int_0^T s\,m(s)\,ds$$

in the case where reproduction occurs at a constant rate, and by

$$M = \sum_{s=1}^{T} m_s \quad \textit{and} \quad M^* = \sum_{s=1}^{T} s\,m_s$$

in the case where it takes place at integer-valued ages $t \geqslant 1$, so that, in both cases

$$\mu_1 \;=\; \frac{\mathbb{E}(M^*)}{\mathbb{E}(M)} \quad \textit{and} \quad \tau \;=\; \frac{\mathbb{E}\big(\frac{M^*}{M}(1-e^{-M})\big)}{\mathbb{E}(1-e^{-M})}\,.$$

Then, $\tau \leqslant \mu_1$.

Proof. First, observe that M^* is actually a deterministic function of M. Indeed, let ψ (resp. ψ^*) denote the function such that $M = \psi(T)$ (resp. $M^* = \psi^*(T)$). Since ψ is nondecreasing, if we define θ by

$$\theta(x) = \inf\{t \geqslant 0 : \psi(t) \geqslant x\}\,,$$

then we have $M^* = \psi^*(\theta(M))$. To see this, note that $\theta(M) \leqslant T$ by construction and that $\theta(M) < T$ implies $\int_{\theta(M)}^{T} m(s)\,ds = 0$ (resp. $\sum_{s=\theta(M)}^{T} m_s = 0$), which in turn implies $\int_{\theta(M)}^{T} s\,m(s)\,ds = 0$ (resp. $\sum_{s=\theta(M)}^{T} s\,m_s = 0$). Thus, writing

$$g(x) = \psi^*(\theta(x)) \quad \text{and} \quad h(x) = 1 - e^{-x}\,,$$

we have to prove

$$\mathbb{E}(g(M))\,\mathbb{E}(h(M)) \;\geqslant\; \mathbb{E}\Big(\frac{g(M)h(M)}{M}\Big)\,\mathbb{E}(M)\,.$$

Clearly, M is a positive random variable, and the functions g and h are positive. Therefore, all we have to do to finish the proof is to show that $x \mapsto h(x)/x$ is nonincreasing and that $x \mapsto g(x)/x$ is nondecreasing, so that we can apply Lemma 6.D.1. First,

$$\frac{d}{dx}\Big(\frac{h(x)}{x}\Big) \;=\; \frac{e^{-x}(1+x)-1}{x^2} \;\leqslant\; 0\,,$$

since $1 + x \leqslant e^x$. Second,

$$\frac{g(x)}{x} \;=\; \frac{\psi^*(\theta(x))}{\psi(\theta(x))} \;=\; F(\theta(x))$$

where $F\colon t \mapsto \psi^*(t)/\psi(t)$. The function θ is nondecreasing by construction. The fact that F is nondecreasing can be shown by straightforward calculations, e.g., in the continuous case,

$$\frac{d}{dt}F(t) \;=\; \frac{m(t)\Big(\int_0^t (t-s)m(s)\,ds\Big)}{\Big(\int_0^t m(s)\,ds\Big)^2} \;\geqslant\; 0\,.$$

However, it is more satisfying to see that $F(t)$ can be interpreted as the expectation of a random variable X_t with density $f_t(s) = m(s)\mathbb{1}_{[0,t]}(s)/\psi(t)$ in the continuous case, and with probability mass function $p_s^{(t)} = m_s/\psi(t)$ for $s = 1, \ldots, t$ in the discrete case. It then is easy to see that X_t is stochastically dominated by $X_{t'}$ for $t < t'$, and so if follows immediately that $F(t) = \mathbb{E}(X_t) \leqslant \mathbb{E}(X_{t'}) = F(t')$. □

6.E Computing μ_1 and τ for Compadre/Comadre

In this section, we detail how the data behind Figure 6.2 and Table 6.1 in the main text were obtained.

The Compadre and Comadre each contain thousands of projection matrices for hundreds of species. However, not all of these matrices are suitable to compute μ_1 and τ. Indeed, for this we need:

(i) The $\mathbf{A} = \mathbf{S} + \mathbf{F}$ decomposition of the projection matrix into its survival and fertility components.

(ii) Non-zero $\mathbf{S}$ and $\mathbf{F}$ matrices.

(iii) A survival matrix $\mathbf{S}$ whose columns all sum to less than one, so that it can be interpreted as a substochastic matrix and that $(\mathbf{I} - \mathbf{S})^{-1}$ is always guaranteed to exist.

This leaves us with 3319 models in Compadre and 1245 models in Comadre. For each of these, μ_1 was computed with formula (11) and τ was estimated by averaging several realizations of the random variable S/N described in the main-text (conditional on $N > 0$). One such realization can be obtained thanks to the following procedure, where $\mathbf{w} = (w_i)$ denotes the stable distribution:

---

parent ← False
while not *parent* **do**
  $i \leftarrow$ random newborn stage chosen proportionally to the entries of $(\sum_k f_{jk} w_k)_j$
  age, N, S ← 0, 0, 0
  alive ← True
  while *alive* **do**
    age ← *age* + 1
    offspring ← Poisson($\sum_j f_{ji}$)
    $N \leftarrow N$ + *offspring*
    $S \leftarrow S$ + *offspring* · *age*
    with probability $1 - \sum_j s_{ji}$ **do**
      alive ← False
    else do
      $i \leftarrow$ random stage chosen proportionally the entries of $(s_{ji})_j$
  end while
  if $N > 0$ **then**
    parent ← True
end while
return S/N

Each estimate $\hat{\tau}$ is associated with a confidence interval $[\hat{\tau} - \varepsilon, \hat{\tau} + \varepsilon]$, where $\varepsilon = 2\hat{\sigma}/\sqrt{n}$ with $\hat{\sigma}$ the empirical standard deviation and $\sqrt{n}$ the number of replicates. In order to get reliable estimates, the number of replicates n was doubled until $\varepsilon < 0.01\hat{\tau}$.

Because the probability of an individual producing some offspring during its lifetime can be arbitrarily small, this means that obtaining one realization of S/N with the procedure above can take an arbitrarily long time. To avoid getting stuck on a computation, all models for which $\hat{\tau}$ could not be computed with the desired precision in a reasonable time were ignored, and 1182 models were thus rejected. Because this is a non-negligible fraction of the 4564 models available, this has the

potential to bias our results. However, note that since those models for which the probability of producing offspring during one's lifetime is very small are precisely those for which we expect μ_1 to differ greatly from τ, if anything this will lead us to *underestimate* the difference between μ_1 and τ.

Finally, after performing these calculations, 11 models (0.28%) were discarded for having biologically unrealistic descriptors (e.g, $\lambda \approx 200$ or an average age of mothers in the stable population $\bar{A} \approx 10000$), leaving us with numerical values of μ_1 and τ for 2706 models of COMPADRE and 1165 models of COMADRE.

6.F Projection matrices for *A. mexicanum*

In this section, we detail a specific example of a real-world model in which μ_1 and τ differ greatly. These particular models were chosen because they have frequently been used as examples of matrix population models. For instance, one of them is shipped with the ULM software for studying population dynamics [19].

The following projection matrices for the tropical palm *Astrocaryum mexicanum* are from Appendix 6 of [5], who averaged them from several projection matrices of [20]. Note that there is a small typo in the projection matrix for the *Low density* model given by [5]: the entry $(9, 8)$ of the projection matrix given is 0.8775, when it should be 0.08775. Correcting this, we find the same descriptors as in their Table 4.

The model is mostly size-based. Stage 1 corresponds to seedlings; stages 2–4 to non-reproducing juveniles and stages 5–10 to full-grown adults. In the matrices below, entries in bold correspond to reproductive transitions.

$$\mathbf{A}^{\text{high}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \mathbf{1.4792} & \mathbf{8.1560} & \mathbf{9.9513} & \mathbf{14.259} & \mathbf{23.594} \\ 0.037349 & 0.83093 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.015881 & 0.89666 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.048969 & 0.95944 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.029778 & 0.90496 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.082074 & 0.91348 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.086520 & 0.90553 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.094467 & 0.87733 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.088200 & 0.88642 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.11358 & 0.9950 \end{pmatrix}$$

$$\mathbf{A}^{\text{med.}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \mathbf{0.18385} & \mathbf{4.222} & \mathbf{8.41} & \mathbf{8.8405} & \mathbf{16.676} & \mathbf{19.904} \\ 0.03629 & 0.84127 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.014582 & 0.91636 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.058131 & 0.93735 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.051565 & 0.91462 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.065923 & 0.8468 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1424 & 0.8725 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1200 & 0.84332 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.14800 & 0.913030 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.086966 & 0.9950 \end{pmatrix}$$

$$\mathbf{A}^{\text{low}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \mathbf{0.33} & \mathbf{0.918} & \mathbf{8.0875} & \mathbf{16.606} & \mathbf{13.068} & \mathbf{16.875} \\ 0.030332 & 0.850010 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.026738 & 0.93928 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.04966 & 0.94548 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.04804 & 0.9185 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0815 & 0.9313 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0687 & 0.86362 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.13637 & 0.91225 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.08775 & 0.87867 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12133 & 0.9950 \end{pmatrix}$$

Table 6.2 lists some relevant descriptors of these models. The difference between μ_1 and τ is significant in all three scenarios – a factor 2 in the *Medium density* model. Finally, and most surprisingly, we see that in the *High density* case, μ_1 is greater than the expected lifespan conditional on reproduction. This counter-intuitive fact casts serious doubts on the relevance of μ_1 as a measure of reproductive timing in the life-cycle.

	High density	Medium density	Low density
μ_1	275.2	261.8	275.5
τ	152.0	131.3	184.8
T_{R_0}	197.6	169.2	153.6
$\bar{A}$	152.6	122.8	105.1
L	232.2	207.6	296.0

Table 6.2: Comparison of several measures of reproductive timing for three real-world models for the demography of the tropical palm *Astrocaryum mexicanum*, in part taken from Table 4 of [5]: T_{R_0} denotes the R_0 generation time, which corresponds to the time it takes for the population to grow by a factor of its net reproductive rate; $\bar{A}$ is the mean age of parents of offspring in a population that has reached its stable distribution; μ_1 is computed as in formula (11); τ and L are estimates of the mean age at reproduction and of the expected lifespan conditional on producing offspring, respectively. All values are expressed in years.

Bibliography

— The On-Line Encyclopedia of Integer Sequences, published electronically at http://oeis.org, 2019.

— Ahlswede, R. and Daykin, D. E. (1978). An inequality for the weights of two families of sets, their unions and intersections. *Probability Theory and Related Fields*, 43(3):183–185.

— Aigner, M. and Ziegler, G. M. (2018). *Proofs from THE BOOK*. Springer-Verlag Berlin, 6th edition.

— Athreya, K. B. and Lahiri, S. N. (2006). *Measure theory and probability theory*. Springer Science+Business Media.

— Balińska, K. T., Quintas, L. V., and Szymański, J. (1994). Random recursive forests. *Random Structures & Algorithms*, 5(1):3–12.

— Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

— Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson approximation*. Oxford Studies in Probability. Clarendon Press.

— Bergeron, F., Flajolet, P., and Salvy, B. (1992). Varieties of increasing trees. In *CAAP '92*, pages 24–48. Springer Berlin Heidelberg.

— Bienvenu, F. (2019). The equivocal mean age of parents in a cohort. *The American Naturalist*, 194(2).

— Bienvenu, F., Débarre, F., and Lambert, A. (2019). The split-and-drift random graph, a null model for speciation. *Stochastic Processes and their Applications*, 129(6):2010–2048.

— Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, 2nd edition.

— Bollobás, B. (2001). *Random graphs*. Cambridge University Press.

— Broadbent, S. R. and Hammersley, J. M. (1957). Percolation processes: I. crystals and mazes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(3):629–641.

— Cardona, G., Pons, J. C., and Céline, S. (2019). Generation of tree-child phylogenetic networks. *arXiv preprint arXiv:1902.09015*.

— Cardona, G., Rossello, F., and Valiente, G. (2009). Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):552–569.

— Caswell, H. (2001). *Matrix Population Models: Construction, Analysis, and Interpretation*. Sinauer Associates, 2nd edition.

— Charlesworth, B. (1994). *Evolution in Age-Structured Populations*. Cambridge Studies in Mathematical Biology. Cambridge University Press, 2nd edition.

— Chauvin, B., Rouault, A., and Wakolbinger, A. (1991). Growing conditioned trees. *Stochastic Processes and their Applications*, 39(1):117–130.

— Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Annals of Probability*, 3(3):534–545.

— Chung, F., Lu, L., Dewey, T. G., and Galas, D. J. (2003). Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687.

— Coale, A. J. (1972). *The Growth and Structure of Human Populations*. Princeton University Press.

— Cochran, M. E. and Ellner, S. P. (1992). Simple methods for calculating age-based life history parameters for stage-structured populations. *Ecological Monographs*, 62(3):345–364.

— Comadre Animal Matrix Database (Data downloaded on 02/02/2019, version 2.0.1). Max Planck Institute for Demographic Research (Germany). Available at www.compadre-db.org.

— Compadre Plant Matrix Database (Data downloaded on 02/02/2019, version 4.0.1). Max Planck Institute for Demographic Research (Germany). Available at www.compadre-db.org.

— Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates.

— Crump, K. S. and Mode, C. J. (1968). A general age-dependent branching process I. *Journal of Mathematical Analysis and Applications*, 24(3):494–508.

— Crump, K. S. and Mode, C. J. (1969). A general age-dependent branching process II. *Journal of Mathematical Analysis and Applications*, 25(1):8–17.

— Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*. Probability and its Applications. Springer-Verlag New York, 2nd edition.

— Drake, B. (2008). *An inversion theorem for labeled trees and some limits of areas under lattice paths*. PhD thesis, Brandeis University.

— Drmota, M. (2009). *Random trees: an interplay between combinatorics and probability*. Springer-Verlag Vienna.

— Durrett, R. (1984). Oriented percolation in two dimensions. *The Annals of Probability*, 12(4):999–1040.

— Durrett, R. (1996). *Stochastic calculus: a practical introduction*. CRC Press.

— Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer-Verlag New York, 2nd edition.

— Durrett, R. (2010). *Probability: theory and examples*. Cambridge University Press, 4th edition.

— Dyer, R. J. and Nason, J. D. (2004). Population graphs: the graph theoretic shape of genetic structure. *Molecular ecology*, 13(7):1713–1727.

— Eğecioğlu, Ö. and Remmel, J. B. (1986). Bijections for Cayley trees, spanning trees, and their q-analogues. *Journal of Combinatorial Theory, Series A*, 42(1):15–30.

— Ellner, S. P. (2018). Generation time in structured populations. *The American Naturalist,*, 192(1):105–110.

— Ellner, S. P., Childs, D. Z., and Rees, M. (2016). *Data-driven modelling of structured populations*. Springer International Publishing.

— Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–61.

— Erdős, P. (1947). Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, 53(4):292–294.

— Erdős, P. (1959). Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38.

— Erdős, P. (1960). Graph theory and probability II. *Canadian Journal of Mathematics*, 13:346–352.

— Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297.

— Erdős, P. and Rényi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1-2):261–267.

— Esary, J. D., Proschan, F., and Walkup, D. W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5):1466–1474.

— Etheridge, A. (2011). *Some mathematical models from population genetics. École d'été de probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer-Verlag Berlin Heidelberg.

— Ethier, S. N. and Kurtz, T. G. (2005). *Markov processes: characterization and convergence*. John Wiley & Sons, 2nd edition.

— Fill, J. A. and Pemantle, R. (1993a). Percolation, first-passage percolation and covering times for Richardson's model on the n-cube. *The Annals of Applied Probability*, 3(2):593–629.

— Fill, J. A. and Pemantle, R. (1993b). Percolation, first-passage percolation and covering times for richardson's model on the n-cube. *The Annals of Applied Probability*, 3(2):593–629.

— Fortuin, C. M., Kasteleyn, P. W., and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103.

— Fuchs, M., Gittenberger, B., and Mansouri, M. (2018). Counting phylogenetic networks with few reticulation vertices: Tree-child and normal networks.

— Gessel, I. M. and Seo, S. (2006). A refinement of Cayley's formula for trees. *The electronic journal of combinatorics*, 11(2):R27.

— Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

— González-Forero, M. (2009). Removing ambiguity from the biological species concept. *Journal of theoretical biology*, 256(1):76–80.

— Grimmett, G. R. (2001). Infinite paths in randomly oriented lattices. *Random Structures & Algorithms*, 18(3):257–266.

— Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

— Huson, D. H. and Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267.

— Irwin, D. E., Irwin, J. H., and Price, T. D. (2001). Ring species as bridges between microevolution and speciation. In *Microevolution rate, pattern, process*, pages 223–243. Springer Netherlands.

— Ispolatov, I., Krapivsky, P., and Yuryev, A. (2005). Duplication-divergence model of protein interaction network. *Physical Review E*, 71(6):061911.

— Jagers, P. (1969). A general stochastic model for population development. *Scandinavian Actuarial Journal*, 1969(1-2):84–103.

— Jørgensen, S. E. and Fath, B. (2008). *Encyclopedia of Ecology*. Elsevier Science, 1st edition.

— Joyal, A. (1981). Une théorie combinatoire des séries formelles. *Advances in mathematics*, 42(1):1–82.

— Keyfitz, N. (1968). *Introduction to the Mathematics of Population*. Addison-Wesley.

— Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

— Kingman, J. F. C. (1992). *Poisson Processes*. Oxford Studies in Probability. Clarendon Press.

— Kliman, R. M. (2016). *Encyclopedia of Evolutionary Biology*. Academic Press.

— Knuth, D. E. (1997). *The art of computer programming: sorting and searching*, volume 3. Addison-Wesley.

— Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2006). *Power laws, scale-free networks and genome biology*. Springer US.

— Lambert, A. (2017). Probabilistic models for the (sub)tree(s) of life. *Brazilian Journal of Probability and Statistics*, 31(3):415–475.

— Legendre, S. and Clobert, J. (1995). ULM, a software for conservation and evolutionary biologists. *Journal of Applied Statistics*, 22(5-6):817–834. The ULM software can be downloaded at http://www.biologie.ens.fr/~legendre/ulm/ulm.html.

— Linusson, S. (2009). A note on correlations in randomly oriented graphs. *arXiv preprint arXiv:0905.2881*.

— Lyons, R., Pemantle, R., and Peres, Y. (1995). Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Annals of Probability*, 23(3):1125–1138.

— Mahmoud, H. M. and Smythe, R. T. (1991). On the distribution of leaves in rooted subtrees of recursive trees. *The Annals of Applied Probability*, 1(3):406–418.

— Manceau, M. and Lambert, A. (2019). The species problem from the modeler's point of view. *Bulletin of Mathematical Biology*, 81:878–898.

— Martinsson, A. (2016). Unoriented first-passage percolation on the n-cube. *The Annals of Applied Probability*, 26(5):2597–2625.

— May, R. M. (2006). Network structure and the biology of populations. *Trends in Ecology & Evolution*, 21(7):394–399.

— Mayr, E. (1942). *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. Columbia University Press.

— McDiarmid, C., Semple, C., and Welsh, D. (2015). Counting phylogenetic networks. *Annals of Combinatorics*, 19(1):205–224.

— Meir, A. and Moon, J. W. (1974). Cutting down recursive trees. *Mathematical Biosciences*, 21(3):173–181.

— Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71.

— Narayanan, B. (2016). Connections in randomly oriented graphs. *Combinatorics, Probability and Computing*, pages 1–5.

— Newman, M., Barabási, A.-L., and Watts, D. J. (2011). *The structure and dynamics of networks*, volume 12. Princeton University Press.

— Nowak, M. A. (2006). *Evolutionary dynamics*. Harvard University Press.

— Pascual, M., Dunne, J. A., et al. (2006). *Ecological networks: linking structure to dynamics in food webs*. Oxford University Press.

— Pinero, D., Martinez-Ramos, M., and Sarukhan, J. (1984). A population model of astrocaryum mexicanum and a sensitivity analysis of its finite rate of increase. *The Journal of Ecology*, 72(3):977–991.

— Pitman, J. (2006). *Combinatorial Stochastic Processes: École d'été de probabilités de Saint-Flour XXXII-2002*. Springer-Verlag Berlin Heidelberg.

— Poulton, E. B. (1904). What is a species? In *Proceedings of the Entomological Society of London 1903*, pages lxxvii–cxvi.

— Proulx, S. R., Promislow, D. E. L., and Phillips, P. C. (2005). Network thinking in ecology and evolution. *Trends in ecology & evolution*, 20(6):345–353.

— Ratcliff, W. C., Fankhauser, J. D., Rogers, D. W., Greig, D., and Travisano, M. (2015). Origins of multicellular evolvability in snowflake yeast. *Nature Communications*, 6:6102.

— Rockwood, L. L. (2015). *Introduction to Population Ecology*. Wiley–Blackwell, 2nd edition.

— Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys*, 8:201–293.

— Ross, S. M. (1995). *Stochastic Processes*. Wiley, 2nd edition.

— Solé, R., Pastor-Satorras, R., Smith, E., and Kepler, T. B. (2002). A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(1):43–54.

— Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117.

— Steel, M. (2016). *Phylogeny: discrete and random processes in evolution*. SIAM.

— Stein, C. M. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 583–602. University of California Press.

— Stein, W. A. et al. (2019). *Sage Mathematics Software (Version 8.2)*. The Sage Development Team. http://www.sagemath.org.

— Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164.

— Van Der Hofstad, R. (2016). *Random graphs and complex networks*. Cambridge University Press.

— Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Modeling of protein interaction networks. *ComPlexUs*, 1:38–44.

— Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

— Willson, S. J. (2007a). Reconstruction of some hybrid phylogenetic networks with homoplasies from distances. *Bulletin of mathematical biology*, 69(8):2561–2590.

— Willson, S. J. (2007b). Unique determination of some homoplasies at hybridization events. *Bulletin of mathematical biology*, 69(5):1709–1725.