Gene genealogies in a diploid Moran model with selfing conditional on its pedigree

Maximillian Newman[1], John Wakeley[2], and Wai-Tong (Louis) Fan[1,2]

[1]*Department of Mathematics, Indiana University, Bloomington, IN*
[2]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA*

September 13, 2024

Abstract

We introduce a stochastic model of a population with overlapping generations and arbitrary levels of self-fertilization versus outcrossing. We study how the global graph of reproductive relationships, or pedigree, influences the genealogical relationships of a sample of two gene copies at a genetic locus. Specifically, we consider a diploid Moran model with constant population size N over time in which a proportion of offspring are produced by selfing. We show that the conditional distribution of the pairwise coalescent time at the locus given the random pedigree converges to a limit law as N tends to infinity. This limit law generally differs from the corresponding traditional result obtained by averaging over all possible pedigrees. We describe three different behaviors in the limit depending on the relative strengths of selfing versus outcrossing: partial selfing, limited outcrossing, and negligible outcrossing. In the case of partial selfing, coalescence times are related to the Kingman coalescent, similar to what is found by the traditional result. In the case of limited outcrossing, the retained pedigree information forms a random graph coming from a fragmentation-coagulation process, with the limiting coalescence times given by the meeting times of random walks on this graph. In the case negligible outcrossing, which represents nearly complete selfing, coalescence times are determined entirely by the fixed times to common ancestry of individuals in the pedigree.

1 Introduction

Modes of reproduction and genetic transmission are fundamental aspects of evolution (Charlesworth, 2006; Olsen et al., 2021). In this work we consider autogamy or self-fertilization—hereafter just 'selfing'—in which meiosis occurs then a zygote is formed by the union of two gametes from the same individual. The usual alternative to selfing in plants and animals is outcrossing, in which zygotes are formed by the union of two gametes from different individuals. The selfing rate, defined here as the probability that an individual is produced by selfing, ranges from 0 to 1. We describe three distinct kinds of coalescent models which hold for different magnitudes of the selfing rate when the process of coalescence is conditioned on the population pedigree, i.e. the organismal genealogy of the population (Ball et al., 1990) which has mostly been largely ignored in theoretical population genetics (see Diamantidis et al. (2024) and Section 1.3).

Email: `newmama@iu.edu` (Maximillian Newman), `wakeley@fas.harvard.edu` (John Wakeley), `waifan@iu.edu` (Wai-Tong (Louis) Fan)

The conditional coalescent models we describe exist in the limit as the population size N tends to infinity, in the same way that the Wright-Fisher diffusion and the corresponding standard neutral coalescent model exist in this limit (Kingman, 1982; Möhle, 1999; Ewens, 2004). We use α_N to denote the selfing rate, with explicit dependence on N, in place of the more usual notation s. Previous analyses, which have not conditioned on the pedigree, have assumed that N is large as in the diffusion limit and $s \in [0, 1]$ is constant, and have referred to this as 'partial self-fertilization' or 'partial selfing'. Our results conditional on the pedigree depend on how α_N changes as $N \to \infty$. Other notations for the selfing rate exist in the literature: Haldane (1924) used l and Pollak (1987) used β. We use $\alpha \in [0, 1]$ to denote the limit of α_N as $N \to \infty$. Disregarding the pedigree, our model coincides with previous ones when $\alpha_N = \alpha$ is a fixed constant. Then α corresponds to s (or l or β).

Conditioning coalescence on the pedigree reveals three markedly different behaviors near the boundary $\alpha = 1$, depending on how quickly $\alpha_N \to 1$. The critical case, which we call 'limited outcrossing', is when $1-\alpha_N$ is of order $1/N$. Then the conditional coalescent process is characterized by a system of coalescing random walks on a random ancestral graph like the ones previously described for recombination (Griffiths, 1991; Griffiths and Marjoram, 1997) and selection (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). If outcrossing is even less frequent, for example $1-\alpha_N$ of order $1/N^2$, the limiting conditional coalescent process is fully determined by the random pedigree of the population, which becomes just a Kingman coalescent tree. We call this 'negligible outcrossing'. If outcrossing is more frequent than in the critical case, for example $1 - \alpha_N$ of order $1/\sqrt{N}$, the limiting conditional coalescent model resembles the previously described (unconditional) coalescent model Nordborg and Donnelly (1997) and Möhle (1998). Here we adopt the previous term 'partial selfing'. This model holds for all $\alpha \in [0, 1]$ as long as the rate of approach to $\alpha = 1$ is not too fast, whereas limited outcrossing and negligible outcrossing hold just for $\alpha = 1$.

Because we will be reviewing previous works in the next two sections, which overwhelmingly use s for the selfing rate, take $s \in [0, 1]$ and do not condition on the population pedigree, we will reserve our notation α_N for Section 1.3 and later. In Section 1.3, we explain how previous population-genetic models in fact average over pedigrees. In Section 1.1 and Section 1.2 we use s for continuity with what one finds in most of the evolutionary and population-genetic literature.

1.1 Prevalence, causes and consequences of selfing in evolution

Although the ability to self-fertilize is rare among animals (Jarne and Auld, 2006; Avise and Mank, 2009; Escobar et al., 2011) it is found in other taxa including eukaryotic microbes and marine invertebrates (Sasson and Ryan, 2017; Yadav et al., 2023) and is common in plants (Abbott and Gomes, 1989; Hartfield et al., 2017; Teterina et al., 2023). About 10-15% of plant species are predominantly selfing (Wright et al., 2013) as are two of the best studied genetic model species, *Arabidopsis thaliana* and *Caenorhabditis elegans* (Abbott and Gomes, 1989; Sellinger et al., 2020; Barrière and Félix, 2005).

The empirical distribution of selfing rates among plant species is markedly bimodal, with relatively fewer species having $s \in (0.2, 0.8)$ (Schemske and Lande, 1985; Vogler and Kalisz, 2001; Goodwillie et al., 2005). Genetical models of inbreeding depression can explain this bimodality (Lande and Schemske, 1985; Charlesworth and Willis, 2009). But, whether selfing is favored over outcrossing or vice versa in a given species depends on a number of ecological and evolutionary factors (Charlesworth, 2006; Wright et al., 2013).

Selfing has two main population-genetic effects: increased homozygosity via inbreeding and,

consequently, reduced recombination. Haldane (1924) described the increase in homozygosity and decrease in heterozygosity, compared to Hardy-Weinberg expectations, in an infinite population with selfing rate $s \in [0, 1]$. Putting his result in terms of the individual inbreeding coefficient F (Wright, 1931, 1951), what Haldane (1924) showed was that the reduction in the proportion of heterozygous individuals in the population at equilibrium is given by $F = s/(2 - s)$.

Bennett and Binet (1956) considered genotype frequencies at two linked loci and showed that this reduction in heterozygosity due to selfing causes a reduction in the effective rate of recombination. Weir and Cockerham (1973) confirmed this effect in their study of two-locus identity coefficients. Recombination is similarly reduced in finite populations (Golding and Strobeck, 1980; Vitalis and Couvet, 2001). Nordborg (2000) provided an interpretation in terms of coalescence. Backward in time, a recombination event places ancestral genetic material onto two chromosomes. These are necessarily in the same parental individual which might itself have been produced by selfing. If so, the recombination event can get healed, that is undone by a coalescent event in the grandparent. The same coefficient $F = s/(2 - s)$ is the average probability of such healing. The result is that only a fraction $1 - F = 2(1 - s)/(2 - s)$ of recombination events are observable. While constant s is certainly the best-studied case, we note that other assumptions about how selfing and recombination scale with N give different results (Kogan et al., 2023).

In the absence of other evolutionary forces, selfing is strongly favored over outcrossing (Fisher, 1941; Moran, 1962; Nagylaki, 1976; Lloyd, 1979; Wells, 1979). As Maynard Smith (1971) observed in discussing the evolution of sexual reproduction or meiosis, outcrossing individuals are at a disadvantage because they transmit only one half of their genome to each offspring. The comparison in Maynard Smith (1971) was to parthenogenesis or asexual reproduction, but selfing is a way to alleviate this "cost of sex" in organisms which do undergo meiosis. The fact that outcrossing is common can be explained in genetical terms by the avoidance of inbreeding depression (Charlesworth and Willis, 2009; Brown and Kelly, 2020) or the purging or deleterious mutations in the presence of recombination (Kondrashov, 1985; Kamran-Disfani and Agrawal, 2014).

In addition, many ecological factors affect the evolutionary dynamics of selfing species. Pioneering work by Baker (1955) and Stebbins (1957) showed that selfing may be favored when the usual modes of cross-fertilization are blocked, in plants for example by excessive rainfall or lack of pollinators, or when the environment presents opportunities for colonization starting from one or a few individuals, for example if rare long-distance dispersal events can access open habitat or if there is rapid turnover of suitable habitat patches. Ingvarsson (2002) showed that the last of these, specifically extinction-recolonization dynamics in a metapopulation, could explain the common observation that levels of genetic variation in selfing species are often far below what is predicted by simple neutral population-genetic models (see Section 1.2). Selfing species with their characteristic suite of phenotypes have evolved many times from outcrossing ancestors, prompting questions about whether it might be an evolutionary "dead end" (Stebbins, 1957; Ornduff, 1969; Sicard and Lenhard, 2011; Wright et al., 2013; Cutter, 2019). Within a single species, subpopulations of selfers are commonly restricted to patches at the margins of the habitat (Baker, 1955; Stebbins, 1957), as for example in the case of *Leavenworthia alabamica* detailed in Busch (2005).

1.2 Population genetics of selfing under neutrality

Our goal in this work is to understand how selfing affects the sampling structure of gene genealogies under neutrality when the process of coalescence is conditioned on the pedigree of the population. We ask whether results conditional on the pedigree differ from those of previous theoretical treat-

ments which have disregarded this feature of the population. For simplicity we use a diploid Moran model of reproduction (Moran, 1958, 1962; Linder, 2009; Coron and Le Jan, 2022), which we will describe in Section 2. Previous theoretical treatments, including those we review here, have assumed Wright-Fisher reproduction (Fisher, 1930; Wright, 1931).

Based on the considerations of Section 1.1, it may be that selfing species are even less likely than outcrossing ones to meet the assumptions of standard models of population genetics. Both the Wright-Fisher model and the Moran model apply to a single well mixed population of constant in size N without selection. The simplest application relevant to selfing is to diploid, monoecious, randomly mating organisms, in which case selfing happens with probability $1/N$. When N can be assumed to be large, both models converge to the Wright-Fisher diffusion with its corresponding coalescent process (Kingman, 1982; Hudson, 1983a; Tajima, 1983; Möhle, 1999; Ewens, 2004) which exist in the limit $N \to \infty$ with time measured in units proportional to N generations. This provides a common framework for studying more complicated populations. Work on the population genetics of selfing has proceeded in this way, for the most part under the assumption that the selfing rate s is a fixed constant when the population size N tends to infinity in the model.

Pollak (1987) took a forward-time approach to studying, among other things, the equilibrium probabilities of identity by descent of two gene copies in the same individual or in different individuals, when $s \gg N^{-1}$. Let us call these probabilities Φ_1 and Φ_2, respectively. For large N, Pollak (1987) found that the expression for Φ_2 was identical to that for a randomly mating population, if $N_e = N/(1+F) = (2-s)N/2$ is used in place of N. As s increases from 0 to 1, N_e decreases from N to $N/2$, effectively taking the value of Φ_2 from that for a diploid (randomly mating) population to that for a haploid population. Pollak (1987) identified this as a separation-of-time-scales result, stemming from the assumption $s \gg N^{-1}$, and suggested that the two-times-scales diffusion approximation of Ethier and Nagylaki (1980) could be applied to prove convergence to the Wright-Fisher diffusion with this value of N_e. The separation of time scales is evident in the relationship Pollak (1987) found between Φ_1 and Φ_2, namely that $\Phi_1 = F + (1-F)\Phi_2$, although this may be best understood in terms of subsequent descriptions of the coalescent process.

Nordborg and Donnelly (1997) took a backward-time coalescent approach to the same model, and Nordborg (1997) placed it in a more general framework of separation of time scales. Nordborg and Donnelly (1997) described a coalescent process with two time scales: one involving just selfing which is fast and plays out over relatively small numbers of generations, and one involving coalescence which is slow in the usual sense of coalescent theory where time is measured in units proportional to N generations. The fast process occurs when one or more pairs of lineages are in the same individual(s). They showed (though without a formal proof) that in the limit $N \to \infty$, if a sample of size n contains $2m$ gene copies together as pairs in m individuals and $n-2m$ gene copies each in different individuals, there is an instantaneous adjustment in which a random number $X \sim \text{binomial}(m, F)$ of the m pairs coalesce. The rest of the ancestry begins with the resulting $n - X$ lineages and is given by the standard neutral coalescent process (Kingman, 1982; Hudson, 1983a; Tajima, 1983), if time is measured in units of $2N_e = (2-s)N$ generations.

It was in this context that Möhle (1998) proved an important general result (Möhle, 1998, Lemma 1, Theorem 1) for Markov processes with two time scales. Owing to the duality relation between the coalescent and diffusion models (Möhle, 1999) it can be seen as the backward-time counterpart to the result of Ethier and Nagylaki (1980) cited by Pollak (1987). For the selfing model with fixed s as $N \to \infty$, this provided rigorous justification for the model of Nordborg and Donnelly (1997): whenever a pair of lineages is in the same individual, they coalesce with probability $F = s/(2-s)$ instantaneously, while samples from different individuals obey the standard neutral

coalescent process with time in units of $(2-s)N$ generations (Möhle, 1998).

For clarity, we note that two slightly different definitions of s have been employed in models with finite N. Pollak (1987) and Nordborg and Donnelly (1997) take $1-s$ to be the probability an individual is produced by random mating, so that selfing occurs in this case with probability $1/N$. Möhle (1998) takes $1-s$ to be the probability an individual is produced by outcrossing rather than selfing. This distinction makes no difference in the limit $N \to \infty$, and Möhle (1998) details how his method can be applied to the model in Nordborg and Donnelly (1997) if the selfing rate is taken to be $s + (1-s)/N$.

1.3 Accounting for the pedigree of the population

The pedigree of a population is a graph representing the set of reproductive relationships of all individuals in the population for all time. Previous studies of the population-genetic effects of selfing, including those reviewed in Section 1.2, have followed standard practice in population genetics by averaging over all possible outcomes of reproduction, which is equivalent to averaging over the pedigree. See Diamantidis et al. (2024) for a recent detailed discussion of the issues surrounding this tradition in population genetics, which we review briefly here. In what follows, we will use our notation, in which α_N denotes the selfing rate in a population of size N, and α denotes its value in the limit $N \to \infty$ or in the special case that $\alpha_N = \alpha$ is constant.

Averaging over outcomes of reproduction is often done implicitly, for example by stating that two ancestral lines coalesce with probability $1/(2N)$ each generation under the diploid monoecious Wright-Fisher model. Sometimes it is explicit, for example in equation (4.1) in Kingman (1982) for the coalescence probability under Cannings' exchangeable model (Cannings, 1974). The practice of averaging is questionable at best, however, because all reproductive outcomes in the population for all time are encoded in a fixed structure, the genealogy or pedigree of the population through which all genetic transmission must also have occurred (Ball et al., 1990). In the most common application of coalescent models, which is to describe the distribution of gene genealogies or genetic variation among loci within the same genome, the practice of averaging over reproductive outcomes is plainly wrong because the same pedigree holds for all loci.

A better-justified starting point for coalescent theory is to condition the ancestral genetic process on the pedigree. This gives the correct sampling structure for multiple unlinked loci: conditionally independent realizations of the process of genetic transmission within the same population/pedigree. The sampling structure implicit in traditionally formulated coalescent models is that each locus has its own pedigree, as if each locus was sampled from an independent population.

It turns out that the standard neutral coalescent model, at least, can still be applied because the coalescent model conditional on the pedigree is the same as the standard model for populations in the domain of the Kingman coalescent (Tyukin, 2015; Diamantidis et al., 2024). The reason is that pedigrees in large well mixed populations have a characteristic mixing time of $\log_2 N$ generations (Chang, 1999; Derrida et al., 1999; Barton and Etheridge, 2011; Wakeley et al., 2012). But it is not true in general that extensions of the standard neutral coalescent model are similarly applicable to coalescent processes conditional on pedigrees generated under the corresponding population models (Wakeley et al., 2016; Wilton et al., 2017; Diamantidis et al., 2024).

In the case of selfing, one consequence of conditioning on the pedigree is clear: each sampled individual has its own number of generations of selfing before the first outcrossing event in its ancestry. If this number is k for some individual, the probability that two distinct gene copies sampled from this individual remain distinct (do not coalesce) in any of these generations is 2^{-k}.

Let U be the random number of selfing generations for an individual. Then, for α constant,

$$\mathbb{P}(U = k) = \alpha^k(1-\alpha) \qquad \text{for} \quad k = 0, 1, 2, \ldots \tag{1}$$

and the probability F in previous works can be seen as the average

$$F \coloneqq \sum_{k=0}^{\infty} \left(1 - 2^{-k}\right) \alpha^k(1-\alpha) = \frac{\alpha}{2-\alpha}. \tag{2}$$

When averaging over the pedigree, this F can be applied to every individual, as in the coalescent model of Nordborg and Donnelly (1997) and Möhle (1998). But, as we will show, when coalescence is conditioned on the pedigree, each individual has its own realization of U randomly sampled according to (1). This idea of fixed individual ancestries of selfing has been used in the development of methods of inferring selfing rates from genetic data, e.g. by Enjalbert and David (2000) and Gao et al. (2007), a topic we will return to in the Discussion.

When the coalescent process is conditioned on the pedigree and selfing is the primary mode of reproduction, an entirely new sort of model emerges in the limit, one which is completely invisible in the unconditional pedigree-averaging process. This novel limit process when the outcrossing rate $1 - \alpha_N$ is of order N^{-1}. It is parameterized by $\lambda \coloneqq \lim_{N\to\infty} N(1-\alpha_N)$, which is how mutation rates are assumed to scale in the Wright-Fisher diffusion and the standard neutral coalescent model (Kingman, 1982; Hudson, 1983a; Tajima, 1983; Ewens, 2004) and how migration rates are scaled in the structured coalescent model (Takahata, 1988; Notohara, 1990; Herbots, 1997). In the limit we consider, λ is the instantaneous rate of outcrossing events per ancestral lineage. Similar to coalescent models generally, this new conditional coalescent model is meant as an approximation for large populations with small outcrossing rates.

Figure 1 shows simulation results of pedigrees and pairwise coalescence times for the model of selfing used by Möhle (1998), which corresponds to $\alpha_N = \alpha$ constant in our model. Specifically, for each of three selfing rates, 50 independent pedigrees were simulated and for each of these two different individuals were sampled at random. Colored lines, one for each pedigree-plus-sample, depict the cumulative distribution function (CDF) of the coalescence time for two gene copies, one from each of the two sampled individuals. On the scale used in Figure 1, the corresponding expectation for the pedigree-averaging model (not depicted) would be the CDF of an exponential distribution with parameter $(2-\alpha)^{-1}$. Many of the CDFs in Figure 1(a) have something close to this shape, but those in Figure 1(b) and Figure 1(c) look nothing like it. Instead they are characterized by random jumps in probability at random times.

As we will show in detail in what follows, the deviations from the predictions of the unconditional coalescent model displayed in Figure 1 are not due to the relatively small population size used in these simulations ($N = 1000$). Instead they illustrate that the behavior we will prove occurs when $N(1-\alpha_N)$ remains finite in the limit $N \to \infty$ is relevant even for such small populations. The value of $\alpha = 0.99$ used in Figure 1(b) is similar to estimates of the selfing rates for the model species *Arabidopsis thaliana* and *Caenorhabditis elegans* (Abbott and Gomes, 1989; Sellinger et al., 2020; Barrière and Félix, 2005). Of course, $N = 1000$ would be a very small population size for either of these species, though it may be realistic for some local populations. In any case, species or populations for which $N(1-\alpha)$ is not very large will have genetic ancestries very unlike what is expected under the unconditional or pedigree-averaging model. Instead, they are captured by the conditional coalescent model with 'limited outcrossing' which is characterized by a system of coalescing random walks on a random ancestral graph, described in Section 4.

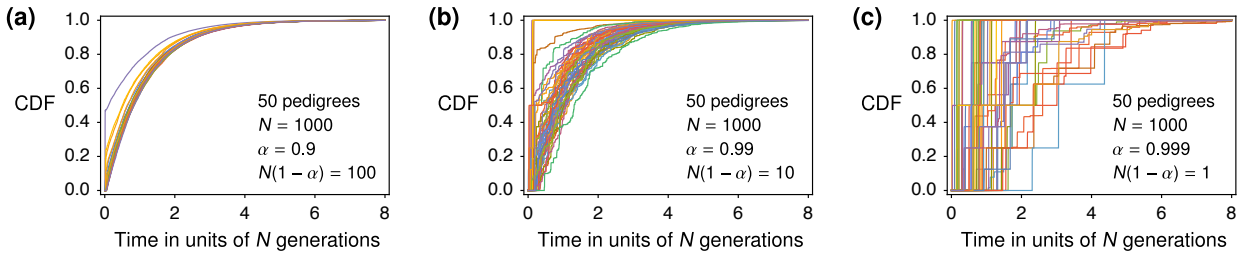


Figure 1: Cumulative distribution functions of the coalescence times for a sample of size $n = 2$ on each of 50 pedigrees simulated under the Wright-Fisher model with partial selfing. Panels (a), (b) and (c) show the results for three different values of the probability of selfing, respectively, for $\alpha \in \{0.9, 0.99, 0.999\}$. Coalescence probabilities in each generation given the pedigree and the sampled individuals were calculated numerically by the method described in Wakeley et al (2012).

1.4 Plan of the present work

To illustrate our key ideas, we focus on the coalescence time for a sample of size $n = 2$, which already reveals the non-trivial effects that pedigree has on gene genealogies. Again, we begin with a diploid Moran model of reproduction with selfing rate α_N and constant size N. The details of this model are presented in Section 2.

To establish the conditional limit for different regimes of α_N, we adapt the method from Diamantidis et al. (2024), which itself was adapted from Birkner et al. (2013). We convert the problem of conditional convergence to L^2 convergence for the distribution of the minimum of two conditionally independent pairwise coalescence times from two independently assorting loci on the pedigree. Unlike the work in Diamantidis et al. (2024), in both the conditional and unconditional case, Möhle's result (Möhle, 1998, Lemma 1, Theorem 1) does not apply, at least not in the obvious way. Some of our results could be obtained using the extension by Möhle and Notohara (2016). Here, we provide an alternate calculation of the coalescence times by examining the random number of opportunities for coalescence between genetic lineages in different individuals, which we will refer to as overlap events, and the timings of these in relation to splitting events in the ancestral process, which result from outcrossing.

The organization of the paper is as follows. In Section 2, after detailing our Moran model, we make precise the pedigree and the gene genealogy. Our results for the unconditional and the conditional distributions of pairwise coalescence times are presented in Section 3 and Section 4, respectively. In the main result section, Section 4, we focus on the case $n = 2$. In Section 4.2 we provide some conjectures about the ancestral process when $n > 2$. In Sections 5, 6 and 7 we give proofs for the cases of limited outcrossing, negligible outcrossing, and partial selfing, respectively.

2 A discrete-time diploid Moran Model with selfing

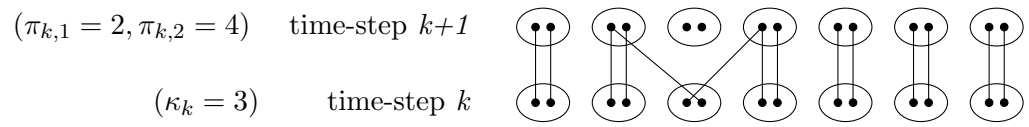
We consider a diploid, monoecious, panmictic (well mixed and randomly mating) population of constant fixed size of N individuals labeled $I = \{1, ..., N\}$ with discrete time-steps and overlapping generations. At each time-step, exactly one reproduction event occurs and one offspring is reproduced. With probability $1 - \alpha_N$ the offspring has two distinct parents; with probability α_N the offspring has exactly one parent. The parent(s) are chosen uniformly at random from the previous time-step. In the case of two parents, these are chosen without replacement. The reproduction

events at different time-steps are independent and identically distributed.

Explicitly, for each non-negative integer $k \in \mathbb{Z}_+ = \{0, 1, 2, 3, ...\}$, the reproduction event between consecutive time-steps k and $k+1$ in the past is as follows:

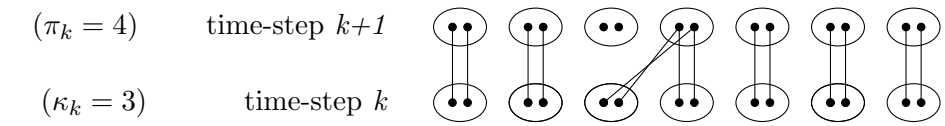
1. (distinct parents) With probability $1 - \alpha_N$, a single offspring is produced by outcrossing. A triplet containing the two parents and one offspring $(\pi_{k,1}, \pi_{k,2}, \kappa_k) \in I^3$ is chosen uniformly at random with $\pi_{k,1} \neq \pi_{k,2}$ (i.e. the parents are distinct). The offspring κ_k is an individual in time-step k, while the offspring's parents are the individuals in time-step $k+1$ with labels $\pi_{k,1}$ and $\pi_{k,2}$.

   The offspring has two genes copies, one inherited from each parent. Genetically, the offspring is produced according to Mendel's laws, which means each of the two gene copies in a parent is equally likely to be the one transmitted to the offspring. An example outcrossing event in a population of size $N = 7$ is depicted below.



2. (selfing) With probability α_N, a parent-offspring pair with labels $(\pi_k, \kappa_k) \in I^2$ is chosen uniformly at random without replacement. The offspring κ_k in time-step k is the offspring of a single individual (the parent) with label π_k at time-step $k+1$. This can be viewed as having $\pi_{k,2} = \pi_{k,1}$ in the previous reproduction event.

   As above, the offspring is produced according to Mendel's laws which means each of the two gene copies in the parent is independently and equally likely to be inherited by each of those of the offspring. An example selfing event in a population of size $N = 7$ is depicted below.



The reproduction events at different time-steps are independent. Note that at time-step k, the offspring κ_k is a new individual introduced and it replaces the one with the same label in the previous step $k+1$ (death-birth for the label κ_k); all the remaining $N-1$ individuals in time-step k are the same individuals as those in time-step $k+1$. Hence this model has overlapping generations.

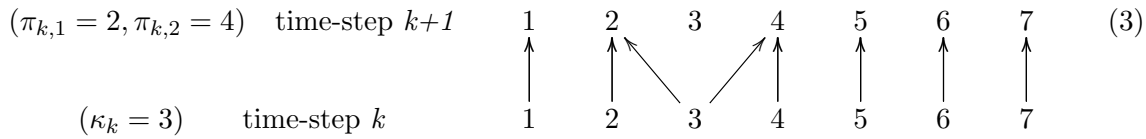
Remark 2.1. When $\alpha_N = 0$, our model is exactly the one introduced in Coron and Le Jan (2022). When α_N is equal to a fixed p in $[0, 1)$, our model is exactly that in Linder (2009).

2.1 The pedigree as important partial information of the population

The above population dynamics generates a random graph, which we will call $\mathcal{G}_N$, with vertex set $I \times \mathbb{Z}_+$, and which represents the pedigree of the population. In the following, we consider the pedigree without specifying outcomes of genetic transmission.

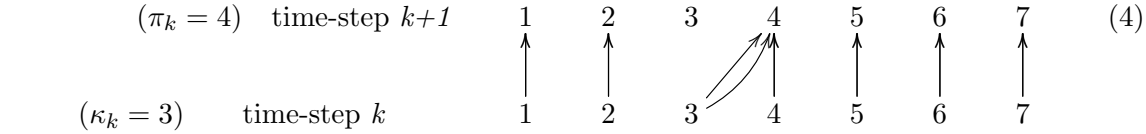
Between consecutive time-steps with two distinct parents, the pedigree has one edge from offspring (κ_k, k) to parent $(\pi_{k,1}, k+1)$, one from offspring (κ_k, k) to parent $(\pi_{k,2}, k+1)$, and $N-1$

single edges for the same individual from (j,k) to $(j,k+1)$ for $j \in I \setminus \{\kappa_k\}$. The portion of the pedigree corresponding to the example above is



(3)

Between consecutive time-steps with selfing, the pedigree has a double edges from (κ_k, k) to $(\pi_k, k+1)$, and again $N-1$ single edges from (j,k) to $(j,k+1)$ for $j \in I \setminus \{\kappa_k\}$. The portion of the pedigree corresponding to the example above is



(4)

Hence, the offspring always has 2 edges coming out of it which trace upward (towards the past); and all the other $N-1$ individuals have a single edge.

Repetition of this process with these two possibilities for each time step backward into the indefinite past results in a single realized pedigree of the population. Patterns of ancestry at each genetic locus are outcomes of Mendelian transmission conditional on this one pedigree. Unlinked loci are transmitted independently through the pedigree. In Figure 2, we illustrate a realization of both the population dynamics with genetic transmission and the corresponding pedigree for the first 4 generations in the past for a population of size 6.

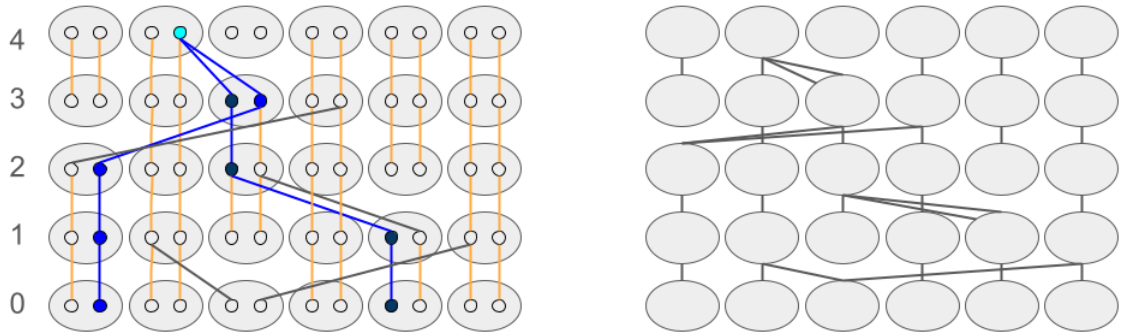


Figure 2: A realization of our Moran process with $N = 6$ individuals (left image) and the corresponding pedigree (right image). On the left, two genes $(X_0, Y_0) = (2, 9)$ are sampled in the present time-step 0, and their lineages are highlighted by the solid dots. Namely, $\{(X_k, Y_k)\}_{k=0}^{4} = \{(2,9),(2,9),(2,5),(6,5),(4,4)\}$. These two lineages coalesce in time-step 4 because $X_4 = Y_4$ but $X_k \neq Y_k$ for $k = 0, 1, 2, 3$.

2.2 Pairwise coalescence time: unconditional versus conditional on the pedigree

In our population model each of the N individuals has two copies of each genetic locus. Suppose we number these $J = \{1, ..., 2N\}$ such that the individual with label $i \in I$ has gene copies $2i-1$ and $2i$. We will generally be interested in computing various statistics for random samples of gene copies from J by following their ancestral genetic lineages backward in time.

Suppose we sample two gene copies X_0 and Y_0 without replacement from the $2N$ present at time $k = 0$. Let X_k and Y_k be their ancestral lineages k time-steps in the past. Then $(X_k)_{k\in\mathbb{Z}_+}$ and $(Y_k)_{k\in\mathbb{Z}_+}$ are two correlated Markov chains with state space J that have the same transition probabilities. Our main object of study, for sample size $n = 2$, is

$$\tau^{(N,2)} := \inf\{k \in \mathbb{Z}_+ : X_k = Y_k\}, \tag{5}$$

the (pairwise) coalescence time of the sample of two genes. The left image in Figure 2 shows a realization of the population process where $\tau^{(N,2)} = 4$ and $N = 6$.

The value of $\tau^{(N,2)}$ is completely determined by the population process. However, knowing the pedigree is *not enough* to determine the exact value of $\tau^{(N,2)}$, due to Mendelian randomness. For instance, even if we know the pedigree is exactly the one the right in Figure 2 and we know that $X_0 = 2$, we do not know what X_3 is: X_3 can be in one of the genes $\{5, 6, 7, 8\}$ with equal probability.

3 Unconditional distribution of pairwise coalescence time

The unconditional distribution (i.e. average over the randomness of the pedigree) of $\tau^{(N,2)}$ can be obtained by considering a Markov process with 3 states {coal, same, diff} representing respectively that the gene copies have coalesced, the gene copies are in the same individual but have not coalesced, and the two gene copies are in two difference individuals.

$$\text{diff} = \left(\bullet\ \right)\left(\bullet\ \right)\ \text{same} = \left(\bullet\ \bullet\right),\ \text{coal} = \left(\bullet\ \right)$$

The one-step transition matrix $\mathbf{\Pi}_N$ of this Markov process, under our diploid Moran model, is

$$\mathbf{\Pi}_N := \begin{array}{c|ccc} & \text{diff} & \text{same} & \text{coal} \\ \hline \text{diff} & 1 - \frac{2}{N(N-1)} & \frac{1}{N(N-1)} & \frac{1}{N(N-1)} \\ \text{same} & (1-\alpha_N)\frac{1}{N} & \frac{N-1}{N} + \alpha_N \frac{1}{2N} & \alpha_N \frac{1}{2N} \\ \text{coal} & 0 & 0 & 1 \end{array}$$

For $q \in \{\text{diff, same, coal}\}$, we let $\mathbb{P}_q$ be the probability under which the initial state (i.e. the sampling configurations) is q. We use $\xrightarrow{d}$ and $\overset{d}{=}$ to denote convergence in distribution and equal in distribution respectively.

Theorem 3.1 (Unconditional limiting distribution)**.** *Suppose $\alpha_N \to \alpha \in [0, 1]$ as $N \to \infty$. Then $N^{-2}\tau^{(N,2)}$ converges in distribution as follows:*

$$N^{-2}\tau^{(N,2)} \xrightarrow{d} \begin{cases} 0 & \text{under the law } \mathbb{P}_{\text{same}}(\cdot|\mathcal{O} = 0) \\ \text{Exp}\left(\frac{2}{2-\alpha}\right) & \text{under the laws } \mathbb{P}_{\text{diff}}(\cdot) \text{ and } \mathbb{P}_{\text{same}}(\cdot|\mathcal{O} \geq 1) \end{cases}$$

where

$$\mathcal{O} := |\{k \geq 1 : \hat{X}_k = \hat{Y}_k,\ \hat{X}_{k-1} \neq \hat{Y}_{k-1}\}|$$

is the total number of time-steps when our two sample lineages transition from belonging to two distinct individuals to belonging to a single individual, and $\hat{X}_k, \hat{Y}_k$ denote the labels of the individuals to which the genes X_k and Y_k belong.

The convergence to an exponential random variable $\mathrm{Exp}\left(\frac{2}{2-\alpha}\right)$ in Theorem 3.1, under $\mathbb{P}_{\mathrm{diff}}$, shows that our Moran model gives the same result as the Wright-Fisher model with partial selfing (Nordborg and Donnelly, 1997; Möhle, 1998) when one averages over the pedigree for a fixed rate of selfing $\alpha \in [0,1]$. Kogan et al. (2023, Lemma 2) recently obtained this result in the broader context of selfing plus recombination under Wright-Fisher reproduction (and averaging over the pedigree).

Two kinds of events are key to our analysis: *overlap events* where two sample lineages in distinct individuals transition to belonging to the same individual, and *splitting events* where two sample lineages in the same individual transition to belonging to two distinct individuals. We use $\mathcal{O}$ to denote the number of overlap events. Note that an overlap event results also in coalescence with probability 1/2. In addition, under $\mathbb{P}_{\mathrm{same}}$ the two lineages may coalesce before they split, in which case $\mathcal{O} = 0$. Under $\mathbb{P}_{\mathrm{diff}}$ it holds that $\mathcal{O} \geq 1$ almost surely. We prove in Lemma 10.1 that

$$\mathbb{P}_{\mathrm{diff}}(\mathcal{O} = 0) = 0 \quad \text{and} \quad \mathbb{P}_{\mathrm{same}}(\mathcal{O} = 0) = \frac{\alpha_N}{2 - \alpha_N}. \tag{6}$$

Remark 3.2. *For any $k \in \mathbb{Z}_+$,*

$$\mathbb{P}_{\mathrm{diff}}(\tau^{(N,2)} > k) = (1,0,0)\, \mathbf{\Pi}_N^k\, (1,1,0)^T$$
$$\mathbb{P}_{\mathrm{same}}(\tau^{(N,2)} > k) = (0,1,0)\, \mathbf{\Pi}_N^k\, (1,1,0)^T$$

One may wish to use (Möhle, 1998, Theorem 1) to calculate the limit of $\mathbf{\Pi}_N^{\lfloor tN^2 \rfloor}$ to obtain the limiting distribution of $N^{-2}\tau^{(N,2)}$ under both $\mathbb{P}_{\mathrm{diff}}$ and $\mathbb{P}_{\mathrm{same}}$. However, this usual approach and its generalization Möhle and Notohara (2016) do not seem to work here directly.

In Section 10.3, we prove Theorem 3.1 without using Möhle's lemma (Möhle, 1998, Theorem 1) nor its generalization (Möhle and Notohara, 2016).

4 Main results: conditional coalescence times given the pedigree

Our main results are about the conditional distribution of $\tau^{(N,2)}$ given the pedigree and the sampled pair of individuals. We consider a sequence of our diploid Moran model indexed by N, under three different regimes for the selfing rate α_N in the limit. Our results are summarized in Table 1 and precised in Theorems 4.3 and 4.4. The first corresponds to the model of partial selfing previously studied by others, only excluding cases in which $\alpha_N \to 1$ more slowly than $1/N$. The second is what we call limited outcrossing, and happens when $\alpha_N \to 1$ at rate λ/N for some $\lambda \in (0, \infty)$. The third is when outcrossing is negligible, because $\alpha_N \to 1$ faster than $1/N$. As we will show, limited outcrossing interpolates between partial selfing and negligible outcrossing. As $\lambda \to \infty$, it becomes indistinguishable from partial selfing with $\alpha = 1$. As $\lambda \to 0$ it becomes indistinguishable from negligible outcrossing.

Following the notation in Diamantidis et al. (2024), we let $\mathcal{A}_N$ be the σ-algebra generated by both the pedigree $\mathcal{G}_N$ and the labels $\hat{X}_0$ and $\hat{Y}_0$ of the sampled individuals. That is, we let

$$\mathcal{A}_N := \sigma\left(\mathcal{G}_N,\, \hat{X}_0,\, \hat{Y}_0\right). \tag{7}$$

Our main results say that the conditional distributions of $N^{-2}\tau^{(N,2)}$ given $\mathcal{A}_N$ converge to different distributions, depending on the sampling configuration (whether the two sampled genes are in the same individual or not) and on the three regimes of the selfing probability. Furthermore, the limiting distributions can retain information of the random pedigree as $N \to \infty$.

Before stating our results rigorously, we first give a quick summary in Table 1 and some illustrations in Figure 3. We are interested in the event (E_t in Table 1) that the rescaled conditional coalescence time $N^{-2}\tau^{(N,2)}$ for a sample in state $q \in \{\text{diff}, \text{same}\}$ is greater than a fixed time t, namely $\mathbb{P}_q(N^{-2}\tau^{(N,2)} > t \mid \mathcal{A}_N)$ which we denote $\mathbb{P}_{\text{diff}}(E_t \mid \mathcal{A}_N)$ and $\mathbb{P}_{\text{same}}(E_t \mid \mathcal{A}_N)$ in Table 1.

regime	selfing rate as $N \to \infty$	$\mathbb{P}_{\text{diff}}(E_t \mid \mathcal{A}_N)$	$\mathbb{P}_{\text{same}}(E_t \mid \mathcal{A}_N)$
partial selfing	$N(1-\alpha_N) \to \infty$	$e^{-\frac{2}{2-\alpha}t}$	$2^{1-\text{Geom}(1-\alpha)}e^{-\frac{2}{2-\alpha}t}$
limited outcrossing	$N(1-\alpha_N) \to \lambda \in (0,\infty)$	"random staircase"	0
negligible outcrossing	$N(1-\alpha_N) \to 0$	$\mathbf{1}_{\{\text{Exp}(2)>t\}}$	0

Table 1: Summary of results for the conditional coalescence time of a sample of size $n = 2$, as $N \to \infty$, where E_t denotes the event $\{N^{-2}\tau^{(N,2)} > t\}$ for a fixed time t. In the last column, $\text{Geom}(1-\alpha)$ is a random variable U satisfying that $\mathbb{P}(U = k) = \alpha^{k-1}(1-\alpha)$ for $k \in \mathbb{Z}_+$. In the middle column, $\mathbf{1}_{\{\text{Exp}(2)>t\}}$ is the indicator of the event $\{\text{Exp}(2) > t\}$, hence it is a Bernoulli random variable B so that $\mathbb{P}(B = 1) = e^{-2t} = 1 - \mathbb{P}(B = 0)$.

From Table 1, we see that under the conditional probability $\mathbb{P}_{\text{diff}}(\cdot \mid \mathcal{A}_N)$, the limiting survival function of $N^{-2}\tau^{(N,2)}$ changes from deterministic to random, as the selfing rate increases from partial selfing to limited outcrossing. In the partial-selfing regime, the deterministic distribution $\text{Exp}(\frac{2}{2-\alpha})$ is the same as that for the unconditional case (when we average over pedigrees as in Theorem 3.1). In the limited-outcrossing regime, the limiting survival function is a random random function which we call a "random staircase" for reasons which will become clear. In the negligible-outcrossing regime, the limiting survival function is the random function $\mathbf{1}_{\{\text{Exp}(2)>t\}}$. Note that the expectation of $\mathbf{1}_{\{\text{Exp}(2)>t\}}$, namely e^{-2t}, is the same as the unconditional probability $\lim_{N\to\infty} \mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} > t)$ when $\alpha_N \to \alpha = 1$. This makes sense because

$$\mathbb{E}[\mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} > t \mid \mathcal{A}_N)] = \mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} > t) \tag{8}$$

for all $N \in \mathbb{N}$ and $t \in \mathbb{R}_+$. It is always true that the expectation of the conditional limit is the same as the corresponding unconditional limit.

Besides, there is a simple explanation for $\mathbf{1}_{\{\text{Exp}(2)>t\}}$ in the negligible-outcrossing regime in Table 1, by considering the extreme case in which $\alpha_N = 1$ for all N. In this extreme case, there is no out-crossing for the two gene copies of any individual, so there is no splitting event for the lineages of the sampled individuals $\hat{X}_0$, $\hat{Y}_0$. The lineages of the sampled individuals $\{\hat{X}_0, \hat{Y}_0\}$ will overlap after a random number of time-steps distributed as a Geometric random variable with mean $\frac{N(N-1)}{2}$, hence they overlap after approximately N^2T_0 many generations, where T_0 is an Exp(2) random variable. Upon the overlap, the two sampled genes will stay in the same individual(s) and coalesce after $O(N)$ many time-steps, i.e. coalesce immediately under the N^2 timescale.

Further, in the case of limited outcrossing, with $N(1-\alpha_N) \to \lambda \in (0,\infty)$, we will see that our random walks on the pedigree will converge to a particle system where each ancestral lineage splits at rate λ and each pair of ancestral lineages coalesce at rate 2. In particular, no matter which ancestral lineage our two sample particles belong, they coalesce together at rate 2, which is what we see in the unconditional case.

Under the conditional probability $\mathbb{P}_{\text{same}}(\cdot \mid \mathcal{A}_N)$, the limiting survival function of $N^{-2}\tau^{(N,2)}$ changes from random to deterministic (namely 0), as the selfing rate increases from partial selfing to limited outcrossing. Note that in the last column of Table 1, the expectation of the conditional probability

$$\mathbb{E}\left[2^{-\text{Geom}(1-\alpha)}e^{-\frac{2}{2-\alpha}t}\right] = \frac{2-2\alpha}{2-\alpha}e^{-\frac{2}{2-\alpha}t}$$

is the same as the unconditional probability

$$\lim_{N\to\infty} \mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t) = \lim_{N\to\infty} \mathbb{P}_{\text{same}}(\mathcal{O} \geq 1)\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t \mid \mathcal{O} \geq 1)$$

according to (6). This makes sense in view of (8).

In Figure 3, we illustrate the limiting behavior of the random, conditional cumulative distribution function (conditional CDF) $t \mapsto \mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} \leq t|\mathcal{A}_N)$ for our three regimes about α_N.

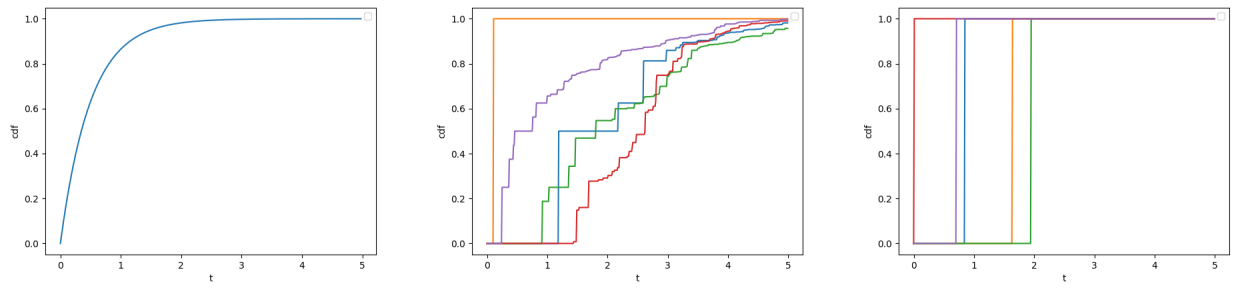


Figure 3: The limiting conditional CDF $t \mapsto \lim_{N\to\infty} \mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} \leq t|\mathcal{A}_N)$ for our three assumptions about α_N: partial selfing, limited outcrossing, and negligible outcrossing. (**Left**) Partial selfing, where $\lim_N N(1-\alpha_N) = \infty$. The limiting CDF is deterministic and is exactly the CDF of an exponential random variable $\text{Exp}\left(\frac{2}{2-\alpha}\right)$; the figure takes $\alpha = 1$. (**Center**) Five realizations of the limiting CDF for $\lambda = 2$ under limited outcrossing, i.e. with $\lim_N N(1-\alpha_N) = \lambda = 2$. The limiting CDF is random and is described precisely in Theorem 4.3, and the corresponding survival function is what we called a "random staircase" in Table 1; cf. also Figure 1. (**Right**) Five realizations of the limiting CDF for the case of negligible outcrossing, where $\lim_N N(1-\alpha_N) = 0$. The limiting CDF is random, and it is the CDF of a random constant that is exponentially distributed with rate 2, i.e. the CDF is a Heaviside function with exponentially distributed jump time.

To describe the "random staircase" for the case of limited outcrossing with $\lambda \in (0, \infty)$ in Table 1, we first consider the lineage of a single sampled gene from a single individual $\hat{X}_0$. We wish to describe a process which encapsulates (i) the set of potential ancestors from whom this present-day gene may have been inherited, and (ii) the corresponding probabilities that each potential ancestor contains the gene. A realization of this process, starting from the individual $\hat{X}_0$ and parameterized by time-step $k \in \mathbb{Z}_+$ in the past, is illustrated in Figure 4.

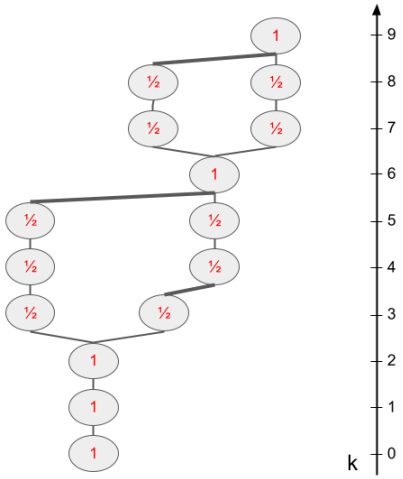


Figure 4: A realization of the set of individuals (potential ancestors) of the sampled gene within the individual $\hat{X}_0$, the oval at the bottom. The three thick edges depict selfing events. The fractional number in each individual is the probability that the sampled gene lies in that individual. After 3 time-steps, the individual to which the sample lineage belongs undergoes a fragmentation due to an outcrossing event.

Consider $n = 2$ sampled genes, one from each of the individuals $\hat{X}_0$ and $\hat{Y}_0$. There is a similar description to the set of potential ancestors, but now each potential ancestor has two probabilities, one for each gene. A realization of this process $\{Z_k^N\}_{k\in\mathbb{Z}_+}$ is illustrated in Figure 5.

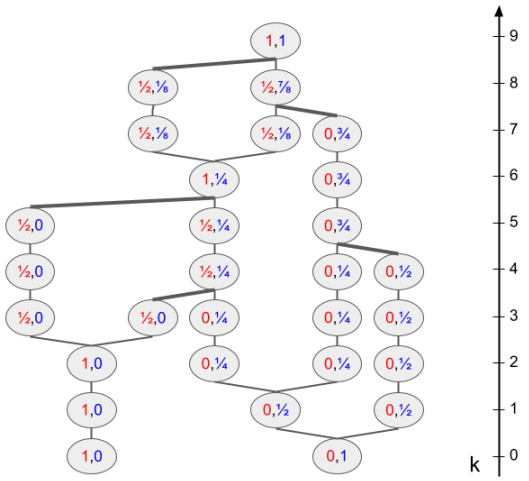


Figure 5: A realization of the process $\{Z_k^N\}_{k\in\mathbb{Z}_+}$, where Z_k^N represents the set of potential ancestors at time-step k in the past of the two genes in the sampled individuals $\hat{X}_0$ and $\hat{Y}_0$, together with the corresponding probabilities (in red and blue respectively) that each potential ancestor contains each of the sample lineages. The individuals at the bottom are $\hat{X}_0$ on the left and $\hat{Y}_0$ on the right. The initial condition is $Z_0 = \{((1,0),1),\,((0,1),2)\}$ where the first triple $((1,0),1)$ indicates that $\hat{X}_0$ has probabilities/masses $(1,0)$ and label 1, and the second triple $((0,1),2)$ indicates that $\hat{Y}_0$ has probabilities/masses $(0,1)$ and label 2.

The discrete-time process $Z^N = (Z_k^N)_{k\in\mathbb{Z}_+}$ is a fragmentation-coagulation process (Bertoin, 2006) starting with two particles $\{((1,0),1),\,((0,1),2)\}$, in which each individual (a potential ancestor) is viewed as a particle and the pairs of probabilities in the first two components within each potential ancestor are the "masses" of the particle. The third component of the particle is its label.

For each $k \in \mathbb{Z}_+$, the value of Z_k^N is a set of 3-tuples, where the sum of all the first two components of each 3-tuple is equal to $(1,1)$ because the total probability/mass for each of the two lineages must be 1. The third component is the label of the particle, which tracks the total structure of inheritance for the process. The state space $\mathcal{P}_{m,2}$ of $Z^N = (Z_k^N)_{k\in\mathbb{Z}_+}$ can be represented as

$$\mathcal{P}_{m,2} := \bigcup_{j=1}^{\infty} \mathcal{P}_{m,2}^{(j)}, \tag{9}$$

where

$$\mathcal{P}_{m,2}^{(j)} := \{(p_i, l_i) : 1 \le i \le j, p_i \in [0,1]^2, l_i \in \mathbb{Z}_+\}.$$

Here j represents the number of particles, and p_i and l_i represent, respectively, the masses and the label of the i-th particle. We endow $\mathcal{P}_{m,2}$ with a metric d such that $(\mathcal{P}_{m,2}, d)$ is a Polish space (a complete and separable metric space); see Section 10.2.

Lemma 4.2 says that if one unit of time is N^2 time-steps and if $N \to \infty$, then under limited outcrossing Z^N converges weakly to a continuous-time process with state space $\mathcal{P}_{m,2}$, described as follows.

Definition 4.1 (Fragmentation-coagulation process with rate λ)**.** *Let $\lambda \in \mathbb{R}_+$ be a fixed deterministic number. A fragmentation-coagulation process with rate λ is a continuous-time process with sample paths in the Skorokhod space $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{m,2})$ almost surely, such that*

(i) Each particle fragments independently at rate λ into two particles with equal masses which is half of that of the fragmenting particle, and each pair of particles coagulate independently at rate 2 into a particle whose masses are the sum of those of the two coagulating particles.

(ii) If i is the label of the fragmenting particle, then the labels of the two resultant particles are i and m, where m is the smallest positive integer such that all labels are distinct; if i and j are the labels of the two coagulating particles, then the resultant particle has label $\min\{i, j\}$.

In this paper, we let $Z_\lambda = (Z_\lambda(t))_{t\in\mathbb{Z}_+}$ be a fragmentation-coagulation process with rate λ, starting with two immiscible unit masses $Z_\lambda(0) = \{((1,0),1), ((0,1),2)\}$; i.e. it starts with two particles with masses $(1,0)$ and $(0,1)$ and with labels 1 and 2 respectively.

The terminology in Definition 4.1 is inspired by Bertoin (2006) who discusses fragmentation and coagulation processes of unit masses. The use of the word "immiscible" reflects the fact that the coagulation of two particles with masses (p,q) and (r,s) into a particle with masses $(p+r, q+s)$ results in no crossover between components; i.e., the two components of the masses do not mix.

Lemma 4.2. *Suppose $\lim_{N\to\infty} N(1-\alpha_N) = \lambda \in \mathbb{R}_+$. Then $(Z^N_{\lfloor tN^2 \rfloor})_{t\in\mathbb{R}_+}$ converges in distribution in $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{m,2})$ under $\mathbb{P}_{\text{diff}}(\cdot|\mathcal{A}_N)$ to the fragmentation-coagulation process Z_λ in Definition 4.1.*

Lemma 4.2 is proven in Section 6. Ignoring the "masses", e.g. ignoring the pairs of probabilities in the individuals in Figure 5, we obtain a simpler process $G^N = (G_k^N)_{k\in\mathbb{Z}_+}$ from Z^N that keeps track of the set of potential ancestors and their ancestral relationships (i.e. which splits into, and which two particles coagulate). It is simply the particle process consisting solely of the labels of the particles of Z^N. This process G^N has a graphical structure that represents all possible trajectories of $(\hat{X}_k, \hat{Y}_k)_{k\in\mathbb{Z}_+}$ and which is a subset of the population pedigree.

Similarly, ignoring the "masses" in Z_λ, we obtain a process which we denote by $G_\lambda = (G_\lambda(t))_{t\in\mathbb{R}_+}$. This continuous-time process is a particle system starting with two different particles, where each

particle splits into two with rate λ and each pair of particles coagulate independently at rate 2. The labelling process of the particles is retained from Z_λ. It is equal in distribution to an ancestral recombination graph (Griffiths, 1991; Griffiths and Marjoram, 1997) or an ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). Throughout this paper, we call G_λ an *ancestral graph*. A realization of an ancestral graph is presented on the left in Figure 6, and further examples are offered in Figure 7.

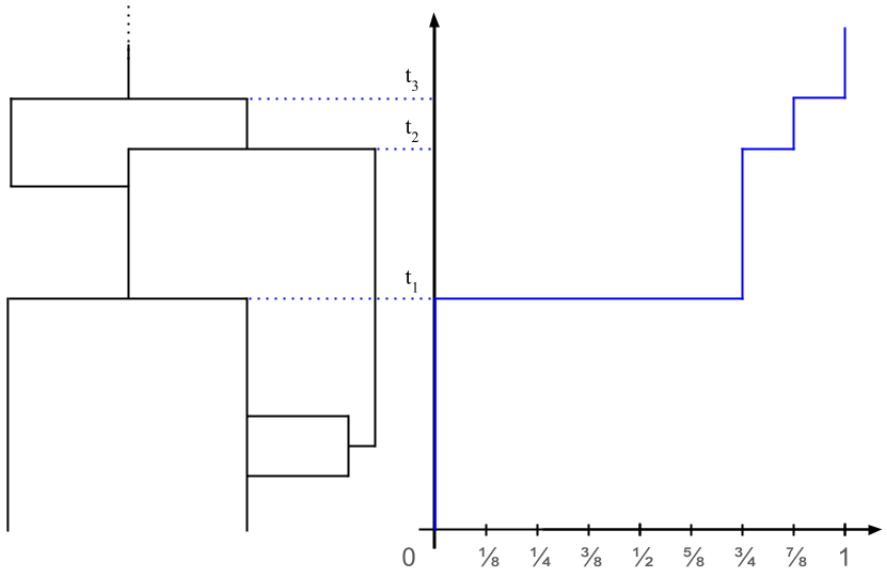


Figure 6: (**Left**) A realization G_λ of the random ancestral graph G_λ starting with two particles/nodes. For this particular realization, the overlap times of the sample lineages are $t_1 < t_2 < t_3$. (**Right**) The "random staircase" mentioned in Table 1 corresponding to this realization of G_λ. The conditional distribution of the limiting coalescence time T_λ, given G_λ, is obtained by tracing ancestral genetic lineages backwards in time along the graph G_λ. Given this G_λ, T_λ must take values in $\{t_1, t_2, t_3\}$. One can read off the conditional CDF of T_λ as follows: The node of G_λ at time t_1 contains $\frac{3}{4}$ of the genetics of the right lineage and the whole of that of the left lineage, so $\mathbb{P}(T_\lambda = t_1 | G_\lambda) = \frac{3}{4}$. Between t_1 and t_2, half of the left lineage splits into the left handle. At t_2, $\frac{1}{4}$ of the right lineage meets half of the left lineage, so $\mathbb{P}(T = t_2 | G_\lambda) = \frac{1}{8}$. The remainder of the right lineage meets the remainder of the left lineage at t_3 so $\mathbb{P}(T_\lambda = t_3) = \frac{1}{8}$.

Next, we consider two conditionally independent random walks on a given ancestral graph. Given $G_\lambda = (G_\lambda(t))_{t\in\mathbb{R}_+}$ and let $G_\lambda(0) = \{x(0), y(0)\}$ be the two different particles at $t = 0$. We let $(x_\lambda(t), y_\lambda(t))_{t\in\mathbb{R}_+}$ be a continuous-time process starting at $(x(0), y(0))$, such that

(i) $\{x_\lambda(t), y_\lambda(t)\} \subset G_\lambda(t)$ for all $t \in \mathbb{R}_+$,

(ii) at any fragmentation event, each of $(x_\lambda(t))_{t\in\mathbb{R}_+}$ and $(y_\lambda(t))_{t\in\mathbb{R}_+}$ will follow each of the two paths available with equal (i.e. 1/2) probability.

In particular, $(x_\lambda(t))_{t\in\mathbb{R}_+}$ and $(y_\lambda(t))_{t\in\mathbb{R}_+}$ are conditionally independent random walks on the ancestral graph G_λ starting at two different particles. The first meeting time of these random walks is

$$T_\lambda := \inf\{t \in \mathbb{R}_+ : x_\lambda(t) = y_\lambda(t)\}. \tag{10}$$

Then T_λ under the law of $\mathbb{P}(\cdot|G_\lambda)$ is the limit law of the pairwise coalescence time of two sample lineages given the pedigree under limited outcrossing.

4.1 Rigorous statements of convergence results

We are now ready to state our main results for coalescence times in Theorems 4.3 and 4.4. Recall the sigma field $\mathcal{A}_N$ in (7) and the pairwise coalescence time $\tau^{(N,2)}$ defined in (5).

Theorem 4.3. *For any $t \in \mathbb{R}_+$, as $N \to \infty$,*

$$\mathbb{P}_{\text{diff}}\left(N^{-2}\tau^{(N,2)} > t \,|\, \mathcal{A}_N\right) \xrightarrow{d} \begin{cases} e^{-\frac{2}{2-\alpha}t} & \text{if} \quad N(1-\alpha_N) \to \infty \ \text{ and } \ \alpha_N \to \alpha \in [0,1] \\ \mathbb{P}(T_\lambda > t \,|\, G_\lambda) & \text{if} \quad N(1-\alpha_N) \to \lambda \in \mathbb{R}_+ \\ \mathbb{1}_{\{\text{Exp}(2) > t\}} & \text{if} \quad N(1-\alpha_N) \to 0 \end{cases} .$$

Consider the random probability measure $\xi_N := \mathbb{P}_{\text{diff}}\left(N^{-2}\tau^{(N,2)} \in \cdot \,|\, \mathcal{A}_N\right)$ which is a random variable taking values in the space $\mathcal{P}(\mathbb{R}_+)$ of probability measures on $\mathbb{R}_+$. Theorem 4.3 precises how the sequence $\{\xi_N\}$ converges, as $N \to \infty$, to a $\mathcal{P}(\mathbb{R}_+)$-valued random variable ξ which happens to be deterministic in the partial selfing case and random in the other two cases. The random variable ξ is the law of an $\text{Exp}\left(\frac{2}{2-\alpha}\right)$ random variable in the first case, the conditional law of T_λ given G_λ in the second case, and the (random) law $\mathbf{1}_{\{\text{Exp}(2)\in\cdot\}}$ in the third case. With extra efforts one can justify the above convergence in the sense of weak convergence in $\mathcal{P}(\mathbb{R}_+)$.

Theorem 4.3 offers a description of the conditional distribution of the coalescence time $\tau^{(N,2)}$ for a sample of two gene copies from different individuals in a population of size N given the pedigree. Simulations of the random CDFs can be seen in Figure 3. It says that the law of $N^{-2}\tau^{(N,2)}$ for two lineages in different individuals is robust to the structure of the pedigree so long as the rate at which any individual undergoes a splitting $(1-\alpha_N)N^{-1}$ is significantly faster than the rate at which any two sample lineages in distinct individuals coalesce, $N^{-2} + O(N^{-3})$, in other words that the ratio of the former over the latter tends to infinity (see also Section 4.3). This robustness for the case of partial selfing is simply that the conditional limit agrees exactly with that in Theorem 3.1.If, on the other hand, the rate of coalescence is comparable to or dominates the rate of splitting of an individual, as in the cases of limited outcrossing and negligible outcrossing, the theorem demonstrates the non-robustness of the pairwise coalescence time to the pedigree. That is, the conditional pairwise coalescence time does not converge to its mean.

Analogously to Theorem 4.3, we obtain the limiting conditional distribution of the coalescence time when our samples are taken from the same individual.

Theorem 4.4. *Suppose $\alpha_N \to \alpha \in [0,1]$ as $N \to \infty$. For all $t \in \mathbb{R}_+$ we have convergence in distribution*

$$\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t | \mathcal{A}_N) \xrightarrow{d} \begin{cases} e^{-2t} & \text{if } \alpha = 0 \\ 2^{-U}e^{-2t} & \text{if } \alpha \in (0,1) \\ 0 & \text{if } \alpha = 1 \end{cases},$$

where U is a random variable satisfying $\mathbb{P}(U = k) = \alpha^k(1-\alpha)$ for $k \in \mathbb{Z}_+$.

Remark 4.5. *The quantity U represents the number of selfing events undergone by the individual from whom our two sample lineages are taken before undergoing a splitting event.*

4.2 Conditional genealogy for general sample size $n > 2$

We anticipate that our main results can readily be extended to general sample size $n > 2$. We do not provide details of the proofs in this paper, but we provide a description to the limiting conditional genealogy in this section.

Suppose we have n samples of genes from the population at time-step 0, where $n > 2$ is an arbitrary fixed number. First consider the coalescent model for lineages in different individuals (corresponding to $\mathbb{P}_{\text{diff}}$ in the $n = 2$ case). The ancestral process $(\Pi_k^N)_{k\in\mathbb{Z}_+}$ takes values in partitions of $\{1, \ldots, n\}$, the set of which is denoted by $\mathcal{E}_n$. The ancestral process partitions labels of our n sample lineages according to whether they have coalesced or not i.e. i belongs to the same partition element of Π_k^N as j if and only if the ith and jth sample lineages have coalesced by time-step k. We provide some conjectures as to the behavior of the ancestral process for the $n > 2$ case.

Recall that G_λ is an ancestral graph starting with two particles, or two initial nodes. Here we let G_λ^n be an ancestral graph starting with n initial nodes, Clearly, $G_\lambda^2 = G_\lambda$.

Given G_λ^n. We let $G_\lambda^n(0) = \{x_i(0)\}_{1\le i\le n}$ be the set of the n particles at $t = 0$, and we let $((x_i(t))_{1\le i\le n})_{t\in\mathbb{R}_+}$ be a continuous time-process that starts at $(x_i(0))_{1\le i\le n}$ and satisfies

(i) $x_i(t) \in G_\lambda(t)$ for all $t \in \mathbb{R}_+$ and all $1 \le i \le n$,

(ii) at any fragmentation event, each $(x_i(t))_{t\in\mathbb{R}_+}$ will follow each of the two paths available with equal (i.e. 1/2) probability,

(iii) when any two different particles meet, they coalesce into one particle.

We can therefore define a $\mathcal{E}_n$-valued partition process $P_\lambda = (P_\lambda(t))_{t\in\mathbb{R}_+}$ defined by the coalescence of the particles i.e. $i \sim_{P_\lambda(t)} j$ if and only if $x_i(t) = x_j(t)$. We call this process the particle partition process.

Let $\mathcal{A}_N$ denote the σ-algebra generated by both the pedigree and the labels of the n individuals from whom we sample the n sample lineages. As before, we use $\mathbb{P}_{\text{diff}}$ to capture conditioning on our n samples coming from n distinct individuals.

Our proofs for the $n = 2$ case in Theorem 4.3 suggest the following conjecture: Under the conditional law $\mathbb{P}_{\text{diff}}(\cdot|\mathcal{A}_N)$, the sequence of $\mathcal{E}_n$-valued processes $\Pi_N = (\Pi^N_{\lfloor tN^2\rfloor})_{t\in\mathbb{R}_+}$ converges in $\mathcal{D}(\mathbb{R}_+, \mathcal{E}_n)$ to a random variable

$$\Pi = \begin{cases} \mathcal{K}_n & \text{if } N(1-\alpha_N) \to \infty \quad \text{and } \alpha_N \to \alpha \\ P_\lambda \quad \text{under } \mathbb{P}(\cdot|G_\lambda^n) & \text{if } N(1-\alpha_N) \to \lambda \in \mathbb{R}_+ \\ K & \text{if } N(1-\alpha_N) \to 0 \end{cases} \tag{11}$$

where $\mathcal{K}_n$ is the standard Kingman n-coalescent with time rescaled by a factor $\frac{2}{2-\alpha}$ and K is a single realization of the Kingman n-coalescent with time rescaled by a factor $\frac{2}{2-1} = 2$.

Furthermore, just as the limit in the case of limited outcrossing garnered the same result in the unconditional limit in Proposition 15 as the splitting rate diverged to infinity, we expect that as $\lambda \to \infty$, P_λ under $\mathbb{P}(\cdot|G_\lambda^n)$ converges in $\mathcal{D}(\mathbb{R}_+, \mathcal{E}_n)$ to the Kingman n-coalescent with time rescaled by a factor 2.

Now consider a sample of size n which contain $2m$ gene copies together as pairs in m individuals and $n - 2m$ gene copies each in different individuals. cf. Nordborg and Donnelly (1997). The most recent part of the ancestry, of duration zero in the limiting model, will involve m independent Bernoulli outcomes such that a random number X of these pairs will coalesce. Under partial selfing, each of the m individuals will have its own number of selfing generations before its first outcrossing event, given by independent draws of the random variable U in Theorem 4.4. Let the outcomes be $\{k_1, k_2, \ldots, k_m\}$. Then $X \sim \sum_{i=1}^m \text{Bernoulli}\left(1 - 2^{-k_i}\right)$. Under limited outcrossing and negligible outcrossing, $\alpha_N \to \alpha = 1$ and $X \equiv m$. Thus, whereas $n > 2$ under partial selfing is a

straightforward extension what is in Theorem 4.4, limited outcrossing and negligible outcrossing introduce an additional Kingman n-coalescent (with time rescaled by a factor 2) for the $n-m$ lines in different individuals following this initial instantaneous adjustment to the sample.

4.3 Why are there different regimes?

As indicated in Table 1, the regimes are distinguished by the magnitude of $\lim_{N\to\infty} N(1-\alpha_N)$.

The quantity $N(1-\alpha_N)$ has a clear interpretation in terms of the population process. In one time step, two ancestral lineages in two distinct individuals coalesce with probability

$$\frac{1}{N(N-1)} \sim \frac{1}{N^2}.$$

The two ancestries overlap as a result of either selfing or outcrossing with probability $2/(N(N-1))$, then the lineages trace back to the same gene copy in the parent (or not) with probability $1/2$. Two sample lineages in the same individual split, in this case necessarily by outcrossing, with probability

$$(1-\alpha_N)\frac{1}{N}.$$

Therefore the limit of $N(1-\alpha_N)$ is equal to the limit of the ratio of the splitting probability of two lineages in one individual and the coalescence probability (or $1/2$ the overlap probability) of two lineages in two individuals.

We can understand the phase transition in the limiting behavior of pairwise coalescence times conditional on the pedigree in terms of the relative magnitudes of the overlap and splitting rates in the population dynamics which produces the pedigree. Partial selfing is where splitting events dominate. Limited outcrossing is where the two are comparable. Negligible outcrossing is where overlap events dominate. Table 2 shows the probabilities of events and two examples.

Event	**Probability**	**Example: $\boldsymbol{\alpha_N = \frac{1}{2}}$**	**Example: $\boldsymbol{\alpha_N = 1 - \frac{\lambda}{N}}$**
Splitting	$(1-\alpha_N)N^{-1}$	$\frac{1}{2}N^{-1}$	λN^{-2}
Overlap	$2N^{-1}(N-1)^{-1}$	$O(N^{-2})$	$2N^{-2}+O(N^{-3})$
by outcrossing	$2(1-\alpha_N)N^{-1}(N-1)^{-1}$	$O(N^{-2})$	$O(N^{-3})$
by selfing	$2\alpha_N N^{-1}(N-1)^{-1}$	$O(N^{-2})$	$2N^{-2}+O(N^{-3})$

Table 2: A splitting event involves a particular individual undergoing an outcrossing event. An overlap event results in two sample lineages in two distinct individuals transitioning to both belonging to the same individual. This can occur when the two individuals are involved in either an outcrossing event or a selfing event. The case $\alpha_N = \frac{1}{2}$ is an example of partial selfing. The case $\alpha_N = 1 - \frac{\lambda}{N}$ is an example of limited outcrossing.

The partial-selfing regime includes $\alpha_N = \alpha \in [0,1)$, which was the domain of the paper by Linder (2009), but also includes α_N converging to 1 sufficiently slowly, for example with $\alpha_N = 1 - \frac{1}{\sqrt{N}}$. In this regime, when our samples begin in distinct individuals, the pedigree in fact gives us no additional information as to the pairwise coalescence times as the population size goes to infinity. As the splitting rate is much faster than the overlap rate, any given sample lineage will undergo many splitting events before it is involved in an overlap. The pedigree has very little to say about "where" the lineage will be by the time we expect there to be an overlap.

On the opposite side of the spectrum is the negligible-outcrossing regime, where $N(1-\alpha_N) \to 0$, which includes the cases $\alpha_N = 1$ and for example $\alpha_N = 1 - \frac{1}{N^2}$. Here, overlap events dominate

splitting events. Sample lineages from distinct individuals will never undergo a splitting before they coalesce. The individual to which each sample lineage belongs is readable directly from the pedigree with high probability for a very long time, long enough for coalescence to occur with high probability. As such, the time it takes for coalescence is, with high probability, directly readable from the pedigree, at least up to a error of order $O(N)$ which is negligible under the usual N^2 timescale of the Moran model.

The limited-outcrossing regime, where $N(1-\alpha_N)$ converges to a finite, non-negative constant λ, exhibits behavior intermediate to partial selfing and negligible outcrossing. Here, as the splitting and overlap rates are comparable, the uncertainty born by splitting in the partial-selfing regime is tempered by the determinism of the negligible-outcrossing regime. The result is that there is a continuous-time directed graph G_λ that encapsulates all possible trajectories of our sample lineages, and that is determined, but not over-determined, by the pedigree. The limiting distribution of the pairwise coalescence time of the two sample lineages in distinct individuals is equal in distribution to the first time when these two random walks on this graph meet.

5 Properties of the scaling limiting under limited outcrossing

The novel scaling limits we obtained in Theorems 4.3 and 4.4, especially the coalescing random walk on the ancestral graph, are worth analysing. This is because they maybe insensitive to changes of the underlying model and therefore hold the promise of explaining a larger class of models than just our Moran model. For example, the simulations for the Wright-Fisher model in Figure 1 in the introduction shows similar behavior as that for our Moran model in Figure 3.

Observe first that the limited-outcrossing regime in Theorem 4.3 includes the case $\lambda = 0$, which means that the distribution of T_λ under $\mathbb{P}(\cdot|G_\lambda)$ (when $\lambda = 0$) is equal to that of H in the negligible-outcrossing regime, where H is a constant random variable whose value is exponentially distributed with rate 2. Corollary 15 below covers the other extreme: as $\lambda \to \infty$, the distribution of T_λ under $\mathbb{P}(\cdot|G_\lambda)$ converges to Exp(2) which corresponds to the partial-selfing regime with $\alpha = 1$. Hence, for lineages in different individuals, the limited-outcrossing regime covers all possible cases for which $\alpha_N \to 1$, as we tune the magnitude of λ. This highlights the significance of the limited-outcrossing regime.

We focus our analysis on the variance of the conditional survival probability $\mathbb{P}(T_\lambda > t|G_\lambda)$ (for all $t \in \mathbb{R}_{>0}$) and that of the conditional expectation $\mathbb{E}[T_\lambda|G_\lambda]$. We obtain explicit formulas and their asymptotics (as $\lambda \to 0$ or $\lambda \to \infty$). The starting point of our analysis is the following observation.

Lemma 5.1. *For any $\lambda \in \mathbb{R}_{>0}$ and $t \in \mathbb{R}_{>0}$,*

$$\mathbb{E}[\mathbb{P}(T_\lambda > t|G_\lambda)^2] = \mathbb{P}(T_\lambda \wedge T'_\lambda > t) \quad \text{and} \quad \mathrm{Var}(\mathbb{E}[T_\lambda|G_\lambda]) = \mathrm{Cov}\big(T_\lambda, T'_\lambda\big), \tag{12}$$

where T_λ and T'_λ are two conditionally independent copies of the pairwise hitting time given the same G_λ; they are defined in (10) with $(x_\lambda(0), y_\lambda(0)) = (x'_\lambda(0), y'_\lambda(0))$ and $x_\lambda(0) \neq y_\lambda(0)$.

Proof. We proof only the first equality in (12), the second follows from the same argument.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{P}_{x_\lambda(0)\neq y_\lambda(0)}(T_\lambda > t|G_\lambda)^2\right] &= \mathbb{E}\left[\mathbb{P}_{x_\lambda(0)\neq y_\lambda(0)}(T_\lambda > t|G_\lambda)\, \mathbb{P}_{x'_\lambda(0)\neq y'_\lambda(0)}(T'_\lambda > t|G_\lambda)\right] \\
&= \mathbb{E}\left[\mathbb{P}_{x'_\lambda(0)=x_\lambda(0)\neq y_\lambda(0)=y'_\lambda(0)}(T_\lambda > t,\, T'_\lambda > t|G_\lambda)\right] \\
&= \mathbb{P}_{x'_\lambda(0)=x_\lambda(0)\neq y_\lambda(0)=y'_\lambda(0)}(T_\lambda \wedge T'_\lambda > t),
\end{aligned}$$

where the second equality above follows from the conditional independence of T_λ and T'_λ. □

It following directly from the first equality in (12) that the variance of $\mathbb{P}(T_\lambda > t|G_\lambda)$ is equal to $\mathbb{P}(T_\lambda \wedge T'_\lambda > t) - e^{-4t}$. We first characterize the second moment (and hence also the variance) of the conditional survival probability $\mathbb{P}(T_\lambda > t|G_\lambda)$.

Lemma 5.2. *For any $\lambda \in \mathbb{R}_{>0}$ and $t \in \mathbb{R}_{>0}$, $\mathbb{E}[\mathbb{P}(T_\lambda > t|G_\lambda)^2]$ is equal to the sum of the entries of the last row of the 3 by 3 matrix e^{tA_λ}, where*

$$A_\lambda = \begin{pmatrix} -12 & 8 & 0 \\ \frac{\lambda}{2} & -6 - \frac{\lambda}{2} & 2 \\ 0 & \lambda & -2 - \lambda \end{pmatrix}. \tag{13}$$

By Mathematica, we obtain that [Louis Give precise reference to an organized Mathematica notebook, perhaps as in Diamantidis et al. (2024).]

$$\mathbb{E}[\mathbb{P}(T_\lambda > t|G_\lambda)^2] = \sum_{i=1}^{3} \frac{e^{t\frac{w_i}{2}} \left(2\lambda^2 + \lambda(56 + 3w_i) + 288 + 36w_i + w_i^2\right)}{432 + 76\lambda + 2\lambda^2 + (80 + 6\lambda)w_i + 3w_i^2}$$

where w_i are solutions to the the cubic polynomial equation in x

$$1152 + 416\lambda + 16\lambda^2 + (432 + 76\lambda + 2\lambda^2)x + (40 + 3\lambda)x^2 + x^3 = 0.$$

This explicit representation is not easy to use, so we provide the following asymptotic formula.

Proposition 5.3. *For any $t \in \mathbb{R}_{>0}$,*

$$\mathbb{E}[\mathbb{P}(T_\lambda > t|G_\lambda)^2] = \begin{cases} e^{-2t} + \lambda\dfrac{e^{-2t}}{8}\left(1 - 4t - e^{-4t}\right) + O(\lambda^2) & \text{as } \lambda \to 0 \\ e^{-4t}\left(1 + 2\lambda^{-1} + 4(3 + 16t)\lambda^{-2}\right) + O(\lambda^{-3}) & \text{as } \lambda \to \infty \end{cases}. \tag{14}$$

In particular, for any $t \in \mathbb{R}_{>0}$,

$$\mathbb{P}(T_\lambda > t|G_\lambda) \xrightarrow{d} \begin{cases} \mathbf{1}_{\{\mathrm{Exp}(2) > t\}} & \text{as } \lambda \to 0 \\ e^{-2t} & \text{as } \lambda \to \infty \end{cases} \tag{15}$$

We can also obtain convergence in L^2 distances; see Remark 5.8.

Lemma 5.2 and Proposition 5.3 will be proved in Section 5.1. Corollary 15 follows directly from Proposition 5.3 because the latter implies that $\mathbb{P}(T_\lambda > t|G_\lambda)$ converges to e^{-2t} in $L^2(\mathbb{P})$ under the probability measure for G_λ.

Lemma 5.4. *For any $\lambda \in \mathbb{R}_+$, the variance of $\mathbb{E}[T_\lambda|G_\lambda]$ is equal to*

$$\mathrm{Var}(\mathbb{E}[T_\lambda|G_\lambda]) = \mathrm{Cov}\left(T_\lambda, T'_\lambda\right) = \frac{1}{2}\frac{\lambda + 36}{\lambda^2 + 26\lambda + 72}. \tag{16}$$

Similar to (14), we can specify the asymptotic behavior of (16) as

$$\mathrm{Var}(\mathbb{E}[T_\lambda|G_\lambda]) = \begin{cases} \dfrac{1}{4} - \dfrac{1}{12}\lambda + O(\lambda^2) & \text{as } \lambda \to 0 \\ \dfrac{1}{2\lambda} + \dfrac{5}{\lambda^2} + O(\lambda^{-3}) & \text{as } \lambda \to \infty \end{cases}.$$

In the Discussion, we consider the significance of (16) for an understanding how coalescence times depend on the pedigree, represented in this case by G_λ.

Proof of Lemma 5.4. Recall from (12) that $\mathrm{Var}(\mathbb{E}[T_\lambda|G_\lambda]) = \mathrm{Cov}(T_\lambda, T'_\lambda)$, and that T_λ and T'_λ are defined only starting from state s_2. Let T_1 and T_2 be two conditionally independent coalesce times, where now we allow any starting state, s_i, for $i \in \{0,1,2\}$. Note that $\mathbb{E}[T_1] = \mathbb{E}[T_2] = 1/2$ regardless of the starting state. Let $\mathrm{Cov}_{s_i}(T_1, T_2) = \mathbb{E}_{s_i}[T_1T_2] - \mathbb{E}[T_1]^2$ be the covariance starting from state s_i. Further, let W be the (exponential) waiting time to exit the current state, which note is one of s_i for $i \in \{0,1,2\}$ with rate parameters specified in Q_λ in (21).

By conditioning on the first step out of state s_i for $i \in \{0,1,2\}$ and using the Markov property,

$$\mathbb{E}_{s_i}[T_1T_2] = \mathbb{E}_{s_i}\left[W^2\right] + \mathbb{E}_{s_i}[W]\sum_{j\neq i}\frac{q_{ij}}{|q_{ii}|}\mathbb{E}_{s_j}[T_1+T_2] + \sum_{j\neq i}\frac{q_{ij}}{|q_{ii}|}\mathbb{E}_{s_j}[T_1T_2] \tag{17}$$

in which q_{ij} are the entries of Q_λ. We write $\mathbb{E}_{s_j}[T_1+T_2]$ in (17) because T_1 and T_2 both depend on the new state s_j. In particular, if the first step is to state $s_{\Delta,1}$, then one of these is equal to zero and other has expectation $\mathbb{E}[T_1]$. If the first step is to state $s_{\Delta,2}$, then both are equal to zero.

Putting in the values of $\mathbb{E}_{s_i}\left[W^2\right]$, $\mathbb{E}_{s_i}[W]$ and q_{ij}, and simplifying gives

$$\begin{aligned}
\mathbb{E}_{s_0}[T_1,T_2] &= \frac{1}{12} + \frac{2}{3}\mathbb{E}_{s_2}[T_1,T_2]\\
\mathbb{E}_{s_1}[T_1,T_2] &= \frac{2}{\lambda+12} + \frac{\lambda}{\lambda+12}\mathbb{E}_{s_1}[T_1,T_2] + \frac{4}{\lambda+12}\mathbb{E}_{s_3}[T_1,T_2]\\
\mathbb{E}_{s_2}[T_1,T_2] &= \frac{1}{\lambda+2} + \frac{\lambda}{\lambda+2}\mathbb{E}_{s_2}[T_1,T_2]
\end{aligned}$$

whose solution gives

$$\mathrm{Cov}_{s_0}(T_1,T_2) = \frac{4}{\lambda^2+26\lambda+72} \tag{18}$$

$$\mathrm{Cov}_{s_1}(T_1,T_2) = \frac{6}{\lambda^2+26\lambda+72} \tag{19}$$

$$\mathrm{Cov}_{s_2}(T_1,T_2) = \frac{1}{2}\frac{\lambda+36}{\lambda^2+26\lambda+72}. \tag{20}$$

Since $(T_\lambda, T'_\lambda) = (T_1, T_2)$ given state s_2, (20) is precisely what is claimed in (16). □

5.1 Further characterization of the limited-outcrossing regime

In this section we prove Lemma 5.2, Proposition 5.3 and Lemma 5.4. The arguments in this section are in the spirit of Diamantidis et al. (2024) and Birkner et al. (2013), and are simpler versions of those we use in the partial-selfing regime in Section 8.

Recall from Lemma 5.1 that the random variables T_λ and T'_λ are the pairwise meeting times for two conditionally independent processes (x_λ, y_λ) and (x'_λ, y'_λ) with the same starting point, i.e. with $(x_\lambda(0),\, y_\lambda(0)) = (x'_\lambda(0),\, y'_\lambda(0))$. Each of these two processes represents the dynamics of a pair of particles on the ancestral graph G_λ. Note also that in Lemma 5.1 we required $x_\lambda \neq x'_\lambda$, meaning that T_λ and T'_λ apply to the case in which two individuals are sampled and two copies of the pairwise coalescence process are followed starting at these same two individuals.

In order to model the full two-pairs process, we need to consider all possible samples and all possible ancestral states of the samples. For simplicity, we combine states where possible by symmetry and restrict ourselves to a process with the following five states.

$$s_0 = (\bullet)\,(\bullet)\,(\bullet)\,(\bullet)$$

$$s_1 = (\bullet)\,(\bullet)\,(\bullet\,\bullet)$$

$$s_2 = (\bullet\,\bullet)\,(\bullet\,\bullet)$$

$$s_{\Delta,1} = (\bullet)\,(\bullet)\,(\bullet) \text{ or } (\bullet)\,(\bullet)\,(\bullet) \text{ or } (\bullet\,\bullet)\,(\bullet) \text{ or } (\bullet)\,(\bullet\,\bullet)$$

$$s_{\Delta,2} = (\bullet)\,(\bullet) \text{ or } (\bullet\,\bullet)$$

Here, two dots of the same color (red or blue) correspond to two particles of the same pair; say, the two red dots correspond to (x_λ, y_λ) and the two blue dots correspond to (x'_λ, y'_λ). Two dots within a parenthesis correspond to taking the same value in G_λ, i.e. being in the same individual. Hence, for $i \in \{0, 1, 2\}$, the state s_i corresponds to when i individuals or particles in G_λ contain two conditionally independent lineages.

States s_0, s_1, and s_2 represent possible configurations of the two pairs among individuals. Both member of both pairs are distinct in states s_0, s_1, and s_2 because $(\bullet\bullet) \to (\bullet)$ and $(\bullet\bullet) \to (\bullet)$ instantaneously under limited outcrossing. Note that state s_2 is the one for which T_λ and T'_λ are defined in Lemma 5.1. The composite states $s_{\Delta,1}$ and $s_{\Delta,2}$ are when one or both pairs have coalesced, respectively. The previously defined time $T_\lambda \wedge T'_\lambda$, which is when at least one of the two pairs coalesces, is the time to enter state $s_{\Delta,1}$ or $s_{\Delta,2}$ starting from state s_2. We take state $s_{\Delta,2}$ to be the absorbing state of this five-state process.

Lemma 5.5. *Let $S = (S_t)_{t\in\mathbb{R}_+}$ be the process, with state space $\{s_0, s_1, s_2, s_{\Delta,1}, s_{\Delta,2}\}$, which tracks the two pairs (x_λ, y_λ) and (x'_λ, y'_λ) at time t (backward) on the coalescence time-scale, then its rate matrix Q_λ is*

$$Q_\lambda = \begin{pmatrix} -12 & 8 & 0 & 4 & 0 \\ \frac{\lambda}{2} & -6-\frac{\lambda}{2} & 2 & 4 & 0 \\ 0 & \lambda & -2-\lambda & 0 & 2 \\ 0 & 0 & 0 & -2 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{21}$$

Proof. [Louis The proof is essentially contained in Figure 7.7 in Wakeley's book.] At each fragmentation event where we have two lineages from different pairs, they have a $\frac{1}{2}$ chance to split due to conditional independence. Therefore the transition from s_1 to s_0 is half the total fragmentation rate of the particle containing both lineages i.e. $\frac{\lambda}{2}$. Similarly, the rate at which s_2 transitions to s_1 is half the rate at which either particle fragments i.e. at rate λ.

A jump from s_2 to the absorbing state $s_{\Delta,2}$ can only happen by both particles containing two sample lineages coagulating, which occurs at rate 2. Since in this case both pairs coalesce, no jumps happen from s_2 to s_0 or $s_{\Delta,1}$. A transition from s_1 to s_2 happens when the two particles containing a single lineage coagulate, which occurs at rate 2. s_1 transitions to s_{Δ_1} by either of the particles containing a single lineages coagulating with the particle containing two lineages, which occurs at rate 4 as each pair transitions at rate 2 and there are 2 pairs. s_1 cannot transition to $s_{\Delta,2}$. s_0 cannot transition to s_2 or $s_{\Delta,2}$. s_0 transitions to s_1 if two particles containing conditionally independent lineages coagulate. There are $\frac{4\cdot 2}{2!} = 4$ such pairs of particles, so the rate from s_0 to s_1 is 8. Similarly, there are just 2 pairs of particles whose coagulation results in s_Δ, so the rate of going from s_0 to s_Δ is 4. □

Observe that the two pairs of particles (x_λ, y_λ) and (x'_λ, y'_λ) absorbed at $T_\lambda \wedge T'_\lambda$ is a continuous-time Markov chain, with state space $\{s_0, s_1, s_2, s_\Delta\}$, where we collapsed $s_{\Delta,2}$ and $s_{\Delta,2}$ into a single state called s_Δ. This absorbed process, called $\widetilde{S} = (\widetilde{S}_t)_{t\in\mathbb{R}_+}$, has 4 states and has transition rate matrix R_λ obtained from (21) by collapsing $s_{\Delta,2}$ and $s_{\Delta,2}$ into a single state s_Δ. By Lemma 5.5, the process $\widetilde{S}$ has rate matrix

$$R_\lambda = \begin{pmatrix} -12 & 8 & 0 & 4 \\ \frac{\lambda}{2} & -6-\frac{\lambda}{2} & 2 & 4 \\ 0 & \lambda & -2-\lambda & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} A_\lambda & \mathbf{v}^{\mathrm{T}} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{22}$$

where A_λ is the matrix (13) and $\mathbf{v} = \begin{pmatrix} 4 & 4 & 2 \end{pmatrix}$.

Note that

$$T_\lambda \wedge T'_\lambda = \inf\{t \in \mathbb{R}_+ : S_t \in \{s_{\Delta,1}, s_{\Delta,2}\}\} = \inf\{t \in \mathbb{R}_+ : \widetilde{S}_t = s_\Delta\} \tag{23}$$

is the first time the process $\widetilde{S}$ reaches s_Δ. The joint distribution of T_λ and T'_λ can be computed as in (Diamantidis et al., 2024, Lemma A1) but we omit this here since it is not needed.

Proof of Lemma 5.2. We first show that

$$\mathbb{E}[\mathbb{P}(T_\lambda > t|G_\lambda)^2] = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} e^{tR_\lambda} \begin{pmatrix} 1 & 1 & 1 & 0 \end{pmatrix}^T. \tag{24}$$

Recall the absorbed chain $\widetilde{S}_\lambda = (\widetilde{S}_\lambda(t))_{t\in\mathbb{R}_+}$ mentioned earlier in this section. Since $x_\lambda(0) = x'_\lambda(0) \neq y_\lambda(0) = y'_\lambda(0)$, we have that the initial condition $\widetilde{S}_\lambda(0) = s_2$. Then for any $t \in \mathbb{R}_+$, by the first equality in (12),

$$\begin{aligned} \mathbb{E}\left[\mathbb{P}(T_\lambda > t|G_\lambda)^2\right] &= \mathbb{P}(T_\lambda \wedge T'_\lambda > t) \\ &= \mathbb{P}_{s_2}(\widetilde{S}_\lambda(t) \in \{s_0, s_1, s_2\}) \\ &= (0,0,1,0)\, e^{tR_\lambda}\, (1,1,1,0)^T, \end{aligned}$$

where the second last equality follows from (23). Hence (24) is established. The right hand side of (24) is the sum of the entries of the last row of the 3 by 3 matrix e^{tA_λ}. Hence Lemma 5.2 follows. □

Lemma 5.2 will be employed to obtain asymtptotic when $\lambda \to 0$ and $\lambda \to \infty$, described in Proposition 5.3.

Proof of Proposition 5.3. We begin by establishing the rate of convergence as $\lambda \to 0$. We write

$$R_\lambda = C + D\lambda \tag{25}$$

for

$$C := \begin{pmatrix} -12 & 8 & 0 & 4 \\ 0 & -6 & 2 & 4 \\ 0 & 0 & -2 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } D := \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{26}$$

For a block matrix

$$A := \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

we have that

$$e^{tA} = \begin{pmatrix} e^{tA_{11}} & F(t) \\ 0 & e^{tA_{22}} \end{pmatrix},$$

where $F(t)$ is given by the matrix integral

$$\int_0^t e^{(t-s)A_{11}} A_{12} e^{sA_{22}} ds$$

(see equation 10.40 of Higham (2008).)

In particular,

$$\exp\left(t \begin{pmatrix} C & D \\ 0 & C \end{pmatrix}\right) = \begin{pmatrix} e^{tC} & \int_0^t e^{C(t-s)} D e^{As} ds \\ 0 & e^{tC} \end{pmatrix} \tag{27}$$

By applying the derivative in λ at $\lambda = 0$ to the Suzuki-Trotter identity (see equation 10.9 of Higham (2008)) for the matrix exponential we simultaneously obtain that

$$\left.\frac{de^{tR_\lambda}}{d\lambda}\right|_{\lambda=0} = \frac{d}{d\lambda} \lim_{N\to\infty} \left[e^{\frac{C}{N}} e^{\frac{D\lambda}{N}}\right]^N \Bigg|_{\lambda=0} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{(1-\frac{i}{N})C} D e^{\frac{i}{N}C} = \int_0^t e^{C(t-s)} D e^{As} ds. \tag{28}$$

Combining equations (28) and (27) gives

$$\left.\frac{de^{tR_\lambda}}{d\lambda}\right|_{\lambda=0} = (I, 0) \ \exp\left(t \begin{pmatrix} C & D \\ 0 & C \end{pmatrix}\right) \begin{pmatrix} 0 \\ I \end{pmatrix}.$$

By a Wolfram Research, Inc., Mathematica calculation,

$$(0,0,1,0)\,(I,0) \exp\left(t \begin{pmatrix} C & D \\ 0 & C \end{pmatrix}\right) \begin{pmatrix} 0 \\ I \end{pmatrix} (1,1,1,0)^T = \frac{e^{-2t}}{8}\left(1 - 4t - e^{-4t}\right) \tag{29}$$

The result for $\lambda \to 0$ thus follows by Taylor's theorem for matrix functions.

The proof for $\lambda \to \infty$ is done via a Wolfram Research, Inc. calculation. [Louis Give precise reference to an organized Mathematica notebook, perhaps as in Diamantidis et al. (2024).]

The proof of Proposition 5.3 is complete. □

To compute $\mathrm{Var}(\mathbb{E}[T_\lambda | G_\lambda])$, we also quantify the probability that two conditionally independent coalescence times on the same pedigree are identical.

Lemma 5.6. *For any $\lambda \in \mathbb{R}_+$,*

$$\mathbb{P}(T_\lambda = T'_\lambda) = \frac{2\lambda + 72}{\lambda^2 + 26\lambda + 72}.. \tag{30}$$

Proof. Define $\hat{S} := (\hat{S}_k)_{k\in\mathbb{Z}_+}$, where $\hat{S}_k := S(t_k)$ for all $k \in \mathbb{Z}_+$. Let T denote the first k at which $S_k \in \{s_{\Delta_1}, s_{\Delta_2}\}$. By construction of $\hat{S}$ we have that

$$\mathbb{P}(T_\lambda = T'_\lambda) = \mathbb{P}(\hat{S}_T = s_{\Delta_2}).$$

$\hat{S}_{k\wedge T}$ is a non-reversible discrete-time Markov chain whose transition matrix K can be recovered from Q 21 directly. Indeed, by definition of $\hat{S}_{k\wedge T}$ at every $k < T$ we have $\hat{S}_{k\wedge T}$ changes states. For

each of the states s_0, s_1, s_2, the corresponding rows in K are taken by setting the diagonal element in Q to zero and then renormalizing the rows to be stochastic. The states $s_{\Delta,1}$ and $s_{\Delta,2}$ for the stopped process are absorbing so their rows are zero everywhere except the diagonal, where they are 1. Combining these gives

$$K = \begin{pmatrix} 0 & \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ \frac{\lambda}{12+\lambda} & 0 & \frac{4}{12+\lambda} & \frac{8}{12+\lambda} & 0 \\ 0 & \frac{\lambda}{\lambda+2} & 0 & 0 & \frac{2}{2+\lambda} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{31}$$

The initial condition of our Markov chain is $\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \end{pmatrix}$ so the distribution P_T of the end states of $(\tilde{S}_{k\wedge T})_{k\in\mathbb{Z}_+}$ given this initial condition is

$$\lim_{n\to\infty} \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \end{pmatrix} K^n.$$

The matrix K admits an SVD decomposition, which can be calculated in Wolfram Research, Inc., and used to find that

$$P_T = \begin{pmatrix} 0 & 0 & 0 & \frac{\lambda(\lambda+24)}{\lambda^2+26\lambda+72} & \frac{2\lambda+72}{\lambda^2+26\lambda+72} \end{pmatrix}.$$

This gives the claim. □

Lemma 5.7. *For any $\lambda \in R_+$,*

$$\mathbb{E}_{s_2}[T_\lambda \wedge T'_\lambda] = \frac{1}{4} + \frac{1}{2}\frac{\lambda+36}{\lambda^2+26\lambda+72}. \tag{32}$$

Proof. We proceed by a first-step analysis argument. Recall from (23) that $T_\lambda \wedge T'_\lambda$ is the first time the process $\widetilde{S}$ from reaches s_Δ. For $i \in \{0,1,2\}$ we let $t_i := \mathbb{E}_{s_i}[T_\lambda \wedge T'_\lambda]$ be the expected time starting from state s_i. Let p_{ij} be the probability of transitioning from state s_i to state s_j. Clearly, $p_{ij} = K_{ij}$ for $i, j \in \{0,1,2\}$, where $K = (K_{ij})$ is in (31). We also let w_i be the expected holding time (the time until the next jump) of $\widetilde{S}$ starting at state s_i. Clearly, $w_i = (-R_{\lambda,ii})^{-1}$ where $R_{\lambda,ii}$ is the i-th diagonal entry of the matrix R_λ in (22). That is, $w_0 = 1/12$, $w_1 = 2/(12+\lambda)$ and $w_2 = 1/(2+\lambda)$.

The process $\widetilde{S}$ can only transition from s_0 to s_1 and s_Δ, s_1 to s_0, s_2, and s_Δ, and s_2 to s_1 and s_Δ. First-step analysis then yields a system of linear equations

$$\begin{pmatrix} 1 & -p_{01} & 0 \\ -p_{10} & 1 & -p_{12} \\ 0 & -p_{21} & 1 \end{pmatrix} \begin{pmatrix} t_0 \\ t_1 \\ t_2 \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} \tag{33}$$

Solving this equation for t_2 yields equation (32). □

We can proceed with the proof of Lemma 5.4.

Remark 5.8. There are some interesting co-incidences. By (30), (20) and the fact $\mathrm{Var}(T_\lambda) = 1/4$, we see that the correlation coefficient $\mathrm{Corr}(T_\lambda, T'_\lambda)$ is exactly the same as $\mathbb{P}(T_\lambda = T'_\lambda)$. Equivalently, $\mathrm{Cov}(T_\lambda, T'_\lambda) = \frac{1}{4}\mathbb{P}(T_\lambda = T'_\lambda)$. Furthermore, by Lemma 5.1 and Lemma 5.7, the expected squared

L^2 distance between the survival function of the conditional coalescence time and that of a rate 2 exponential random variable is

$$\mathbb{E}\left[\int_0^\infty \left(\mathbb{P}(T_\lambda > t|G_\lambda) - e^{-2t}\right)^2 dt\right] = \int_0^\infty (\mathbb{P}(T_\lambda \wedge T'_\lambda > t) - e^{-4t})dt = \frac{1}{2}\frac{\lambda+36}{\lambda^2+26\lambda+72}$$

which is also equal to the covariance $\mathrm{Cov}(T_\lambda, T'_\lambda)$.

6 Proofs for the limited-outcrossing regime

In this section, we prove our main results for the limited-outcrossing regime, in which $N(1-\alpha_N) \to \lambda \in \mathbb{R}_+$.

We first describe a discrete-time Markov process $\tilde{Z}^N$, for each integer $N > 1$, which not only keeps track of the labels of the potential ancestors of our sample, but also yields the fragmentation-coagulation process Z^N described before Definition 4.1 via projection. This process takes values in finite subsets of $[0,1]^2 \times \mathbb{Z}_+ \times \mathbb{Z}_+$ and has initial state

$$\tilde{Z}_0^N = \{((1,0), 1, \hat{X}_0), ((0,1), 2, \hat{Y}_0)\}.$$

The third component in each set tracks a new label that we will create and the fourth component corresponds to the label given by the population. For simplicity we denote by $P(k)$ the collection of possible progenitors in the past, i.e. all those elements i in I_N such that i is the fourth component of some 4-tuple in $\tilde{Z}_k^N$. A new label will be created whenever the fourth component of one of the four-tuples undergoes a split where neither parent belongs to P.

To understand the transition of this process in a single time-step, we first demonstrate its transition from time-step $k = 0$ to time-step $k = 1$. Precisely, $\tilde{Z}_1^N$ is equal to

$$\begin{cases} \{((1,0), 1, \hat{X}_0), ((0,1), 2, \hat{Y}_0)\}, & \text{if neither } \hat{X}_0, \hat{Y}_0 \text{ is } \kappa_0 \\ \{((1,0), 1, \pi_{0,1}), ((0,1), 2, \hat{Y}_0)\}, & \text{if } \kappa_0 = \hat{X}_0 \text{ self-fertilizes and } \pi_{0,1} \neq \hat{Y}_0 \\ \{((1,0), 1, \hat{X}_0), ((0,1), 2, \pi_{0,1})\}, & \text{if } \kappa_0 = \hat{Y}_0 \text{ self-fertilizes and } \pi_{0,1} \neq \hat{X}_0 \\ \{((\frac{1}{2},0), 1, \pi_{0,1}), ((\frac{1}{2},0), 3, \pi_{0,2}), ((0,1), 2, \hat{Y}_0)\}, & \text{if } \kappa_k = \hat{X}_0,\ \hat{Y}_0 \notin \{\pi_{0,1}, \pi_{0,2}\} \\ \{((1,0), 1, \hat{X}_0), ((0,\frac{1}{2}), 2, \pi_{0,1}), ((0,\frac{1}{2}), 3, \pi_{0,2})\}, & \text{if } \kappa_k = \hat{X}_0,\ \hat{Y}_0 \notin \{\pi_{0,1}, \pi_{0,2}\} \\ \{(\frac{1}{2}, 0, 1, j), ((\frac{1}{2},1), 2, \hat{Y}_0)\}, & \text{if } \kappa_k = \hat{X}_0,\ \hat{Y}_0 \in \{\pi_{0,1}, \pi_{0,2}\}, j \in \{\pi_{0,1}, \pi_{0,2}\} \setminus \{\hat{Y}_0\} \\ \{((1,\frac{1}{2}), 1, \hat{X}_0), ((0,\frac{1}{2}), 2, j)\}, & \text{if } \kappa_k = \hat{Y}_0,\ \hat{X}_0 \in \{\pi_{0,1}, \pi_{0,2}\}, j \in \{\pi_{0,1}, \pi_{0,2}\} \setminus \{\hat{X}_0\} \\ \{((1,\frac{1}{2}), 1, \hat{X}_0), ((0,\frac{1}{2}), 2, j)\}, & \text{if } \kappa_k = \hat{Y}_0,\ \hat{X}_0 \in \{\pi_{0,1}, \pi_{0,2}\}, j \in \{\pi_{0,1}, \pi_{0,2}\} \setminus \{\hat{X}_0\} \\ \{((1,1), 1, \pi_{0,1}))\}, & \text{if } \kappa_0 \in \{\hat{X}_0, \hat{Y}_0\} \text{ and } \pi_{0,1} = \pi_{0,2} \in \{\hat{X}_0, \hat{Y}_0\} \end{cases}$$

Suppose now that $\tilde{Z}_j^N$ is defined for all $j \leq k$. Write $\tilde{Z}_k^N = z = \{z_i\}_{i=1}^m$ and $\hat{z}_{r,\ldots,i_r} = z \setminus \{z_{i_n}\}_{n=1}^r$, where each $z_i = (p_i, q_i, l_i, j_i)$. Let m_k denote the smallest positive integer not contained in l_i, and $P(k) = \{j_i\}_{i=1}^m$. Then the transition from time-step k to $k+1$ is defined in such a way that $\tilde{Z}_{k+1}^N$

is equal to

$$\begin{cases} z, & \text{if } \kappa_k \notin P(k) \\ \hat{z}_r \cup \{((p_r, q_r), l_r, \pi_{k,1})\}, & \text{if } \kappa_0 = j_r, \text{ and } \pi_{k,1} = \pi_{k,2} \notin P(k) \\ \hat{z}_r \cup \{((\frac{p_r}{2}, \frac{q_r}{2}), l_r, \pi_{k,1}), (\frac{p_r}{2}, \frac{q_r}{2}), m_k, \pi_{k,2})\}, & \text{if } \kappa_0 = j_r, \pi_{k,1} \neq \pi_{k,2} \\ & \text{ and } \pi_{k,1}, \pi_{k,2} \notin P(k) \\ \hat{z}_{rs} \cup \{((p_r + \frac{p_s}{2}, q_r + \frac{q_s}{2}), l_r, j_r), ((\frac{p_s}{2}, \frac{q_s}{2}), l_s, j)\}, & \text{if } \kappa_0 = j_s, \{\pi_{k,1}, \pi_{k,2}\} \cap P(k) = \{j_r\} \\ & \text{ and } j \in \{\pi_{k,1}, \pi_{k,2}\} \setminus \{j_r\} \\ \hat{z}_{rst} \cup \{((p_r + \frac{p_t}{2}, q_r + \frac{q_t}{2}), l_r, j_r), ((p_s + \frac{p_t}{2}, q_s + \frac{q_t}{2}), l_s, j_s)\}, & \text{if } \kappa_0 = j_t, \{\pi_{k,1}, \pi_{k,2}\} \cap P(k) = \{j_r, j_s\} \\ \hat{z}_{rs} \cup \{((p_r + p_s, q_r + q_s), l_r, \pi_{k,1})\}, & \text{if } \kappa_k = j_s \text{ and } \pi_{k,1} = \pi_{k,2} = j_r \in P(k) \end{cases}$$

where we assume $r < s < t$. The transition rates can be calculated as

$$\tilde{Z}^N_{k+1} = \begin{cases} z, & \text{with probability } \frac{N-m}{N} \\ \hat{z}_r \cup \{((p_r, q_r), l_r, \pi_{k,1})\}, & \text{with probability } \alpha_N \frac{1}{N} \frac{N-m-1}{N-1} \\ \hat{z}_r \cup \{((\frac{p_r}{2}, \frac{q_r}{2}), l_r, \pi_{k,1}), (\frac{p_r}{2}, \frac{q_r}{2}), m_k + 1, \pi_{k,2})\}, & \text{with probability } (1-\alpha_N) \frac{1}{N} \frac{N-m-1}{N-1} \frac{N-m-2}{N-2} \\ \hat{z}_{rs} \cup \{((p_r + \frac{p_s}{2}, q_r + \frac{q_s}{2}), l_r, j_r), ((\frac{p_s}{2}, \frac{q_s}{2}), l_s, j)\}, & \text{with probability } (1-\alpha_N) \frac{2}{N} \frac{1}{N-1} \frac{N-m-2}{N-2} \\ \hat{z}_{rst} \cup \{((p_r + \frac{p_t}{2}, q_r + \frac{q_t}{2}), l_r, j_r), ((p_s + \frac{p_t}{2}, q_s + \frac{q_t}{2}), l_s, j_s)\}, & \text{with probability } (1-\alpha_N) \frac{1}{N} \frac{2}{(N-1)(N-2)} \\ \hat{z}_{rs} \cup \{((p_r + p_s, q_r + q_s), l_r, \pi_{k,1})\}, & \text{with probability } \alpha_N \frac{1}{N} \frac{1}{N-1} \end{cases}$$

We can now define the process Z^N from $\tilde{Z}^N$. Denote by $\pi : [0,1]^2 \times \mathbb{Z}_+ \times \mathbb{Z}_+ \to [0,1]^2 \times \mathbb{Z}_+$ the projection to the first three components (i.e. ignoring the last component). Then we define $Z^N = (\pi(\tilde{Z}^N_k))_{k \in \mathbb{Z}_+}$. Z^N is simply $\tilde{Z}^N$ where we forget the labels at the population level. Note that, as $\tilde{Z}^N$ is, conditional on $\mathcal{A}_N$, deterministic, so is Z^N.

Note that Z^N is in distribution the same as the discrete-time Markov process with state space $\mathcal{P}_{m,2}$ defined in (9). $\mathcal{P}_{m,2}$ is a Polish space. We shall first describe the transition of Z^N, and then prove weak convergence using standard theory of Markov processes.

$$Z^N_1 = \begin{cases} \{((1,0),1),((0,1),2)\} & \text{neither } \hat{X}_0 \text{ nor } \hat{Y}_0 \text{ is a offspring} \\ \{((\frac{1}{2},0),1),((\frac{1}{2},0),3)((0,1),2)\} & \hat{X}_0 \text{ splits without overlap with } \hat{Y}_0 \\ \{((1,0),1)((0,\frac{1}{2}),2),((0,\frac{1}{2}),3)\} & \hat{Y}_0 \text{ splits without overlap with } \hat{X}_0 \\ \{((\frac{1}{2},0),1),((\frac{1}{2},1),2)\} & \hat{X}_0 \text{ splits and one parent is } \hat{Y}_0 \\ \{((1,\frac{1}{2}),1),((0,\frac{1}{2}),2)\} & \hat{Y}_0 \text{ splits and one parent is } \hat{X}_0 \\ \{((1,1),1)\}, & \text{selfing event between } \hat{X}_0 \text{ and } \hat{Y}_0 \end{cases}$$

Suppose $Z^N_k = z = \{(p_i, l_i)\}_{i=1}^m$ for $p_i \in [0,1]^2$ and $l_i \in \mathbb{Z}_+$. Let m_k denote the smallest positive integer not contained in $\{l_i\}$. Recall $P(k) = \{j_i\}_{1 \le i \le m}$ denotes the corresponding set of population-level labels such that l_i is the label of the individual to whom the the label l_i (i.e. such that $(p_i, l_i, j_i) \in \tilde{Z}^N_k$). By $\hat{z}_{rs\ldots i_j}$ we denote the set $z \setminus [z_{i_l} : 1 \le l \le j]$. The one-step dynamics are described as

$$Z^N_{k+1} = \begin{cases} z, & \text{if } \kappa_k \notin l, \text{ or } \{\kappa_k, \pi_{k,1}, \pi_{k,2}\} \cap P(k) = \{\kappa_k\}, \pi_{k,1} = \pi_{k,2} \\ \{(p_r + p_s, l_r)\} \cup \hat{z}_{rs}, & \text{if } \kappa_k = j_s, \pi_{k,1} = \pi_{k,2} = j_r \\ \{(p_r + \frac{1}{2}p_t, l_r), ((p_s + \frac{1}{2}p_t), l_s)\} \cup \hat{z}_{rst}, & \text{if } l_t = \kappa_k, P(k) \cap \{\pi_{k,2}, \pi_{k,1}\} = \{l_s, l_t\} \\ \{(p_r + \frac{1}{2}p_s, l_r), ((\frac{1}{2}p_s), l_s)\} \cup \hat{z}_{rs}, & \text{if } l_s = \kappa_k, P(k) \cap \{\pi_{k,1}, \pi_{k,2}\} = \{l_r\} \\ \{(\frac{1}{2}p_r, l_r), (\frac{1}{2}p_r, m_k)\} \cup \hat{z}_r, & \text{if } l_r = \kappa_k, P(k) \cap \{\pi_{k,1}, \pi_{k,2}\} = \emptyset. \end{cases}$$

for any $r < s < t$.

The one-step transition probabilities $\{\mathbb{P}(Z_{k+1}^N = \cdot | Z_k^N = q)\}$ can thus be calculated as follows:

$$z \to \begin{cases} z, & \text{with probability } \frac{N-m}{N} + \alpha_N \frac{m}{N} \frac{N-m-1}{N-1} \\ \{(p_r + p_s, l_r)\} \cup \hat{z}_{rs}, & \text{with probability } \alpha_N \frac{2}{N} \frac{1}{N-1} \\ \{(p_r + \frac{1}{2}p_t, l_r), ((p_s + \frac{1}{2}p_t), l_s)\} \cup \hat{z}_{rst}, & \text{with probability } (1-\alpha_N) \frac{2}{N(N-1)(N-2)} \\ \{(p_r + \frac{1}{2}p_s, l_r), ((\frac{1}{2}p_s), l_s)\} \cup \hat{z}_{rs}, & \text{with probability } (1-\alpha_N) \frac{2}{N(N-1)} \frac{N-m-2}{N-2} \\ \{(\frac{1}{2}p_r, l_r), (\frac{1}{2}p_r, m_k)\} \cup \hat{z}_r, & \text{with probability } (1-\alpha_N) \frac{1}{N} \frac{N-m-1}{N-1} \frac{n-M-2}{n-2} \end{cases} \tag{34}$$

for each fixed r and $s \neq r, s \neq t, t \neq r$.

In the critical regime, these transitions are

$$z \to \begin{cases} z, & \text{with probability } 1 - O(N^{-2}) \\ \{(p_r + p_s, l_r)\} \cup \hat{z}_{rs}, & \text{with probability } 2N^{-2} + O(N^{-3}) \\ \{(p_r + \frac{1}{2}p_t, l_r), ((p_s + \frac{1}{2}p_t), l_s)\} \cup \hat{z}_{rst}, & \text{with probability } o(N^{-3}) \\ \{(p_r + \frac{1}{2}p_s, l_r), ((\frac{1}{2}p_s), l_s)\} \cup \hat{z}_{rs}, & \text{with probability } o(N^{-3}) \\ \{(\frac{1}{2}p_r, l_r), (\frac{1}{2}p_r, m_k)\} \cup \hat{z}_r, & \text{with probability } \lambda N^{-2} + o(N^{-2}) \end{cases}$$

Remark 6.1. *The labels of Z_k^N, for any $k \in \mathbb{Z}_+$, are contained in $\{1, \ldots, \sup_{i \leq k} |Z_i^N|\}$.*

We let G_k^N be the projection of the set Z_k^N to its labels in $\mathbb{Z}_+$. From the above calculations, $(G_k^N)_{k \in \mathbb{Z}_+}$ is itself a Markov process with one-step transition probabilities

$$l \to \begin{cases} \hat{l}_r \cup \{m_k\} & \text{with probability} \quad \lambda |G_k^N| N^{-2} + o(N^{-2}) \\ l & \text{with probability} \quad 1 - O(N^{-2}) \\ \hat{l}_r & \text{with probability} \quad |G_k^N|(|G_k^N| - 1)N^{-2} + O(N^{-2}) \end{cases} \tag{35}$$

for some r. G_k^N takes values in finite subsets of $\mathbb{Z}_+$ $\Lambda = \bigsqcup_{m=1}^{\infty} \{l_i\}_{i=1}^m$ for $l_i \in \mathbb{Z}_+$, which is a Polish space under the metric $d(l, m) = |l \cap m|$. Importantly, this metric engenders the same topology as by the projection map from $\mathcal{P}_{m,2}$. We denote the time-rescaling of $(G_k^N)_{k \in \mathbb{Z}_+}$ by $G^N := (G_{\lfloor tN^2 \rfloor}^N)_{t \in \mathbb{R}_+}$.

With the above precise description of the process G^N, for each $N \geq 2$, we proceed with a proof of Lemma 4.2.

Proof of Lemma 4.2. We first prove the weak convergence of the processes $\{G^N\}_{N \in \mathbb{N}}$ (i.e. ignoring the masses). Precisely, we shall show that if $N(1 - \alpha_N) \to \lambda \in \mathbb{R}_+$, then $(G_{\lfloor tN^2 \rfloor}^N)_{t \in \mathbb{R}_+}$ converges in distribution under $\mathbb{P}_{\text{diff}}(\cdot | \mathcal{A}_N)$ in $\mathcal{D}(\mathbb{R}_+, \Lambda)$ to a continuous-time Markov process $G_\lambda = (G_\lambda(t))_{t \geq 0}$ with initial distribution $G_\lambda(0) = \{1, 2\}$ and transition rates

$$l \to \begin{cases} l \cup \{m(l)\} & \text{with rate} \quad \lambda |l| \\ \hat{l}_r & \text{with rate} \quad |l|(|l| - 1) \end{cases}, \tag{36}$$

where $m(l)$ denotes the smallest positive integer not contained in l.

Let $T^{(N)}$ be the linear operator on the space $\mathcal{C}_b(\Lambda)$ of bounded continuous functions on Λ defined by $T^{(N)} f(n) = \mathbb{E}\left[f(G_1^N) \,|\, G_0^N = l\right]$. The generator $\mathcal{L}^N$ of the discrete-time process G^N is given by

$$\mathcal{L}^N f(n) := \mathbb{E}\left[f(G_1^N) - f(G_0^N) \,|\, G_0^N = l] = (T^{(N)} - I) f(l)\right] \tag{37}$$

which can be explicitly computed using (35). Let $\mathcal{L}$ be the infinitesimal generator of the continuous-time process $G_\lambda = (G_\lambda(t))_{t\geq 0}$. That is, by (36),

$$\mathcal{L}f(l) = \lambda|l|[f(l \cup \{m(l)\}) - f(l)] + \sum_{r=1}^{|l|} f(\hat{l}_r)r(r-1)]. \tag{38}$$

It follows from (35) that for all $f \in \mathcal{C}_b(\Lambda)$ with finite support,

$$\sup_{l\in\Lambda} |N^2\mathcal{L}^N f(l) - \mathcal{L}f(l)| \to 0 \quad \text{as } N \to \infty. \tag{39}$$

Let $\{T(t)\}_{t\in\mathbb{R}_+}$ be the semigroup on $\mathcal{C}_b(\Lambda)$ of the continuous-time process G_λ. By Theorem 6.5 of Ethier and Kurtz (2009, chapter 1) and (39), it holds that

$$\lim_N \sup_{0\leq t\leq t_0} \sup_{l\in\Lambda} |\left(T^{(N)}\right)^{\lfloor tN^2 \rfloor} f(l) - T(t)f(l)| = 0 \tag{40}$$

for all f in the domain of $\mathcal{L}$.

Next, since the fragmentation rate is linear while the coagulation rate is quadratic, we can show as in Griffiths (1991, Theorem 3) that the compact containment condition (Ethier and Kurtz (2009, (7.9), page 129)) holds. That is, for any $\varepsilon \in (0,1)$ and $T \in (0,\infty)$, there is a finite deterministic constant $K = K_{\varepsilon,T}$ such that

$$\limsup_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} |G^N_{[tN^2]}| \geq K\right) \leq \varepsilon. \tag{41}$$

This suffices for the compact containment condition of G^N as we recall from remark 6.1 that G^N_k is contained in

$$\left\{1, 2, \ldots, \sup_{0\leq i\leq k} |G^N_i|\right\}$$

for all $k \in \mathbb{Z}_+$, and is clearly equivalent to a compact containment condition for the process $L^N := (|G^N_{\lfloor tN^2 \rfloor}|)_{t\in\mathbb{R}_+}$ that counts the number of particles at each cross-section.

We now demonstrate how to prove (41) by a martingale argument by examining the continuous-time Markovian process $L := |G_\lambda|$ taking values in $\mathbb{Z}_+$ capturing the size of the cross-section of G_λ at any time-point. Let $\mathcal{Q}$ denote the generator of L. We can see that, for any f in $\mathcal{C}_b(\mathbb{Z}_+)$ that

$$\mathcal{Q}f(n) = \lambda n(f(n+1) - f(n)) + n(n-1)(f(n-1) - f(n)) \tag{42}$$

First, we invoke a general relation between Markov processes and martingales. For any $f \in \mathcal{C}_b(\mathbb{Z}_+)$,

$$M(t) := f(L_t) - f(L_0) - \int_0^t \mathcal{Q}f(L_s)\, ds \tag{43}$$

is a martingale with quadratic variation $\langle M\rangle_t = \int_0^t \mathcal{Q}(f^2)(L_s) - 2f(L_s)\mathcal{Q}f(L_s)\, ds$; see Kipnis and Landim (1998, Lemma 5.1) or Ethier and Kurtz (2009, Proposition 4.1.7). A truncation argument will enable us to take f to be the identity function (i.e. when $f(n) = n$ for all $n \in \mathbb{Z}_+$) to obtain from (38) that

$$L_t = L_0 + \int_0^t -L_s^2 + (\lambda+1)L_s\, ds + M^{\mathrm{Id}}(t) \quad \text{for } t \in \mathbb{R}_+, \tag{44}$$

where M^{Id} is a martingale with quadratic variation $\langle M^{\mathrm{Id}}\rangle_t = \int_0^t L_s^2 + (\lambda-1)L_s\,ds$. In the above, we used the fact that, from (38), when f is the identity function,

$$\begin{aligned}
\mathcal{Q}f(n) =& \lambda n - n(n-1) = -n^2 + (\lambda+1)n \leq \frac{(\lambda+1)^2}{4} && (45)\\
\mathcal{Q}f^2(n) =& \lambda n[(n+1)^2 - n^2] + n(n-1)[(n-1)^2 - n^2]\\
=& -2n^3 + (3+2\lambda)n^2 + (\lambda-1)n && (46)\\
\mathcal{Q}f^2(n) - 2f(n)\mathcal{Q}f(n) =& n^2 + (\lambda-1)n
\end{aligned}$$

Therefore, from (44) and (45),

$$\sup_{t\in[0,T]} L_t \leq L_0 + \frac{(\lambda+1)^2}{4}\,T + \sup_{t\in[0,T]} M^{\mathrm{Id}}(t).$$

By Doob's maximal inequality, the above formula for $\langle M^{\mathrm{Id}}\rangle_t$, and the fact that $\sup_{t\in[0,T]}\mathbb{E}[L_t^2] < \infty$, which follows from (46), there exists a constant $C_T \in (0,\infty)$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]} M^{\mathrm{Id}}(t) \geq K\right) \leq \frac{C_T}{K^2} \quad \text{for all } K \in (0,\infty).$$

By the last two displayed inequalities, for any $\varepsilon \in (0,1)$ there exists a constant $K_{\varepsilon,T}$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]} L_t \geq K_{\varepsilon,T}\right) \leq \varepsilon. \tag{47}$$

Finally, (41) can be obtained by exactly the same argument above, where $L_t = |L_t|$ and $\int_0^t \mathcal{Q}f(L_s)\,ds$ are replaced, respectively, by $L_k^N := |G_k^N|$ and $\sum_{i=0}^{k-1}\mathcal{Q}^N f(L_i^N)$ for $\mathcal{Q}^N$ the generator of $L^N = (L_k^N)_{k\in\mathbb{Z}_+}$.

Equipped with (41), it then follows from Corollary 8.9 of Ethier and Kurtz (2009, Chapter 4) and (40) that G_λ converges in distribution in $\mathcal{D}(\mathbb{R}_+,\Lambda)$ to the process $G_\lambda = (G_\lambda(t))_{t\geq 0}$.

The proof for the process $\{Z^N\}$ follows from the same argument as the above, using (34) instead of (35). □

Next, we couple random walks on $(Z_k^N)_{k\in\mathbb{Z}_+}$ with those on $\mathcal{G}_N$ to ensure that their limits conditional on the pedigree are identical. Using an argument like that of Birkner et al. (2013) and Diamantidis et al. (2024), we show that, as λ goes to infinity, the limits conditional on the pedigree converge to the unconditional limit. This proof technique is further elaborated in Section 8 to establish the limit conditional on the pedigree for the partial-selfing regime.

6.1 Potential ancestors of two sample lineages under limited outcrossing

To model the ancestry of our two sample lineages, we use a branching-coalescing-overlapping graph which differs from the ancestral recombination graph (Griffiths, 1991; Griffiths and Marjoram, 1997; Hudson, 1983b) and the ancestral selection graph Krone and Neuhauser (1997); Neuhauser and Krone (1997) only in how ancestral genetic material is traced back in time through the graph.

The process $Z^N = (Z_k^N)_{k\in\mathbb{Z}_+}$, as we have already seen, can be viewed as a graphically as having two initial nodes $(1,0)$ and $(0,1)$ under the sampling scheme $\mathbb{P}_{\mathrm{diff}}$. These two nodes correspond to

the masses of the two lineages in the two individuals from whom we sampled the two lineages at time-step 0, namely $\hat{X}_0$ and $\hat{Y}_0$. We establish a coupling between our pair of random walks on the pedigree and some new pair of random walks on Z^N using this correspondence.

Let $x_0^N = (1,0)$ and $y_0^N = (0,1)$. Suppose that x_j^N is defined for all $j \le k$, for induction. If $\hat{X}_k$ undergoes an outcrossing at time-step k, then we define x_k^N to follow the path of the outcrossing at Z_k^N that corresponds to the edge along which $\hat{X}_k^N$ travels. y_k^N can be continuously defined in the same way. This generates two discrete-time random walks (x^N, y^N) on Z^N starting at $((1,0),(0,1))$ and satisfying

(i) $\{x_k^N, y_k^N\} \subset Z_k^N$ for all $k \in \mathbb{Z}_+$,

(ii) at any fragmentation event, each of $(x_k^N)_{k\in\mathbb{Z}_+}$ and $(y_k^N)_{k\in\mathbb{Z}_+}$ will follow each of the two paths available with equal (i.e. 1/2) probability.

The proof of Lemma 4.2 can readily be extended to obtain the following joint convergence.

Lemma 6.2. *Suppose* $\lim_{N\to\infty} N(1-\alpha_N) = \lambda \in \mathbb{R}_+$. *Then* $(Z^N_{\lfloor tN^2\rfloor}, x^N_{\lfloor tN^2\rfloor}, y^N_{\lfloor tN^2\rfloor})_{t\in\mathbb{R}_+}$ *converges in distribution in* $\mathcal{D}(\mathbb{R}_+, \mathcal{P}_{m,2}\times[0,1]^2\times[0,1]^2)$ *under* $\mathbb{P}_{\text{diff}}(\cdot|\mathcal{A}_N)$ *to a process* $(Z_\lambda, x_\lambda, y_\lambda)$, *where* $Z_\lambda = (Z_\lambda(t))_{t\in\mathbb{R}_+}$ *is a fragmentation-coagulation process of two immiscible unit masses with rate* λ. *The joint process* $(x_\lambda, y_\lambda) = (x_\lambda(t), y_\lambda(t))_{t\in\mathbb{R}_+}$ *is a continuous-time* $Z_\lambda \times Z_\lambda$*-valued process with*

(i) $\{x_\lambda(t), y_\lambda(t)\} \subset Z_\lambda(t)$ *for all* $t \in \mathbb{R}_+$,

(ii) at any fragmentation event, each of $(x_\lambda(t))_{t\in\mathbb{R}_+}$ *and* $(y_\lambda(t))_{t\in\mathbb{R}_+}$ *will follow each of the two paths available with equal (i.e. 1/2) probability.*

We omit the proof of Lemma 6.2 since it follows from the same standard argument to that of Lemma 4.2.

6.2 Conditional limit in the limited-outcrossing regime for distinct individuals

Here, in the continuous-time limit, coalescence between overlapping particles is instantaneous, while this is in general not true in the discrete-time model. We first show that this behavior is true in the limit for our random walks on the pedigree by the following lemma.

Lemma 6.3. *Suppose* $\lim_{N\to\infty} N(1-\alpha_N) = \lambda \in \mathbb{R}_+$. *Let* τ_O^N *be the time of the first overlap of our two sample lineages*

$$\inf\{k \in \mathbb{Z}_+ : \hat{X}_k = \hat{Y}_k\}.$$

Then $\tau^{(N,2)} - \tau_O^N$ *converges to* 0 *in distribution as* N *goes to infinity under* $\mathbb{P}_{\text{diff}}(\cdot|\mathcal{A}_N)$.

Proof. Fix $\varepsilon > 0$. Observe that

$$\mathbb{P}_{\text{diff}}(N^{-2}|\tau^{(N,2)} - \tau_O^N| > \varepsilon|\mathcal{A}_N) \le \frac{1}{\varepsilon}\mathbb{P}_{\text{diff}}(N^{-2}|\tau^{(N,2)} - \tau_O^N| > \varepsilon)$$

by the conditional Markov inequality. The upper bound established here goes to zero as the probability that there is only a single overlap before coalescence is $\frac{1}{2-\alpha_N}$, which converges to one, by Lemma 10.1 and the time it takes between coalescence and that last overlap has mean going to zero by Lemma 10.2. □

We can represent the resulting process Z of Lemma 4.2 as a graph starting with two nodes whereby each node fragments at rate λ and each pair of distinct nodes coagulates at rate 2. We can understand the underlying graph structure G_λ as a limit of a sample of two lineages on $\mathcal{G}_N$ by the following theorem. This may be visually interpreted as in Figure 7.

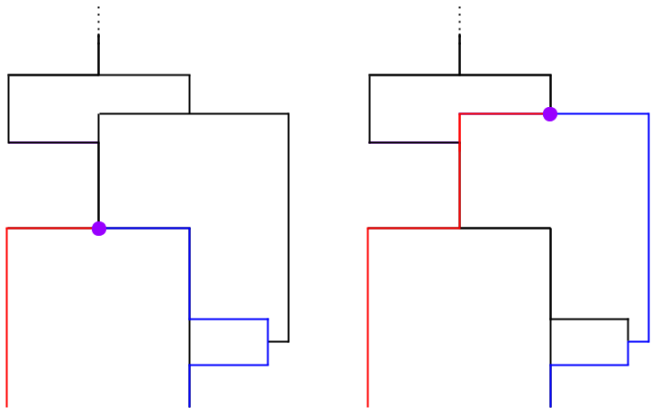


Figure 7: Two different realizations, one on the left and the other on right, of the pair (x_λ, y_λ) of random walk paths on the same ancestral graph G_λ shown before in Figure 6. In each realization of the paths, x_λ is in blue and y_λ in red. At each fragmentation of the graph, where a particle splits into two, each walk chooses which particle path to follow fairly i.e. with probability $\frac{1}{2}$ each. The thick purple dot represents the first meeting time T_λ of the two paths x_λ and y_λ. For these two realizations, $T_\lambda = t_1$ on the left and $T_\lambda = t_2$ on the right, where $\{t_1, t_2, t_3\}$ are shown in Figure 6,

[Max Tentative addition.]

Lemma 6.4. *For any finite collection of Borel subsets A_i of $\mathbb{R}$ for $1 \le i \le k$ the tuple*

$$\left(\mathbb{P}_{\text{diff}}\left(N^{-2}\tau^{N,2} \in A_i | \mathcal{A}_N\right)\right)_{1\le i\le k}$$

converges in distribution to

$$\left(\mathbb{P}\left(T_\lambda \in A_i | G_\lambda\right)\right)_{1\le i\le k}. \tag{48}$$

Proof. By linearity of probability and continuity of addition we can take, without loss of generality, that the A_i are disjoint. Further, as open intervals form a basis for $\mathbb{R}$, we can in fact take the A_i to be disjoint intervals.

Let x_k^N, y_k^N denote the projection of $\hat{X}_k$ and $\hat{Y}_k$ to G^N, which follow each path of each outcrossing with equal likelihood. Let T_λ^N denote the pairwise coalescence time of x_k^N and y_k^N on G_k^N i.e.

$$T_\lambda^N := \inf\{k \in \mathbb{Z}_+ : \quad x_k^N = y_k^N\}.$$

Further, let

$$\tau_O^N := \inf\{k \in \mathbb{Z}_+ : \quad \hat{X}_k = \hat{Y}_k\}$$

be the first time to overlap for the random walks on the pedigree.

By construction, $\tau_O^N = T_\lambda^N$ pointwise and knowing the pedigree tells exactly as much information about τ_O^N as knowing G_λ. In particular,

$$\left(\mathbb{P}_{\text{diff}}\left(N^{-2}\tau_O^N \in A_i \mid \mathcal{A}_N\right)\right)_{1\le i\le k} = \left(\mathbb{P}\left(N^{-2}T_\lambda^N \in A_i\right)\right)_{1\le i\le k}. \tag{49}$$

It follows from Lemma 6.2 that the right hands side of equation 49 converses as N goes to infinity to 48.

Indeed, fix a graph $\tilde{G}$ in $\mathcal{D}(\mathbb{R}_+, \Lambda)$ with particles trajectories $\tilde{x}$ and $\tilde{y}$. Define, as in 10, the hitting time $\tilde{T}_\lambda$ of these particles as

$$\tilde{T}_\lambda := \inf\{t \geq 0 : \quad \tilde{x}(t) = \tilde{y}(t)\}.$$

For each A_i, the functional Ψ_i defined by

$$\Psi_i(\tilde{G}) := \mathbb{P}(\tilde{T}_\lambda \in A_i | \tilde{G})$$

is a continuous functional of the graph $\tilde{G}$. Ψ_i depends only on the discrete structure of the graph $\tilde{G}$ and its edge lengths up to time t. By proposition 5.3 of Ethier and Kurtz (2009, Chapter 3), for any sequence of graphs $\tilde{G}^N$ converging to $\tilde{G}$ in $\mathcal{D}(\mathbb{R}_+, \Lambda)$ the discrete structure of $\tilde{G}^N$ is eventually fixed for large enough N on any compact interval K of which A_i is a proper subspace. Further, on K the edge lengths converge uniformly in N. This demonstrates continuity of Ψ_i. Continuity in $\mathbb{R}^k$ is equivalent to continuity in each component so the right hands side of equation 49 converses as N goes to infinity to 48.

By definition of Ψ_i, we have $\Psi_i(G^N) = \mathbb{P}(\tilde{T}^N_\lambda | G^N)$ where

$$\tilde{T}^N_\lambda := \inf\{t \geq 0 x^N_{\lfloor tN^2 \rfloor} = y^N_{\lfloor tN^2 \rfloor}\}.$$

Note that $\tilde{T}^N_\lambda \geq N^{-2} T^N_\lambda$ for all N and their difference $\tilde{T}^N_\lambda - N^{-2} T^N_\lambda$ is pointwise bounded above by N^{-2}. Thus their distributional limits conditional on the pedigree are the same. Therefore

$$\left(\left(\mathbb{P}_{\text{diff}}(N^{-2}\tau^N_O \in A_i | \mathcal{A}_N)\right)\right)_{1 \leq i \leq k}$$

converges as N goes to infinity to 48.

The result follows from the fact that the limits conditional on the pedigree of τ^N_O and $\tau^{N,2}$ under $\mathbb{P}_{\text{diff}}(\cdot | \mathcal{A}_N)$ must be the same by Lemma 6.3. □

Theorem 6.5. *The sequence of measures ξ_N defined by*

$$\xi_N(\cdot) := \mathbb{P}_{\text{diff}}(N^{-2}\tau^{N,2} \in \cdot | \mathcal{A}_N)$$

converges weakly in distribution to

$$\xi = \mathbb{P}(\cdot | G_\lambda).$$

Proof. By Theorem 4.2 of Kallenberg (2017), weak convergence in the vague topology follows by demonstrating that, for any fixed f in $C_c(\mathbb{R})$,

$$\xi_N f = \int_{\mathbb{R}} f(x) d\xi_N(x)$$

converges in distribution to

$$\xi f = \int_{\mathbb{R}} f(x) d\xi(x).$$

However, this is equivalent to the finite dimensional convergence

$$(\zeta_N B_i)_{1 \leq i \leq n} \xrightarrow{d} (\zeta B_i)_{1 \leq i \leq n}$$

for any ζ-continuity sets B_i (see pp.109-110 in Kallenberg (2017)). This follows immediately from Lemma 6.4.

As each ζ_N is a probability measure, it follows by Theorem 4.18 in Kallenberg (2017) that in fact the sequence of ζ_N converge weakly in distribution.

□

[Max End of tentative addition.]

We are now prepared to prove Theorem 4.3 in the limited-outcrossing regime.

Theorem 6.6. *Let x_λ and y_λ be independent random walks on G_λ and $T_\lambda := \inf\{t \in \mathbb{R}_+ : x_\lambda(t) = y_\lambda(t)\}$ be their first meeting time defined in (10). Suppose $N(1-\alpha_N) \to \lambda \in \mathbb{R}_{>0}$. Then, for any fixed $t > 0$,*

$$\mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N) \xrightarrow{d} \mathbb{P}_{x_0 \neq y_0}(T_\lambda > t|G_\lambda).$$

Remark 6.7. *Note that this implies exactly the statement of Theorem 4.3 in the case of limited outcrossing.*

Proof. Let x_k^N and y_k^N denote the projection of $\hat{X}_k$ and $\hat{Y}_k$ to Z_k^N. Without fear of confusion, we can forget their masses and view them as walks on G^N which follow each path of each outcrossing with equal likelihood. Let T_λ^N denote the pairwise coalescence time of x_k^N and y_k^N on G_k^N i.e.

$$T_\lambda^N := \inf\{k \in \mathbb{Z}_+ : x_k^N = y_k^N\}.$$

Further, let

$$\tau_O^N := \inf\{k \in \mathbb{Z}_+ : \hat{X}_k = \hat{Y}_k\}$$

be the first time to overlap for the random walks on the pedigree.

By construction $\tau_O^N = T_\lambda^N$ pointwise and knowing the pedigree tells exactly as much information about τ_O^N as knowing G_λ. In particular,

$$\mathbb{P}_{\text{diff}}(N^{-2}\tau_O^N > t|\mathcal{A}_N) = \mathbb{P}_{x_0^N \neq y_0^N}(N^{-2}T_\lambda^N > t|G^N). \tag{50}$$

It follows from Lemma 6.2 that the right hand side of (50) converges as N goes to infinity to $\mathbb{P}(T_\lambda > t|G_\lambda)$.

Indeed, for any graph $\tilde{G}$ in $\mathcal{D}(\mathbb{R}_+, \Lambda)$ with initial particles $\tilde{x}, \tilde{y}$, we can define $\tilde{T}$, as in (10),

$$\tilde{T} := \inf\{t \geq 0 : \tilde{x}(t) = \tilde{y}(t)\}.$$

The functional Ψ, defined by

$$\Psi(\tilde{G}) := \mathbb{P}(\tilde{T} > t \mid \tilde{G}),$$

is a continuous functional of the graph $\tilde{G}$. Ψ depends only on the discrete structure of the graph $\tilde{G}$ and its edge lengths up to time t. By lemma 6.2 the convergence of G^N to G_λ in $\mathcal{D}(\mathbb{R}_+, \Lambda)$ and Proposition 5.3 of Ethier and Kurtz (2009, Chapter 3), the discrete structure of G^N is eventually fixed for large enough N on any compact interval K of which $[0, t]$ is a proper subspace. Further, on K the edge lengths converge uniformly in N.

From the definition of Ψ, we have $\Psi(G^N) = \mathbb{P}(\tilde{T}_\lambda^N > t \mid G^N)$, where $\tilde{T}_\lambda^N$ is the continuous-time analogue of T_λ^N; i.e.

$$\tilde{T}_\lambda^N := \inf\{t \geq 0 : x^N_{\lfloor tN^2 \rfloor} = y^N_{\lfloor tN^2 \rfloor}\}.$$

Note that $\tilde{T}^N_\lambda \geq N^{-2}T^N_\lambda$ for all N and their difference $\tilde{T}^N_\lambda - N^{-2}T^N_\lambda$ is pointwise bounded by $\frac{1}{N^2}$. Thus, their limits conditional on the pedigree are the same. Therefore $\mathbb{P}(N^{-2}\tau^N_O > t|\mathcal{A}_N)$ converges as N goes to infinity to $\mathbb{P}(T_\lambda > t|G_\lambda)$.

The result follows from the fact that the limits conditional on the pedigree of τ^N_O and $\tau^{(N,2)}$ under $\mathbb{P}_{\text{diff}}(\cdot \mid \mathcal{A}_N)$ must be the same by Lemma 6.3. □

6.3 Conditional limit under $\mathbb{P}_{\text{same}}$

Selfing is ubiquitous in the limited-outcrossing regime. In particular, any sample of two lineages in the same individual will undergo very many selfing reproduction events before they have the chance to undergo a split. This is reflected in Lemma 6.8.

Lemma 6.8. *If $N(1-\alpha_N) \to \lambda \in \mathbb{R}_+$, then $\tau^{(N,2)}$ converges under $\mathbb{P}_{\text{same}}(\cdot|\mathcal{A}_N)$ to 0 in distribution.*

Proof. Fix $\varepsilon > 0$. By the conditional Markov inequality we have

$$\mathbb{P}(\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N) > \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t). \tag{51}$$

The limit conditional on the pedigree therefore follows from the unconditional limit established in Theorem 10.3 as $\mathbb{P}_{\text{same}}(\mathcal{O} = 0) = \frac{\alpha_n}{2-\alpha_N}$, which converges to 1, and under the law of $\mathbb{P}_{\text{same}}(\cdot|\mathcal{O} = 1)$ $\tau^{(N,2)}$ converges in distribution to 0 by Theorem 3.1. □

7 Proofs for the negligible-outcrossing regime

The proofs for the case of negligible outcrossing are straightforward. They follow directly from the results in the limited-outcrossing regime.

Corollary 7.1. *If $N(1-\alpha_N) \to 0$, then $\mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N)$ converges in distribution as N goes to infinity to a Heaviside function with its jump time exponentially distributed with rate 2.*

Proof. By Theorem 6.6, $\mathbb{P}_{\text{diff}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N)$ converges as N goes to infinity to

$$\mathbb{P}_{x_0 \neq y_0}(T_0 > t|Z_0).$$

Z_0 branches at rate 0 and coagulates the two existing nodes at rate 2. In particular, conditional on Z_0, T_0 is known exactly, and it distributed unconditionally as an exponential random variable with rate 2. Therefore

$$\mathbb{P}(\mathbb{P}_{x_0 \neq y_0}(T_0 > t|Z_0) = 1) = e^{-2t} = 1 - \mathbb{P}(\mathbb{P}_{x_0 \neq y_0}(T_0 > t|Z_0) = 0)$$

This is exactly the claim. □

Corollary 7.2. *If $N(1-\alpha_N) \to 0$, then $\tau^{(N,2)}$ converges in distribution to 0 under $\mathbb{P}_{\text{same}}(\cdot\,|\mathcal{A}_N)$.*

The proof of this corollary is omitted as it is contained in the statement of the Lemma 6.8.

8 Proofs for the partial-selfing regime

In this section, we assume that $N(1-\alpha_N) \to \infty$ and $\alpha_N \to \alpha \in [0,1]$ as $N \to \infty$.

We shall see that, for samples in different individuals, the distribution of $\tau^{(N,2)}$ conditioned on the pedigree is related to the minimum of two pairwise coalescence times conditionally independent with respect to the pedigree. These two conditionally independent pairwise coalescence times are the pairwise coalescence times of two conditionally independent pairs of random walks, (X,Y) and (X',Y'), on $J \times \mathbb{Z}_+$. This is a slightly more complicated argument than was presented in the proof of Corollary 15 in Section 5.1.

8.1 Two conditionally independent *pairs* of genes

In the partial-selfing, direct analysis regime of $\tau^{(N,2)}$ conditional on $\mathcal{A}_N$ is more difficult. We can proceed by connecting the behavior of $\tau^{(N,2)}$ to that of two copies of $\tau^{(N,2)}$ conditionally independent with respect to the pedigree with the following lemma.

Lemma 8.1. *For any fixed $t \in \mathbb{R}$ and $N \geq 2$,*

$$\mathbb{E}_{\text{diff}}\left[\mathbb{P}_{\text{diff}}(N^{-2}\tau^{N,2} > t|\mathcal{A}_N)^2\right] = \mathbb{P}_{\text{diff}}\left(N^{-2}\tau \wedge \tau' > t\right).$$

Proof. By conditional independence we have

$$\mathbb{E}_{\text{diff}}\left[\mathbb{P}_{\text{diff}}(N^{-2}\tau > t|\mathcal{A}_N)\right]\mathbb{E}_{\text{diff}}\left[\mathbb{P}_{\text{diff}}(N^{-2}\tau' > t|\mathcal{A}_N)\right] = \mathbb{E}_{\text{diff}}\left[\mathbb{P}_{\text{diff}}(N^{-2}\tau \wedge \tau' > t|\mathcal{A}_N)\right].$$

The result follows immediately. □

Let τ and τ' be conditionally independent pairwise coalescence times with respect to $\mathcal{A}_N$. τ and τ' are the pairwise coalescence times of two conditionally independent pairs of random walks on the same pedigree, denoted by (X,Y) and (X',Y'). The collection of all individuals occupied by one of these random walks k time-steps in the past is denoted by $p_k = \{\hat{X}_k, \hat{Y}_k, \hat{X'}_k, \hat{Y'}_k\}$. Just as in the unconditional case, the key objects of study are overlaps and splits. In this case, though, we need to track 4 sample lineages, not simply two. Luckily, at least in the subcritical regime, our analysis will work in much the same way.

We now precisely define the overlap and splitting times. We let $\tau_0^O = 0$ by convention, and, for $i \geq 1$ we let

$$\tau_i^O = \inf\{k \in \mathbb{Z}_+ : k > \tau_{i-1}^O, |p_k| < |p_{k-1}|\}.$$

These can be understood as the overlap times, where (at least) two of our four sample lineages transition to belonging in fewer individuals than before, with each other.

With respect to splits we define τ_0^D to be

$$\tau_0^D := \inf\{k \in \mathbb{Z}_+ : |p_k| = 4\},$$

the first time that all four sample lineages are in four distinct individuals, if it exists. We then iteratively define, for $i \geq 1$,

$$\tau_i^D := \inf\{k \in \mathbb{Z}_+ : |p_k| > |p_{k-1}|\}.$$

The infimum of an empty set is taken to be infinite, by convention.

We let $\mathcal{O}$ denote the total number of finite overlap times before $\tau \wedge \tau'$ i.e.,

$$\mathcal{O} := |\{i \in \mathbb{Z}_> : \tau_i^O \leq \tau \wedge \tau'\}|.$$

Lemma 8.2. *Let Γ_N denote the set on which*

$$0 = \tau_0^O < \tau_0^D < \tau_1^O < \tau_1^D < \ldots < \tau_{\mathcal{O}-1}^D < \tau_{\mathcal{O}}^O \leq \tau \wedge \tau' \quad \mathbb{P}_{\text{diff}} - a.s. \tag{52}$$

Then $\mathbb{P}_{\text{diff}}(\Gamma_N)$ converges to 1 as N tends to infinity if $N(1-\alpha_N) \to \infty$. Further, $\mathcal{O}$ under $\mathbb{P}_{\text{diff}}(\cdot|\Gamma_N)$ is geometric with parameter $\frac{1}{6}\frac{\alpha_N}{2-\alpha_N}$.

The primary use of a lemma such as this is to get a telescoping sum representation of $\tau \wedge \tau'$ as follows:

$$\tau \wedge \tau' = \tau \wedge \tau' - \tau_{\mathcal{O}}^O + \sum_{i=0}^{\mathcal{O}} \left(\tau_i^O - \tau_{i-1}^D\right) + \sum_{i=1}^{\mathcal{O}-1} \left(\tau_i^D - \tau_i^O\right). \tag{53}$$

In order to prove Lemma 8.2, we first prove two other lemmas. We begin by calculating the probability that $\tau_0^D < \tau_1^O$. We then calculate the probability that any overlap which occurs before $\tau \wedge \tau'$ is followed either by a split or coalescence. We also quantify the times of the splits.

Lemma 8.3. *The probability that $\tau_0^D < \tau_1^O$ under the law $\mathbb{P}_{\text{diff}}$ converges to one as N goes to infinity. Further, $N^{-2}\tau_0^D$ converges to 0 in distribution conditional on being less than τ_1^O.*

Proof. The rate λ_N at which either individual splits without an overlap event is

$$(1-\alpha_N)\frac{2}{N}\frac{N-2}{N-3}\frac{N-3}{N-2}\frac{1}{2} + (1-\alpha_N)\frac{2}{N}\frac{2}{N-1}\frac{1}{4} = (1-\alpha_N)\frac{N-2}{N(N-1)}.$$

The rate μ_N at which there is an overlap event is

$$(1-\alpha_N)\frac{2}{N}\frac{2}{N-1}\frac{3}{4} + \alpha_N\frac{2}{N(N-1)} = \frac{3-\alpha_N}{N(N-1)}.$$

The probability, therefore, that either individual splits without an overlap event is

$$1 - \frac{\mu_N}{\mu_N + \lambda_N} = 1 - \frac{3-\alpha_N}{3-\alpha_N + (1-\alpha_N)(N-2)}.$$

The time it takes for such a desired split to occur, b_0, can therefore be observed to be geometric with parameter

$$\frac{\lambda_N}{1-\frac{\mu_N}{\lambda_N}} = \frac{\lambda_N^2}{\lambda_N - \mu_N}.$$

Once either individual splits without an overlap event, we have 4 sample lineages in 3 individuals. The rate λ'_N at which the occupied individual containing two sample lineages splits without an overlap event is

$$(1-\alpha_N)\frac{1}{N}\left(\frac{2}{N-1}\frac{N-3}{N-2} + \frac{N-3}{N-1}\frac{N-4}{N-2}\right) = \frac{(1-\alpha_N)(N-3)(2N-7)}{2N(N-1)(N-2)}.$$

The rate μ'_N depends more precisely on which individual is chosen as the offspring. If the offspring contains two conditionally independent sample lineages and one parent is occupied, then there is a $\frac{3}{4}$ chance of having an overlap event. Regardless of which occupied offspring is chosen, if both parents are occupied then an overlap event is guaranteed. If an overlap is due to selfing then an overlap occurs if and only if both parent and offspring are occupied. Thus, μ'_N can be calculated as

$$(1-\alpha_N)\left(\frac{1}{N}\frac{4}{N-1}\frac{N-3}{N-2}\frac{3}{4} + \frac{2}{N}\frac{4}{N-1}\frac{N-3}{N-2}\frac{1}{2} + \frac{3}{N}\frac{2}{N-1}\frac{1}{N-2}\right) + \alpha_N\frac{6}{N(N-1)}.$$

By some simplification this becomes

$$\frac{N(7-\alpha_N)-15+3\alpha_N}{N(N-1)(N-2)}.$$

Therefore the probability that we have four lineages in four individuals before any overlap event is

$$1-\frac{\mu'_N}{\lambda'_N+\mu'_N}=1-\frac{2(N(7-\alpha_N)-15+3\alpha_N)}{(1-\alpha_N)(2N^2-15N+18)+12N-12},$$

and the time, b_1 it takes for this to occur is geometric with parameter $\frac{(\lambda'_N)^2}{\lambda'_N-\mu'_N}$. This gives the claim using the fact that $N(1-\alpha_N)\to\infty$ □

Lemma 8.4. *The probability*

$$r_N=\mathbb{P}_{\text{diff}}(\tau\wedge\tau'<\tau^O_{i+1}\ \textit{or}\ \tau^D_{i+1}<\tau\wedge\tau'||p_{\tau^O_i}|=3,\hat{X}_{\tau^O_i}=\hat{Y}_{\tau^O_i}\ \textit{or}\ \hat{X'}_{\tau^O_i}=\hat{Y'}_{\tau^O_i})$$

that a split in an individual containing two sample lineages, given that individual contains one of the two possibly coalescent pairs x,y or x',y', occurs before the next overlap or that coalescence occurs before that next split converges to one as N goes to infinity.

Proof. The probability λ that an occupied individual containing two sample lineages splits or else has these two lineages coalesce via splitting in a single time-step, both without an overlap event, is

$$(1-\alpha_N)\frac{1}{N}\frac{N-3}{N-1}\frac{N-4}{N-2}+(1-\alpha_N)\frac{1}{N}\frac{4}{N-1}\frac{N-3}{N-2}\frac{1}{2}+\alpha_N\frac{1}{N}\frac{N-3}{N-1}\frac{1}{2} \tag{54}$$

$$=\frac{(N-3)((2-\alpha_N)N+2(1-\alpha_N)-2)}{2N(N-1)(N-2)} \tag{55}$$

The probability the coalescent event occurs before the splitting event can be calculated as the ratio of selfing coalescence probability and λ. The probability μ that there is an overlap event involving the three occupied individuals is

$$\frac{N(7-\alpha_N)-3(5-\alpha_N)}{N(N-1)(N-2)}.$$

Therefore the probability of the two sample lineages in the same individual coalescing or splitting before another overlap event is

$$\frac{\lambda}{\lambda+\mu}=1-\frac{\mu}{\lambda+\mu} \tag{56}$$

$$=1-\frac{(14-2\alpha_N)N-15+3\alpha_N}{(2-\alpha_N)N^2+(8-\alpha_N)N-30+12\alpha_N}. \tag{57}$$

This gives the claim. □

We can now proceed with the proof of Lemma 8.2.

Proof of Lemma 8.2. We begin with $\mathcal{O}$. The behavior of $\mathcal{O}$ is determined by the structure of the representation (52). Indeed, as τ^D_0 is finite on Γ_N we reach a state where all four lineages are in four distinct individuals. From this state, it is only possible for $|p_k|$ to transition from four to

three. The presence of splits between each overlap indicates that $|p_k|$ oscillates between 3 and 4 after τ_0^D and before $\tau \wedge \tau'$.

Therefore, the relevant overlaps of pairs occur are those that occur from the state in which all four lineages are in distinct individuals. There are $\binom{4}{2}$ equally likely possible pairs of lineages to overlap, but only two, those between x and y or x' and y', that could result in coalescence. Therefore there is a $\frac{1}{3}$ chance of either of these two potentially coalescent pairs overlapping at an overlap time. These lineages coalesce instantaneously with a further probability of $\frac{1}{2}$, but they may coalesce before the next split with some positive probability.

Let U denote the number of selfing events, in the single individual containing two sample lineages, before the next split, conditional on Γ_N. The probability, conditional on Γ_N, that we have a selfing event before a split is α_N. Therefore

$$\mathbb{P}(U = j) = \alpha_N^j(1 - \alpha_N).$$

When $U = k$ there is a

$$2^{-1} + ... + 2^{-k} = 1 - 2^{-k}$$

chance of coalescing. Therefore the probability of coalescing due to any given overlap, conditional on Γ_N, is

$$\frac{1}{3}\left(\frac{1}{2} + \frac{1}{2}\mathbb{E}\left[1 - 2^{-U})\right]\right) = \frac{1}{6}\frac{\alpha_N}{2 - \alpha_N}.$$

It suffices now to show that the probability of Γ_N converges to one as N goes to infinity. This is precisely that probability that $\tau_0^D < \tau_1^O$ multiplied by the probability that each of the next $\mathcal{O} - 1$ splits each occur before the next overlap, or else coalesce. The probability that $\tau_0^D < \tau_1^O$ converges to one by Lemma 8.3.

The probability that we have the structure in Γ_N after τ_0^D is exactly $\mathbb{E}\left[r_N^{\mathcal{O}}\right]$, where r_N is the probability given by Lemma 8.4.

As r_N converges to one as N goes to infinity and $\mathcal{O}$ is simply a geometric random variable with parameter strictly bounded away from 0, $\mathbb{E}\left[r_N^{\mathcal{O}}\right]$ converges to one as N goes to infinity. This gives the claim. □

It remains simply to quantify the gap between overlaps and splits given by the following lemma, and the gap between the final overlap and coalescence.

Lemma 8.5. *As $N \to \infty$, the conditional expectations $\left\{\mathbb{E}\left[N^{-2}(\tau_i^D - \tau_{i-1}^O)|\tau_i^D < \tau_i^O\right]\right\}_{i\in\mathbb{Z}_{>0}}$ and $\mathbb{E}\left[N^{-2}(\tau^{(N,2)} - \tau_{\mathcal{O}}^O))\,\middle|\, |p_{\tau_{\mathcal{O}}^O}| = 3\right]$ all converge to 0.*

Proof. We demonstrate the first limit. The probability of splitting of the multiply-occupied individual in a single step without any overlap event is

$$\frac{(1 - \alpha_N)(N - 3)(N - 4)}{N(N - 1)(N - 2)}.$$

Therefore, conditioning on knowing $\tau_i^D < \tau_i^O$ implies $(\tau_i^D - \tau_{i-1}^O)$ is geometric with this parameter. This gives the claim. We now prove the second limit. The difference $\tau^{(N,2)} - \tau_{\mathcal{O}}^O$ is stochastically dominated by the time it takes for the relevant occupied individual to split, which is geometric with parameter $(1 - \alpha_N)N^{-1}$. Therefore,

$$\mathbb{E}\left[N^{-2}(\tau^{(N,2)} - \tau_{\mathcal{O}}^O)\right] \leq \frac{N}{1 - \alpha_N}N^{-2} = \frac{1}{N(1 - \alpha_N)}$$

which converges to 0 as N goes to infinity as $N(1-\alpha_N) \to \infty$. □

We are now prepared to prove the limiting behavior of $\tau \wedge \tau'$ in the subcritical regime.

Theorem 8.6. *Suppose $\alpha_N \to \alpha \in [0,1]$ and $N(1-\alpha_N) \to \infty$. Then for any fixed $t > 0$, as $N \to \infty$, $N^{-2}(\tau \wedge \tau')$ converges in distribution to an exponential random variable with rate $\frac{4}{2-\alpha}$.*

Proof. By Lemma 8.2

$$\mathbb{E}_{\text{diff}}\left[e^{itN^{-2}\tau\wedge\tau'}\right] = \mathbb{E}_{\text{diff}}\left[e^{itN^{-2}\tau\wedge\tau'}|\Gamma_N\right] + o(1). \tag{58}$$

We therefore can decompose $\tau \wedge \tau'$ as in (53). Let σ_N denote the characteristic function of $\tau_i^D - \tau_i^O$ conditional on Γ_N for $i \geq 1$, σ'_N denote the characteristic function of τ_0^D, and φ_N denote the characteristic function of each of the $\tau_i^O - \tau_{i-1}^D$, conditional on Γ_N. They all have the same characteristic function as they are all independent and identically distributed. Finally, let ψ_N denote the characteristic function of $\tau \wedge \tau' - \tau_{\mathcal{O}}^O$, again conditional on Γ_N. By Lemmas 8.5 and 8.3 we have that $\psi_N(tN^{-2})$, $\sigma_N(tN^{-2})$, $\sigma'_N(tN^{-2})$ all converge to 1 as $N \to \infty$.

By Lemma 8.3,

$$\mathbb{E}_{\text{diff}}\left[e^{itN^{-2}\tau\wedge\tau'}|\Gamma_N\right] = \sigma'_N(tN^{-2})\psi_N(tN^{-2})\sum_{j=1}^{\infty}\mathbb{P}_{\text{diff}}(\mathcal{O}=j|\Gamma_N)\varphi_N(tN^{-2})^j\sigma_N^{j-1}. \tag{59}$$

As $\mathbb{P}_{\text{diff}}(\mathcal{O}=j|\Gamma_N) = (\beta_N)^j(1-\beta_N)$ for $\beta_N = \frac{1}{6}\frac{\alpha_N}{2-\alpha_N}$ by Lemma 8.3, (59) becomes

$$\sigma'_N(tN^{-2})\psi_N(tN^{-2})\frac{\beta_N\varphi_N(tN^{-2})}{1-(1-\beta_N)\varphi_N(tN^{-2})\psi_N(tN^{-2})}. \tag{60}$$

Note, φ_N is the characteristic function of a geometric random variable with parameter $2\binom{4}{2}N^{-2} + O(N^{-3})$. Therefore $\varphi_N(tN^{-2})$ converges to the characteristic function φ of an exponential random variable with rate $2\binom{4}{2} = 12$. Further, β_N converges to $\beta = \frac{1}{6}\frac{\alpha}{2-\alpha}$.

Therefore (60) converges as N goes to infinity to

$$\frac{\beta\varphi}{1-(1-\beta)\varphi}.$$

By some algebra this is the characteristic function of an exponential random variable with rate $12\beta = \frac{4}{2-\alpha}$. This is the claim. □

We can now arrive at the main proof of this section, the limit conditional on the pedigree of $N^{-2}\tau^{(N,2)}$.

Theorem 8.7. *Suppose $\alpha_N \to \alpha \in [0,1]$ and $N(1-\alpha_N) \to \infty$. Then for any fixed $t > 0$, as $N \to \infty$, $\mathbb{P}_{\text{diff}}(N^{-2}\tau^{N,2} > t \mid \mathcal{A}_N)$ converges to $e^{-\frac{2}{2-\alpha}t}$ in L^2.*

Proof. By expanding out the squared L^2 distance between $\mathbb{P}_{\text{diff}}(N^{-2}\tau^{N,2} > t \mid \mathcal{A}_N)$ and e^{-2t} and applying lemma 8.1 yields

$$\mathbb{P}_{\text{diff}}(N^{-2}\tau\wedge\tau' > t) - 2\mathbb{P}_{\text{diff}}(N^{-2}\tau^{N,2} > t)e^{-\frac{2}{2-\alpha}} + e^{-\frac{4}{2-\alpha}t}. \tag{61}$$

By 3.1 it follows that the second summand in 61 converges, as N goes to infinity, to $e^{-\frac{4}{w-\alpha}t}$. It follows, therefore, that the squared L^2 distance in equation 61 is

$$\mathbb{P}_{\text{diff}}(N^{-2}\tau \wedge \tau' > t) - e^{-\frac{4}{2-\alpha}t} + o(1). \tag{62}$$

That is, we have L^2 convergence so long as $N^{-2}\tau \wedge \tau'$ converges in distribution to an exponential random variable with rate $\frac{4}{2-\alpha}$, but this is exactly the result of Lemma 8.6. □

In particular, in the partial-selfing regime, we have that the classical approximation by averaging over the pedigrees when we begin in distinct individuals is robust to the pedigree.

8.2 Pairwise conditional convergence starting from a single individual

We proceed now to continue our analysis to the case where both sample lineages are taken from the same individual. Thankfully, much of the analysis in this section relies greatly on the work we have already established for $\mathbb{P}_{\text{diff}}$.

Lemma 8.8. *Let $\tilde{b}$ denote the infimum over the time-steps k such that $\hat{X}_k \neq \hat{Y}_k$. The infimum of an empty set is infinity. Then*

$$\mathbb{P}_{\text{same}}(\tau \wedge \tau' > \tilde{b}) = \frac{4(1-\alpha_N)}{4-\alpha_N}.$$

Further, $\mathbb{E}_{\text{same}}\left[N^{-2}\tilde{b} \mid \tilde{b} < \infty\right]$ converges to 0 as N goes to infinity.

Proof. Let U denote the number of selfing events of the single occupied individual before splitting. The rate at which selfing events in this individual occur is $\alpha_N N^{-1}$. The rate of splitting is $(1-\alpha_N)N^{-1}$. Therefore the probability of a splitting event before coalescing is $1-\alpha_N$. Therefore $\mathbb{P}(U = k) = \alpha_N^k(1-\alpha_N)$. That is, U is 0 with probability $1-\alpha_N$, else it is a geometric random variable with parameter $1-\alpha_N$.

The probability that neither pair coalesces during a selfing event is $\frac{1}{4}$. Therefore

$$\mathbb{P}_{\text{same}}(\tau \wedge \tau' > \tilde{b} | U = k) = 4^{-k}.$$

$\mathbb{E}_{\text{same}}\left[4^{-U}\right] = \frac{4(1-\alpha_N)}{4-\alpha_N}$.

That $N^{-2}\tilde{b}$ converges to 0 follows from it being geometric with parameter $(1-\alpha_N)N^{-1}$ and $N(1-\alpha_N) \to 0$. □

Corollary 8.9. *As $N \to \infty$, $\mathbb{E}_{\text{same}}\left[e^{itN^{-2}\tau\wedge\tau'}\right]$ converges to*

$$\frac{3\alpha}{4-\alpha} + (1 - \frac{3\alpha}{4-\alpha})(1 + it(\frac{4}{2-\alpha})^{-1})^{-1}.$$

Here a geometric random variable with parameter 0 is taken simply to be infinite.

Proof. We can decompose $\mathbb{E}_{\text{same}}\left[e^{itN^{-2}\tau\wedge\tau'}\right]$ into where coalescence occurs before the first splitting time $\tilde{b}$. The probability that it occurs before the first split is $\frac{3\alpha_N}{4-\alpha_N}$ and the time for it to occur, given that we have coalesce converges in distribution to 0 with the time rescaling by

Lemma 8.8. Therefore, conditional on $\{\tau\wedge\tau' < \tilde{b}\}$, with the $\mathbb{P}_{\text{same}}$ sampling scheme $\tau\wedge\tau'$ converges in distribution to 0.

Similarly we can condition on $\{\tau \wedge \tau' > \tilde{b}\}$, where we transition into the initial conditions for the two pairs under $\mathbb{P}_{\text{diff}}$. Therefore

$$\mathbb{P}_{\text{same}}(N^{-2}\tau \wedge \tau' > t|\tau \wedge \tau' > \tilde{b}) = \mathbb{P}_{\text{diff}}(N^{-2}(\tau \wedge \tau' + \tilde{b}) > t). \tag{63}$$

Again by Lemma 8.8 we know N^{-2} converges to 0 in distribution as N goes to infinity so (63) converges, by Theorem 8.7 to $e^{-\frac{4}{2-\alpha}t}$ as N goes to infinity. In particular, then, $\tau \wedge \tau'$ under the law $\mathbb{P}_{\text{same}}(\cdot|\tau \wedge \tau' > \tilde{b})$ converges to an exponential random variable with rate $\frac{4}{2-\alpha}$, who has a characteristic function $(1 + it(\frac{4}{2-\alpha})^{-1})^{-1}$.

Combining the two conditional limits gives the result. □

Theorem 8.10. *Suppose $\alpha_N \to \alpha \in [0, 1]$ and $N(1 - \alpha_N) \to \infty$. Then for any fixed $t > 0$, as $N \to \infty$,*

$$\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N) \to 2^{-U}e^{-\frac{2}{2-\alpha}t}$$

where $\mathbb{P}(U = k) = \alpha^k(1 - \alpha)$ for k in $\mathbb{Z}_+$.

Proof. Let U_N denote the number of overlap events that occur in the single occupied individual before this individual undergoes a splitting event and let $\tilde{b}$ denote the time-step of this split. The probability of k overlaps before the first split in the single occupied individual is $\alpha_N^k(1 - \alpha_N)$. Therefore U_N converges in distribution to a random variable U satisfying $\mathbb{P}(U = k) = \alpha^k(1 - \alpha)$, the same U as in the statement of the theorem.

The key observation is that $\tau^{(N,2)}$ coalesces "instantaneously" with probability $1 - 2^{-U_N}$, else it reaches a state where the two lineages are in distinct individuals and that these two distinct individuals are directly readable from the pedigree, reducing to the case where we started in distinct individuals. Equivalent to the statement is showing that $\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N)$ converges to $2^{-U}e^{-\frac{2}{2-\alpha}t}$ in L^2.

Let B_N denote the set where $N^{-2}\tilde{b} < t$. B_N converges to one in probability by Lemma 8.8. In particular,

$$\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N) = \mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N, B_N) + o(1).$$

Further,

$$\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N, B_N) = 2^{-U_N}\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N, B_N, \tau^{(N,2)} > \tilde{b}).$$

Therefore

$$\mathbb{E}_{\text{same}}\left[(\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N) - 2^{-U_N}e^{-\frac{2}{2-\alpha}t})^2\right] = \mathbb{E}_{\text{same}}\left[(\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t|\mathcal{A}_N) - 2^{-U_N}e^{-\frac{2}{2-\alpha}t})^2\middle|B_N\right] + o(1$$

is

$$\begin{aligned}&\mathbb{P}_{\text{same}}(N^{-2}\tau \wedge \tau' > t) + \mathbb{E}_{\text{same}}\left[4^{-U_N}\right]e^{-\frac{4}{2-\alpha}t}\\ &- 2\mathbb{E}_{\text{same}}\left[4^{-U_N}\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t\middle|\tau^{(N,2)} > \tilde{b}, B_N, \mathcal{A}_N)\right]e^{-\frac{2}{2-\alpha}t} + o(1).\end{aligned} \tag{64}$$

See that the sum of the first and second summands converges to

$$\frac{8(1-\alpha)}{2-\alpha}e^{-\frac{4}{2-\alpha}t}$$

by Corollary 8.9 and that $\mathbb{E}_{\text{same}}\left[4^{-U_N}\right] = \frac{4(1-\alpha_N)}{2-\alpha}$. It suffices therefore to show that the third summand converges to the negation of this term.

The key observation is that 4^{-U_N} is independent to the pairwise coalescence time conditional on the pedigree given that the coalescence time occurs after the split. In particular,

$$\mathbb{E}_{\text{same}}\left[4^{-U_N}\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t | N^{-2}\tau^{(N,2)} > \tilde{b}, B_N) \mid \mathcal{A}_N\right]$$

decomposes into a product

$$\mathbb{E}_{\text{same}}\left[4^{-U_N}\right]\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t | \tau^{(N,2)} > \tilde{b}, B_N).$$

$\mathbb{E}_{\text{same}}\left[4^{-U_N}\right]$ may be calculated directly as $\frac{4(1-\alpha_N)}{4-\alpha_N}$ and $\mathbb{P}_{\text{same}}(N^{-2}\tau^{(N,2)} > t | N^{-2}\tau^{(N,2)} > N^{-2}\tilde{b}, B_N)$, which converges to $e^{-\frac{2}{2-\alpha}t}$ by Corollary 10.4 and the fact that conditioning on $N^{-2}\tau^{(N,2)} > \tilde{b}$ is equivalent to conditioning on there being a split before coalescence. Therefore the third summand in (64) converges to

$$-\frac{8(1-\alpha)}{2-\alpha}e^{-\frac{4}{2-\alpha}t},$$

which we established was sufficient for the claim. □

8.3 Proofs for Theorem 4.3 and Theorem 4.4

Proof of Theorem 4.3. The proof of Theorem 4.3 in the partial selfing regime is by expanding the L^2 distance between the conditional survival probability and the survival probability of an exponential random variable. This is performed using a second moment estimate of the conditional survival probability provided by Lemma 8.1. The cross-term and the square of the survival probability of the exponential simplify by Theorem 3.1, reducing the problem to convergence in distribution of the minimum of two conditionally independent copies of the pairwise coalescence time. This convergence is established in Theorem 8.6.

The proof in the limited outcrossing and negligible outcrossing regimes are the same. In this case, the subgraph of the pedigree generated by following all outcrossings and the random walks thereon are shown to converge in Skorokhod space by Lemma 6.2. This generates a coupling between random walks on the pedigree and random walks on this convergent subgraph. The convergence given by Lemma 6.2 is sufficient to show that the conditional survival probability converges due to its dependence only on the discrete structure of the subgraph and the lengths of its edges, which converge jointly.

□

Proof of Theorem 4.4. The proof in the partial selfing regime of Theorem 4.4 follows the same L^2 expansion argument as in the partial selfing regime of Theorem 4.3, including the same use of a second moment calculation as in Lemma 8.1. The first and third summands in the expansion can be calculated to be the same by showing that the minimum of two conditionally independent realizations of $\tau^{N,2}$ under the "same" sampling configuration is 0 with positive probability, and is otherwise exponential with rate $\frac{4}{2-\alpha}$ using Lemma 8.8. The cross term can be shown to converge to negative of the above sum using independence of the number of selfing events before the two sample lineages split and the pairwise coalescence time conditional on undergoing a split before

coalescence, and then also using the fact that the time of the first splitting event converges to 0 with the N^2 time rescaling by Lemma 8.8.

In the case where α_N converges to 1, it follows that the conditional coalescence time converges to 0 in distribution by using Theorem 3.1 and the conditional Markov inequality, per Lemma 6.8.

□

9 Discussion

In this work, we have assumed that the pedigree of a population of constant size N is the outcome of a random process of reproduction in which offspring are produced by self-fertilization with probability α_N. By conditioning on the pedigree and considering pairwise times to common ancestry, we found three different coalescent models in the limit $N \to \infty$, when time is measured in units of N generations. The deciding factor is $N(1-\alpha_N)$, which may be interpreted either as the rate of outcrossing events along an ancestral lineage or as the expected number of outcrossed offspring in the population each generation. The three models are 'negligible outcrossing' which applies when $N(1-\alpha_N) \to 0$, 'limited outcrossing' which applies when $N(1-\alpha_N) \to \lambda \in (0,\infty)$, and 'partial selfing' which applies when $N(1-\alpha_N) \to \infty$. Negligible outcrossing and limited outcrossing both require $\alpha_N \to 1$, while partial selfing includes this as a special case.

Previous population-genetic analyses of selfing populations have assumed a fixed selfing rate, $\alpha_N = \alpha$ in our notation, and have implicitly averaged over pedigrees. When this averaging is done, just one coalescent model for all $\alpha \in [0,1]$ emerges in the limit, specifically the coalescent model with partial selfing described by Nordborg and Donnelly (1997) and Möhle (1998). When coalescence is conditioned on the pedigree and α is fixed, our very similar partial-selfing model is obtained but only for $\alpha \in [0,1)$. Assuming a fixed selfing rate of $\alpha = 1$ leads to an entirely different model, namely negligible outcrossing.

A more detailed look at how α_N approaches 1 reveals the critical case of limited outcrossing in between negligible outcrossing and partial selfing and characterized by the parameter $N(1-\alpha_N) \to \lambda \in (0,\infty)$. In this regime, the parts of the population pedigree which might be in the ancestry of the sampled individuals are replaced by a random graph like those previously used to model recombination (Griffiths, 1991; Griffiths and Marjoram, 1997) and selection (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). It only differs in its construction by having λ as the rate at which ancestral lineages split. Each splitting event in the graph corresponds to an outcrossing event in the genealogical ancestry of the sampled individuals.

Depending on λ, limited outcrossing displays a range of behaviors between the two qualitatively different extremes of negligible outcrossing and partial selfing. Figure 8 illustrates this using the law of total variance for the coalescence time of two gene copies sampled from two different individuals. Let T and G represent this "diff" coalescence time and its associated ancestral graph, for simplicity dropping the subscript λ we attached to these previously in Section 5. Then we can write

$$\mathrm{Var}(T) = \mathbb{E}[\mathrm{Var}(T|G)] + \mathrm{Var}(\mathbb{E}[T|G]) \tag{65}$$

and note that $\mathrm{Var}(T) = 1/4$, because T by itself is exponentially distributed with rate parameter $2/(2-\alpha_N) \to 2$. The second term on the right-hand side of (65), $\mathrm{Var}(\mathbb{E}[T|G])$, is given by (16). The first term on the right-hand side of (65) is then simply $\mathbb{E}[\mathrm{Var}(T|G)] = 1/4 - \mathrm{Var}(\mathbb{E}[T|G])$, but both terms are displayed in Figure 8 for the sake of illustration.

When λ is very small, on the far left in Figure 8, variation in the mean coalescence time

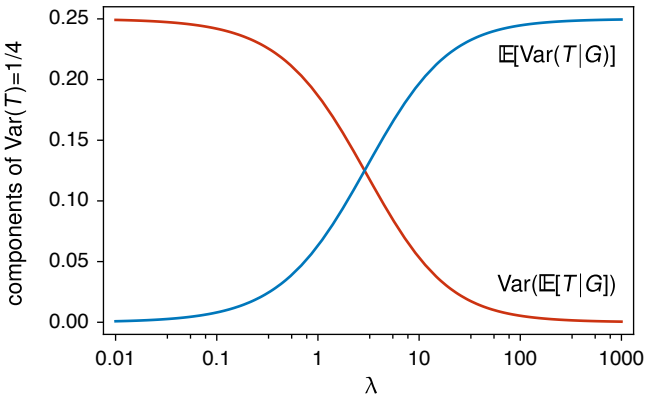


Figure 8: The two components, $\mathbb{E}[\mathrm{Var}(T|G)]$ and $\mathrm{Var}(\mathbb{E}[T|G])$, of $\mathrm{Var}(T)$ as functions of λ under limited outcrossing. By the law of total variance, $\mathrm{Var}(T) = \mathbb{E}[\mathrm{Var}(T|G)] + \mathrm{Var}(\mathbb{E}[T|G])$. Also, $\mathrm{Var}(T) = 1/4$ in this case because it is the average over ancestral graphs (G), that is to say over pedigrees, and because $\alpha_N \to 1$.

among graphs, $\mathrm{Var}(\mathbb{E}[T|G])$, accounts for the bulk of variation in T. Correspondingly, there is little variation in coalescence times given the graph. Figure 8 displays the expectation of the latter, $\mathbb{E}[\mathrm{Var}(T|G)]$, but $\mathrm{Var}(T|G)$ itself will be small on any graph in this case. With probability approaching one as $\lambda \to 0$, the individuals containing the two ancestral lineages will experience only selfing events in their ancestries, up to and including the time they descend from a common ancestral individual. Pairwise coalescence times at every locus in the genome starting from these two individuals will converge on the time of this ancestor and $\mathrm{Var}(T|G)$ should be minimal. Variation in the waiting time to this first common ancestral individual among graphs will be the primary source of variation in T. In the limit $\lambda \to 0$, this waiting time is exponentially distributed with rate parameter $2/(2 - \alpha_N) \to 2$, so $\mathrm{Var}(\mathbb{E}[T|G])$ approaches $\mathrm{Var}(T) = 1/4$. It is given by a Kingman coalescent process made effectively haploid by selfing, just as under negligible outcrossing.

Across the middle of Figure 8, ancestral graphs will include increasing numbers of splitting or outcrossing events as λ increases. Such graphs contain multiple pathways to coalescence because genetic lineages may trace back to either parent at each outcrossing event. They include multiple possible coalescence times, as already evident in the CDFs of $T|G$ in Figure 6 for a hypothetical graph with three splitting events and three possible coalescence times, in Figure 3 for each of five simulated ancestral graphs with $\lambda = 2$, and in Figure 1 for each of fifty pedigrees in three versions of a finite-N Wright-Fisher model corresponding to $\lambda \in \{1, 10, 100\}$. As a result, $\mathrm{Var}(T|G)$ is positive and $\mathbb{E}[\mathrm{Var}(T|G)]$ becomes the main source of variation in T as λ increases, eventually swamping variation of the mean coalescence times among graphs.

When λ is very large, on the far right in Figure 8, single ancestral graphs contain no such features constraining $T|G$ to take particular values. These graphs are large. In their construction backward in time, the number of ancestral lineages grows quickly to reach a quasi-stable distribution, specifically a zero-truncated Poisson distribution with mean λ (Mano, 2009). The waiting time to the event that completes the graph, that is when there is just one lineage left, will be long, $\sim 2e^{\lambda}/\lambda^2$ in expectation using equation (1.1) in Griffiths and Marjoram (1997). Given the graph, each of the pair of genetic lineages starting from the two sampled individuals will trace back through a large number of splitting events before the two meet in a single individual and coalesce. As $\lambda \to \infty$, all variation in T will be due to this random process of coalescence given the graph. Corollary 15 shows that this conditional coalescent process converges on that of the partial-selfing model with $\alpha_N \to 1$. Lemma **??** [Max This has now been removed as a lemma] shows further that

$\mathrm{Var}(\mathbb{E}[T|G])$ in (16) is also the expected squared L^2 distance of the CDF under limited outcrossing to that under the corresponding model of partial selfing.

To summarize what is shown in Figure 8, limited outcrossing converges on negligible outcrossing as $\lambda \to 0$ and on partial selfing as $\lambda \to \infty$. With respect to the pre-limiting Moran model, the results and intuitions about Figure 8 concern the behavior at the boundary of complete or nearly complete selfing. We have first negligible outcrossing, then limited outcrossing, then an extreme case of partial selfing with $\alpha = 1$. The entire rest of the range of selfing rates $\alpha \in [0, 1]$ is covered by the partial-selfing model.

As with the well known Wright-Fisher diffusion and corresponding standard neutral coalescent model, these new models of coalescence conditional on the pedigree are meant as robust approximations for large populations. We expect, for example, that the Wright-Fisher model with selfing used to produce Figure 1 will have the same three limiting cases we found for a Moran model with selfing. Thus, Figure 1 and Figure 8 both suggest that if $\lambda := N(1 - \alpha_N)$ is greater than about 100 and α_N is very small, then the partial-selfing model can safely be used in place of the limited-outcrossing model. But another implication of Figure 1 is that when the population size is not particularly large, the partial-selfing model may be a better description of the ancestral process even if α_N is not particularly close to 1 as in Figure 1a and Figure 1b.

We chose to use the name partial selfing because in this case the coalescent process conditional on the pedigree is very similar to the previously described partial-selfing model which did not condition on the pedigree (Nordborg and Donnelly, 1997; Möhle, 1998). In both cases, times to common ancestry for "diff" samples follows the Kingman coalescent process with effective population size $N_e = (2 - s)N/2$. In our formulation, partial selfing is obtained when $N(1 - \alpha_N) \to \infty$. We may also recall that for "diff" samples limited outcrossing converges on partial selfing as $\lambda \to \infty$. The reason pedigrees do not constrain "diff" coalescence times under the partial-selfing model is that outcrossing events dominate the ancestry of the population. The same underlying phenomena are are work here as in the standard models of population genetics, where the fact that individuals generally have two parents strongly influences the pedigree (Chang, 1999; Derrida et al., 1999; Coron and Le Jan, 2022) resulting in a fast mixing time of ancestral genetic processes given the pedigree (Barton and Etheridge, 2011) and in the conditional and unconditional coalescent processes converging to the same Kingman coalescent process (Tyukin, 2015; Diamantidis et al., 2024).

There is an important difference between the conditional and unconditional models of partial selfing, already anticipated in (1) and (2). Both involve a separation of time scales between fast individual-level and slow population-level processes but only the latter is the same in the two formulations. The latter is all that matters for "diff" samples. However, conditioning on the pedigree means fixing the outcome of the fast process, in particular the number of generations of selfing in the ancestry of each individual. In our model, a sampled individual has a single realization of the random variable U in (1) or in Theorem 4.4. Given a particular outcome k, a pair of gene copies in the individual coalesces within k generations with probability

$$F_k := 1 - 2^{-k} \qquad \text{for} \quad k = 0, 1, 2, \ldots$$

and if so has a negligible coalescence time. If it does not coalesce, it enters the "diff" configuration and has a Kingman coalescent time. So there is variation in the "same" coalescent process among individuals when coalescence is conditional on the pedigree.

[WORK IN PROGRESS BELOW]

This has implications for the inference of selfing rates from genetic data.

Then, in place of what Pollak (1987) found for equilibrium probabilities of identity by descent

for "same" and "diff" samples (Φ_1 and Φ_2, respectively) in the unconditional process, we have

$$\Phi_{1,k} = F_k + (1 - F_k)\,\Phi_2 \qquad \text{for} \quad k = 0, 1, 2, \ldots$$

in the process conditional on the pedigree.

Enjalbert and David (2000); David et al. (2007); McClure and Whitlock (2012) use $s^{k-1}(1-s)$ in methods of using multilocus data to estimate selfing rates (s only, not k, so also averaging rather than keeping k fixed)

Gao et al. (2007); Wang et al. (2012); Redelings et al. (2015) do something akin to this, without averaging, and develop methods of estimating both the selfing rates and number of generations back to the most recent outcrossing event for each individual

Motivated by the idea of a recently established local population of a wind-pollinated plant, Wilson and Dawson (2007) developed a pedigree-based method of inference in which the contribution from one of the two parents at each outcrossing event was assumed to have come from a location outside the sampled population. The method infers the local pedigree which is tree-like, or loop-free, and the plus immigrating lineages at each outcrossing, up to the time the local population was established. In this context, multilocus patterns of genetic identity are determined by the number of generations back to the most recent outcrossing event for each individual.

Acknowledgements

This work was supported by National Science Foundation grants DMS-2152103 and DMS-2348164.

References

Abbott R. J. and Gomes M. F. Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh. *Heredity*, 62(3):411–418, 1989. doi: 10.1038/hdy.1989.56.

Avise J. C. and Mank J. E. Evolutionary perspectives on hermaphroditism in fishes. *Sexual Development*, 3(2-3):152–163, 2009. doi: 10.1159/000223079.

Baker H. G. Self-compatibility and establishment after "long-distance" dispersal. *Evolution*, 9(3): 347–349, 1955. doi: 10.1111/j.1558-5646.1955.tb01544.x.

Ball M. R., Nigel J. E., and Avise J. C. Gene genealogies within the organismal pedigrees of random-mating populations. *Evolution; international journal of organic evolution*, 44(2):360–370, 1990. doi: 10.1111/j.1558-5646.1990.tb05205.x.

Barrière A. and Félix M.-A. High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. *Current Biology*, 15(13):1176–1184, 2005. doi: 10.1016/j.cub.2005.06.022.

Barton N. H. and Etheridge A. M. The relation between reproductive value and genetic contribution. *Genetics*, 188(4):953–973, 2011. doi: 10.1534/genetics.111.127555.

Bennett J. H. and Binet F. E. Association between mendelian factors with mixed selfing and random mating. *Heredity*, 10(1):51–55, 1956. doi: 10.1038/hdy.1956.3.

Bertoin J. *Random fragmentation and coagulation processes*, volume 102. Cambridge University Press, 2006.

Birkner M., Blath J., and Eldon B. An Ancestral Recombination Graph for Diploid Populations with Skewed Offspring Distribution. *Genetics*, 193(1):255–290, 01 2013. ISSN 1943-2631. doi: 10.1534/genetics.112.144329. URL `https://doi.org/10.1534/genetics.112.144329`.

Brown K. E. and Kelly J. K. Severe inbreeding depression is predicted by the "rare allele load" in Mimulus guttatus. *Evolution*, 74(3):587–596, 2020. doi: 10.1111/evo.13876.

Busch J. W. The evolution of self-compatibility in geographically peripheral populations of Leavenworthia alabamica (Brassicaceae). *American Journal of Botany*, 92(9):1503–1512, 2005. doi: https://doi.org/10.3732/ajb.92.9.1503.

Cannings C. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Advances in Applied Probability*, 6(2):260–290, 1974. doi: 10.2307/1426293.

Chang J. T. Recent common ancestors of all present-day individuals. *Advances in Applied Probability*, 31(4):1002–1026, 1999. doi: 10.1239/aap/1029955256.

Charlesworth D. Evolution of plant breeding systems. *Current Biology*, 16(17):R726–R735, 2006. doi: 10.1016/j.cub.2006.07.068.

Charlesworth D. and Willis J. H. The genetics of inbreeding depression. *Nature Reviews Genetics*, 10(11):783–796, 2009. doi: 10.1038/nrg2664.

Coron C. and Le Jan Y. Pedigree in the biparental Moran model. *Journal of Mathematical Biology*, 84(6):51, 2022. doi: 10.1007/s00285-022-01752-0.

Cutter A. D. Reproductive transitions in plants and animals: selfing syndrome, sexual selection and speciation. *New Phytologist*, 224(3):1080–1094, 2019. doi: 10.1111/nph.16075.

David P., Pujol B., Viard F., Castella V., and Goudet J. Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology*, 16(12):2474–2487, 2007. doi: 10.1111/j.1365-294X.2007.03330.x.

Derrida B., Manrubia S. C., and Zanette D. H. Statistical properties of genealogical trees. *Physical Review Letters*, 82:1987–1990, 1999. doi: 10.1103/PhysRevLett.82.1987.

Diamantidis D., Fan W.-T. L., Birkner M., and Wakeley J. Bursts of coalescence within population pedigrees whenever big families occur. *Genetics*, page iyae030, 02 2024. ISSN 1943-2631. doi: 10.1093/genetics/iyae030. URL `https://doi.org/10.1093/genetics/iyae030`.

Enjalbert J. and David J. L. Inferring recent outcrossing rates using multilocus individual heterozygosity: Application to evolving wheat populations. *Genetics*, 156(4):1973–1982, 2000. doi: 10.1093/genetics/156.4.1973.

Escobar J. S., Auld J. R., Correa A. C., Alonso J. M., Bony Y. K., Coutellec M., Koene J. M., Pointier J., Jarne P., and David P. Patterns of mating-system evolution in hermaphroditic animals: Correlations among selfing rate, inbreeding depression and the timing of reproduction. *Evolution*, 65(5):1233–1253, 2011. doi: 10.1111/j.1558-5646.2011.01218.x.

Ethier S. N. and Nagylaki T. Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Advances in Applied Probability*, 12(1):14–49, 1980. doi: 10.2307/1426492.

Ethier S. N. and Kurtz T. G. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

Ewens W. J. *Mathematical Population Genetics, Volume I: Theoretical Foundations*. Springer-Verlag, Berlin, 2004.

Fisher R. A. Average excess and average effect of a gene substitution. *Annals of Eugenics*, 11(1): 53–63, 1941. doi: 10.1111/j.1469-1809.1941.tb02272.x.

Fisher R. A. *The Genetical Theory of Natural Selection*. Clarendon, Oxford, 1930.

Gao H., Williamson S., and Bustamante C. D. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, 176(3):1635–1651, 2007. doi: 10.1534/genetics.107.072371.

Golding G. B. and Strobeck C. Linkage disequilibrium in a finite population that is partially selfing. *Genetics*, 94(3):777–789, 1980. doi: 10.1093/genetics/94.3.777.

Goodwillie C., Kalisz S., and Eckert C. G. The evolutionary enigma of mixed mating systems in plants: Occurrence, theoretical explanations, and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, 36:47–79, 2005. doi: 10.1146/annurev.ecolsys.36.091704.175539.

Griffiths R. C. The two-locus ancestral graph. In Basawa I. V. and Taylor R. L., editors, *Selected Proceedings of the Symposium on Applied Probability*, pages 100–117. Institute of Mathematical Statistics, Hayward, CA, USA, 1991.

Griffiths R. C. and Marjoram P. An ancestral recombination graph. In Donnelly P. and Tavaré S., editors, *Progress in Population Genetics and Human Evolution (IMA Volumes in Mathematics and its Applications, vol. 87)*, pages 257–270. Springer-Verlag, New York, 1997.

Haldane J. B. S. A mathematical theory of natural and artificial selection. Part II. The influence of partial self-fertilisation, inbreeding, assortative mating, and selective fertilisation on the composition of Mendelian populations, and on natural selection. *Proceedings of the Cambridge Philosophical Society, Biological Sciences*, 1(3):158–163, 1924. doi: 10.1111/j.1469-185X.1924.tb00546.x.

Hartfield M., Bataillon T., and Glémin S. The evolutionary interplay between adaptation and self-fertilization. *Trends in Genetics*, 33(6):420–431, 2017. doi: 10.1016/j.tig.2017.04.002.

Herbots H. M. The structured coalescent. In Donnelly P. and Tavaré S., editors, *Progress in Population Genetics and Human Evolution (IMA Volumes in Mathematics and its Applications, vol. 87*, pages 231–255. Springer-Verlag, New York, 1997.

Higham N. J. *Functions of Matrices*. Society for Industrial and Applied Mathematics, Philadelphia, 2008. doi: 10.1137/1.9780898717778.

Hudson R. R. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37(1):203–217, 1983a. doi: 10.1111/j.1558-5646.1983.tb05528.x.

Hudson R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983b. ISSN 0040-5809. doi: https://doi.org/10.1016/0040-5809(83)90013-8. URL https://www.sciencedirect.com/science/article/pii/0040580983900138.

Ingvarsson P. K. A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. *Evolution*, 56(12):2368–2373, 2002. doi: 10.1111/j.0014-3820.2002.tb00162.x.

Jarne P. and Auld J. R. Animals mix it up too: The distribution of self-fertilization among hermaphroditic animals. *Evolution*, 60(9):1816–1824, 2006. doi: 10.1111/j.0014-3820.2006.tb00525.x.

Kallenberg O. *Random Measures, Theory and Applications*, volume 77 of *Probability Theory and Stochastic Modeling*. Springer Cham, 1st edition, 2017.

Kamran-Disfani A. and Agrawal A. F. Selfing, adaptation and background selection in finite populations. *Journal of Evolutionary Biology*, 27(7):1360–1371, 2014. doi: 10.1111/jeb.12343.

Kingman J. F. C. On the genealogy of large populations. *Journal of Applied Probability*, 19(A): 27–43, 1982. doi: 10.2307/3213548.

Kipnis C. and Landim C. *Scaling limits of interacting particle systems*, volume 320. Springer Science & Business Media, 1998.

Kogan D., Diamantidis D., Wakeley J., and Fan W.-T. L. Correlation of coalescence times in a diploid wright-fisher model with recombination and selfing. *bioRxiv*, 2023. doi: 10.1101/2023.10.18.563014. URL https://www.biorxiv.org/content/early/2023/10/21/2023.10.18.563014.

Kondrashov A. S. Deleterious mutation as an evolutionary factor. II. Facultative apomixis and selfing. *Genetics*, 111(3):635–653, 1985. doi: 10.1093/genetics/111.3.635.

Krone S. M. and Neuhauser C. Ancestral processes with selection. *Theoretical Population Biology*, 51(3):210–237, 1997. ISSN 0040-5809. doi: https://doi.org/10.1006/tpbi.1997.1299. URL https://www.sciencedirect.com/science/article/pii/S0040580997912995.

Lande R. and Schemske D. W. The evolution of self-fertilization and inbreeding depression in plants. I. Genetical models. *Evolution*, 39(1):24–40, 1985. doi: 10.1111/j.1558-5646.1985.tb04077.x.

Linder M. Common ancestors in a generalized moran model, 2009. URL https://www.diva-portal.org/smash/get/diva2:310019/FULLTEXT01.pdf.

Lloyd D. G. Some reproductive factors affecting the selection of self-fertilization in plants. *The American Naturalist*, 113(1):67–79, 1979. doi: 10.1086/283365.

Mano S. Duality, ancestral and diffusion processes in models with selection. *Theoretical Population Biology*, 75(2):164–175, 2009. doi: 10.1016/j.tpb.2009.01.007.

Maynard Smith J. The origin and maintenance of sex. In Williams G. C., editor, *Group Selection*, pages 163–175. Aldine Atherton, Chicago, USA, 1971.

McClure N. S. and Whitlock M. C. Multilocus estimation of selfing and its heritability. *Heredity*, 109(3):173–179, 2012. doi: 10.1038/hdy.2012.27.

Möhle M. A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability*, 30(2):493–512, 1998. doi: 10.1239/aap/1035228080.

Möhle M. The concept of duality and applications to Markov processes arising in neutral population genetics models. *Bernoulli*, 5:761–777, 1999.

Möhle M. and Notohara M. An extension of a convergence theorem for Markov chains arising in population genetics. *Journal of Applied Probability*, 53(3):953–956, 2016. doi: 10.1017/jpr.2016.5.

Moran P. A. P. Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54(1):60–71, 1958. doi: 10.1017/S0305004100033193.

Moran P. A. P. *Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford, 1962.

Nagylaki T. A model for the evolution of self-fertilization and vegetative reproduction. *Journal of Theoretical Biology*, 58(1):55–58, 1976. doi: 10.1016/0022-5193(76)90138-7.

Neuhauser C. and Krone S. M. The genealogy of samples in models with selection. *Genetics*, 145 (2):519–534, 1997. doi: 10.1093/genetics/145.2.519.

Nordborg M. Structured coalescent processes on different time scales. *Genetics*, 146(4):1501–1514, 1997. doi: 10.1093/genetics/146.4.1501.

Nordborg M. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics*, 154(2):923–929, 2000. doi: 10.1093/genetics/154.2.923.

Nordborg M. and Donnelly P. The coalescent process with selfing. *Genetics*, 146(3):1185–1195, 1997. doi: 10.1093/genetics/146.3.1185.

Notohara M. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, 29(1):59–75, 1990. doi: 10.1007/BF00173909.

Olsen K. C., Ryan W. H., Kosman E. T., Moscoso J. A., Levitan D. R., and Winn A. A. Lessons from the study of plant mating systems for exploring the causes and consequences of inbreeding in marine invertebrates. *Marine Biology*, 168(3):39, 2021. doi: 10.1007/s00227-021-03838-7.

Ornduff R. Reproductive biology in relation to systematics. *Taxon*, 18(2):121–133, 1969. doi: 10.2307/1218671.

Pollak E. On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics*, 117 (2):353–360, 1987. doi: 10.1093/genetics/117.2.353.

Redelings B. D., Kumagai S., Tatarenkov A., Wang L., Sakai A. K., Weller S. G., Culley T. M., Avise J. C., and Uyenoyama M. K. A Bayesian approach to inferring rates of selfing and locus-specific mutation. *Genetics*, 201(3):1171–1188, 2015. doi: 10.1534/genetics.115.179093.

Sasson D. A. and Ryan J. F. A reconstruction of sexual modes throughout animal evolution. *BMC Evolutionary Biology*, 17(1):242, 2017. doi: 10.1186/s12862-017-1071-3.

Schemske D. W. and Lande R. The evolution of self-fertilization and inbreeding depression in plants. II. Empirical observations. *Evolution*, 39(1):41–52, 1985. doi: 10.1111/j.1558-5646.1985.tb04078.x.

Sellinger T. P. P., Abu Awad D., Moest M., and Tellier A. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4):1–28, 2020. doi: 10.1371/journal.pgen.1008698.

Sicard A. and Lenhard M. The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Annals of Botany*, 107(9):1433–1443, 2011. doi: 10.1093/aob/mcr023.

Stebbins G. L. Self fertilization and population variability in the higher plants. *The American Naturalist*, 91(861):337–354, 1957. doi: 10.1086/281999.

Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460, 10 1983. ISSN 1943-2631. doi: 10.1093/genetics/105.2.437. URL `https://doi.org/10.1093/genetics/105.2.437`.

Takahata N. The coalescent in two partially isolated diffusion populations. *Genetics Research*, 52 (3):213–222, 1988. doi: 10.1017/S0016672300027683.

Teterina A. A., Willis J. H., Lukac M., Jovelin R., Cutter A. D., and Phillips P. C. Genomic diversity landscapes in outcrossing and selfing Caenorhabditis nematodes. *PLOS Genetics*, 19 (8):1–38, 2023. doi: 10.1371/journal.pgen.1010879.

Tyukin A. Quenched limits of coalescents in fixed pedigrees. Master's thesis, Johannes-Gutenberg-Universität Mainz, Germany, 2015. URL `https://www.glk.uni-mainz.de/files/2018/08/andrey_tyukin_msc.pdf`.

Vitalis R. and Couvet D. Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. *Genetics Research*, 77(1):67–81, 2001. doi: 10.1017/S0016672300004833.

Vogler D. W. and Kalisz S. Sex among the flowers: Th distribution of plant mating systems. *Evolution*, 55(1):202–204, 2001. doi: 10.1111/j.0014-3820.2001.tb01285.x.

Wakeley J., King L., Low B., and Ramachandran S. Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent. *Genetics*, 190:1433–45, 01 2012. doi: 10.1534/genetics.111.135574.

Wakeley J., King L., and Wilton P. R. Effects of the population pedigree on genetic signatures of historical demographic events. *Proceedings of the National Academy of Sciences*, 113(29):7994–8001, 2016. doi: 10.1073/pnas.1601080113. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1601080113`.

Wang J., El-Kassaby Y. A., and Ritland K. Estimating selfing rates from reconstructed pedigrees using multilocus genotype data. *Molecular Ecology*, 21(1):100–116, 2012. doi: 10.1111/j.1365-294X.2011.05373.x.

Weir B. S. and Cockerham C. C. Mixed self and random mating at two loci. *Genetical Research*, 21(3):247–262, 1973. doi: 10.1017/S0016672300013446.

Wells H. Self-fertilization: Advantageous or deleterious? *Evolution*, 33(1):252–255, 1979. doi: 10.1111/j.1558-5646.1979.tb04679.x.

Wilson I. J. and Dawson K. J. A Markov chain monte carlo strategy for sampling from the joint posterior distribution of pedigrees and population parameters under a Fisher–Wright model with partial selfing. *Theoretical Population Biology*, 72(3):436–458, 2007. doi: 10.1016/j.tpb.2007.03.002.

Wilton P. R., Baduel P., Landon M. M., and Wakeley J. Population structure and coalescence in pedigrees: Comparisons to the structured coalescent and a framework for inference. *Theoretical Population Biology*, 115:1–12, 2017. doi: 10.1016/j.tpb.2017.01.004.

Wolfram Research, Inc. Mathematica, Version 14.0. URL `https://www.wolfram.com/mathematica`. Champaign, IL, 2024.

Wright S. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931. doi: 10.1093/genetics/16.2.97.

Wright S. The genetical structure of populations. *Annals of Eugenics*, 15:323–354, 1951. doi: 0.1111/j.1469-1809.1949.tb02451.x.

Wright S. I., Kalisz S., and Slotte T. Evolutionary consequences of self-fertilization in plants. *Proceedings of the Royal Society B: Biological Sciences*, 280(1760):20130133, 2013. doi: 10.1098/rspb.2013.0133.

Yadav V., Sun S., and Heitman J. On the evolution of variation in sexual reproduction through the prism of eukaryotic microbes. *Proceedings of the National Academy of Sciences*, 120(10): e2219120120, 2023. doi: 10.1073/pnas.2219120120.

10 Appendix

In this section, we list some notation used in this paper, and prove Theorem 3.1 for completeness.

10.1 Notation

- $\mathbb{R}_+ = \mathbb{R}_+$
- $\mathbb{Z}_+ = \{0, 1, 2, 3, ...\}$ is the set of all non-negative integers
- $\mathbb{Z}_{>0} = \mathbb{N} = \{1, 2, 3, 4, ...\}$ is the set of all positive integers
- $N \in \mathbb{Z}_{>0}$ is the population size (i.e. total number of individuals)
- $\xrightarrow{d}$ denotes convergence in distribution
- $I = I_N = \{1, 2 \cdots, N\}$ is the set of the labels of the individuals.
- $J = J_N = \{1, 2, \cdots, 2N-1, 2N\}$ is the set of the labels of the genes, where individual i has genes $\{2i-1, 2i\}$ for $i \in I_N$.

10.2 Metric d on $\mathcal{P}_{m,2}$

We define a metric d on $\mathcal{P}_{m,2}$ so that $(\mathcal{P}_{m,2}, d)$ is a Polish space. We first fix a sequence of metrics d_k on $\mathcal{P}^{(k)}_{m,2}$. Suppose x, y in $\mathcal{P}^{(k)}_{m,2}$ are equal to $\{(p_i, l_i)\}$ and $\{(q_i, m_i)\}$ for non-decreasing l_i, m_i, respectively. Finally, we define d_k on $\mathcal{P}^{(k)}_{m,2}$ by

$$d_k(x, y) := \min\{1, \sum_{i=1}^{k} |p_i - q_i| + |l_i - m_i|\} \tag{66}$$

and $(\mathcal{P}^{(k)}_{m,2}, d_k)$ is also a complete metric space. Define then, for x, y in $\mathcal{P}_{m,2}$ the metric

$$d(x, y) := \begin{cases} 1, & \text{if } |x| \neq |y| \\ d_k(x, y), & \text{if } |x| = |y| \end{cases}.$$

If x_n is Cauchy in $\mathcal{P}_{m,2}$, it must eventually be Cauchy inside $\mathcal{P}^{(k)}_{m,2}$ for some $k \in \mathbb{Z}_+$. Completeness follows therefore from completeness of $(\mathcal{P}^{(k)}_{m,2}, d_k)$. Separability is similarly immediate.

10.3 Proofs for the unconditional distribution of $\tau^{(N,2)}$

To give a rigorous proof to Theorem 3.1, we define the timings of overlap events. Set $\tau^O_0 := 0$ and, for $i \geq 1$,

$$\tau^O_i := \inf\{k > \tau^O_{i-1} : \hat{X}_k = \hat{Y}_k, \hat{X}_{k-1} \neq \hat{Y}_{k-1}\}$$

be the time-step of the i-th overlap event in the past. Next, we define the timings of splitting events. We let $\tau^D_1 := \inf\{k \in \mathbb{Z}_+ : \hat{X}_k \neq \hat{Y}_k\}$, which is zero under $\mathbb{P}_{\text{diff}}$ and is the time of the first splitting event under $\mathbb{P}_{\text{same}}$. For $i \geq 2$, we let

$$\tau^D_i := \inf\{k > \tau^D_{i-1} : \hat{X}_k \neq \hat{Y}_k, \hat{X}_{k-1} = \hat{Y}_{k-1}\}$$

be the time-step of the i-th splitting event in the past. By convention, $\inf \emptyset = \infty$.

We decompose $\tau^{(N,2)}$ as in Figure 9 and describe the distribution of $\mathcal{O}$ in the following lemma. For simplicity we write $X \sim Geom(r)$ when a random variable X is a geometric random variable with parameter r, that is, when $\mathbb{P}(X = m) = (1-r)^{m-1}r$ for $m \in \mathbb{Z}_{>0}$.

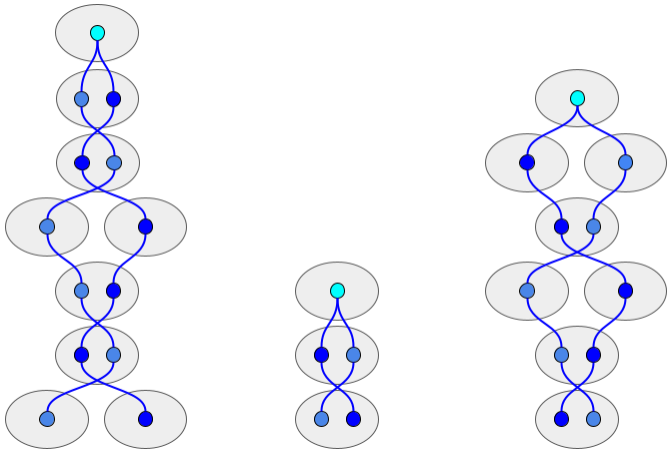


Figure 9: Here we see three different realizations of $\tau^{(N,2)}$. The left realization is under $\mathbb{P}_{\text{diff}}$ with $\mathcal{O} = 2$. The middle realization is under $\mathbb{P}_{\text{same}}$ with $\mathcal{O} = 0$. The right realization is under $\mathbb{P}_{\text{same}}$ with $\mathcal{O} = 2$. Note that the initial state makes no contribution to $\mathcal{O}$ i.e. under the $\mathbb{P}_{\text{same}}$ sample $\mathcal{O}$ may be zero.

Lemma 10.1. *Under both $\mathbb{P}_{\text{diff}}$ and $\mathbb{P}_{\text{same}}$, it holds that*

$$\tau^{(N,2)} = (\tau^{(N,2)} - \tau^O_{\mathcal{O}}) + \sum_{i=1}^{\mathcal{O}} (\tau^O_i - \tau^D_{i-1}) + \sum_{i=1}^{\mathcal{O}-1} (\tau^D_i - \tau^O_{i-1}) \quad \text{on the event } \{\mathcal{O} \geq 1\} \tag{67}$$

and that conditional on the event $\{\mathcal{O} \geq 1\}$, $\mathcal{O} \sim Geom\left(\frac{1}{2-\alpha_N}\right)$. Under $\mathbb{P}_{\text{same}}$, the two lineages coalesce before splitting on the event $\{\mathcal{O} = 0\}$. Furthermore, (6) *holds.*

Proof. (67) holds because on the event $\{\mathcal{O} \geq 1\}$, it holds that

$$0 = \tau^D_1 < \tau^O_1 < \tau^D_2 < \tau^O_2 < \cdots < \tau^O_{\mathcal{O}} \leq \tau^{(N,2)} \quad \mathbb{P}_{\text{diff}} - a.s.$$

and

$$0 = \tau^O_0 < \tau^D_1 < \tau^O_1 < \tau^D_2 < \tau^O_2 < \cdots < \tau^O_{\mathcal{O}} \leq \tau^{(N,2)} \quad \mathbb{P}_{\text{same}} - a.s.$$

The right hand side of (67) is a telescoping sum.

Under $\mathbb{P}_{same}$, the two lineages coalesces before splitting on the event $\{\mathcal{O} = 0\} = \{\tau^D_1 = \infty\}$ because $\tau^{(N,2)} < \infty$ almost surely.

That $\mathbb{P}_{\text{diff}}(\mathcal{O} = 0) = 0$ follows simply from the fact that the two lineages cannot coalesce without belonging to the same individual, whether at coalescence time or not.

We now prove that

$$\mathbb{P}_{\text{same}}(\mathcal{O} = 0) = \frac{\alpha_N}{2 - \alpha_N}.$$

For each selfing event in the occupied individual before a split, there is a $\frac{1}{2}$ chance of coalescing. In particular, if there are k selfing events before a split then there is a

$$\frac{1}{2} + (\frac{1}{2})^2 + ... + (\frac{1}{2})^k = 1 - 2^{-k}$$

chance of coalescing during those selfing events. Now let U denote the number of selfing events before a split event. We can see that the probability that we have a selfing event before a splitting event in this multiply-occupied individual is

$$\frac{\alpha_N N^{-1}}{\alpha_N N^{-1} + (1-\alpha_N)N^{-1}} = \alpha_N.$$

Therefore

$$\mathbb{P}(U = k) = \alpha_N^k (1-\alpha_N).$$

for $k > 0$, and $1 - \alpha_N$ for $k = 0$. It follows that the probability of coalescing before splitting is

$$\mathbb{E}\left[1 - 2^{-U}\right] = \frac{\alpha_N}{2-\alpha_N}.$$

The proof of (6) is complete.

It remains to prove, therefore, that, conditional on the event $\{\mathcal{O} \geq 1\}$, that $\mathcal{O} \sim Geom(\frac{1}{2-\alpha_N})$. In each overlap event, it is equally as likely that we coalesce instantly as we do not. However, if we have not coalesced instantly, we have two sample lineages in the same individual. The probability that these two sample lineages coalesce before a splitting event is given by the (6). Therefore the probability that an overlap event is the final overlap is

$$\frac{1}{2} + \frac{1}{2}\frac{\alpha_N}{2-\alpha_N} = \frac{1}{2-\alpha_N}.$$

This gives the claim. □

Now that we have characterized the distribution of $\mathcal{O}$, we characterize the distribution of each term in the decomposition (67) in lemmas below.

Lemma 10.2. *Under each of $\mathbb{P}_{\text{diff}}$ and $\mathbb{P}_{\text{same}}(\cdot \,|\, \mathcal{O} \geq 1)$, the $2\mathcal{O}$ terms on the right of* (67) *are independent random variables with the following distributions:*

- $\{\tau_i^D - \tau_{i-1}^O\}_{i=1}^{\mathcal{O}}$ *are geometric random variables with parameter* $\frac{2-\alpha_N}{2N}$
- $\{\tau_i^O - \tau_{i-1}^D\}_{i=1}^{\mathcal{O}-1}$ *are geometric random variables with parameter* $2N^{-2}$
- $\tau^{(N,2)} - \tau_{\mathcal{O}}^O$ *is* 0 *with probability* $1 - \frac{1}{2}\alpha_N$ *and conditioned on not being zero is a geometric random variable with parameter* $\frac{\alpha_N(2-\alpha_N)}{2N}$

Proof. We begin by analyzing the splitting gap times $\tau_i^D - \tau_{i-1}^O$. Note by (6) there is a $1 - \frac{\alpha_N}{2-\alpha_N} = \frac{2(1-\alpha_N)}{2-\alpha_N}$ chance of splitting before a selfing coalescence. The odds that we split in the next step given the two sample lineages do not coalesce via selfing before a split event is therefore

$$(1-\alpha_N)N^{-1}\frac{2-\alpha_N}{2(1-\alpha_N)} = \frac{2-\alpha_N}{2N}.$$

We can see then that $\tau_i^D - \tau_{i-1}^O$ is geometric with this same rate.

Now we may characterize the gap between the final overlap and coalescence, $\tau^{(N,2)} - \tau_{\mathcal{O}}^O$.Given that we have reached the final overlap, $\tau^{(N,2)} - \tau_{\mathcal{O}}^O$ is 0 if coalescence occurred at the overlap.

There is, unconditioned, a $\frac{1}{2}$ chance of this occurring. There is, unconditioned, a $\frac{1}{2}\frac{\alpha_N}{2-\alpha_N}$ chance of coalescing due to selfing. This gives the conditioned probability that $\tau^{(N,2)} - \tau^O_{\mathcal{O}}$ is 0 is

$$\frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}\frac{\alpha_N}{2-\alpha_N}} = 1 - \frac{1}{2}\alpha_N.$$

If $\tau^{(N,2)} - \tau^O_{\mathcal{O}} \neq 0$ then the final coalescence was due to a selfing event that occurred before splitting. The probability that this coalescence via selfing occurs in a single generation given that this is the last overlap is

$$\frac{\alpha_N N^{-1}\frac{1}{2}}{\frac{1}{2-\alpha_N}} = \frac{\alpha_N(2-\alpha_N)}{2N}.$$

This completes the characterization of $\tau^{(N,2)} - \tau^O_{\mathcal{O}}$.

It remains simply to characterize $\tau^O_i - \tau^D_{i-1}$. As it is equally likely that one coalesces as that one overlaps without coalescing, and as the odds of coalescing in a single time-step when we have two sample lineages in two distinct individuals is N^{-2}, in each time-step there is a $2N^{-2}$ chance of an overlap. Therefore, $\tau^O_i - \tau^D_{i-1}$ is geometric with parameter $2N^{-2}$. This completes the lemma. □

Lemma 10.3. *Take the decomposition given by Lemma 10.1. Suppose $\alpha_N \to \alpha \in [0,1]$ as N goes to infinity. The under each of $\mathbb{P}_{\text{diff}}$ and $\mathbb{P}_{\text{same}}(\cdot|\mathcal{O} \geq 1)$, the $2\mathcal{O}$ terms on the right of (67) converge with the N^{-2} rescaling as follows:*

- $N^{-2}(\tau^{(N,2)} - \tau^O_{\mathcal{O}})$ *converges to* 0 *in distribution*
- $\{N^{-2}(\tau^D_i - \tau^O_{i-1})\}_{i=1}^{\mathcal{O}}$ *converge in distribution to* 0
- $\{N^{-2}\left(\tau^O_i - \tau^D_{i-1}\right)\}_{i=1}^{\mathcal{O}}$ *converge to independent exponential random variables with rate* 2.

Proof. Notice that $N^{-2}(\tau^D_i - \tau^O_i)$ and $N^{-2}(\tau^{(N,2)} - \tau^O_{\mathcal{O}})$ are non-negative random variables so it suffices to see their mean go to zero. By Lemma 10.2

$$\mathbb{E}_{\text{diff}}\left[N^{-2}(\tau^D_i - \tau^O_i)\right] = \frac{1}{N(N-1)}\frac{2N}{2-\alpha_N} = \frac{2}{2-\alpha_N}\frac{1}{N-1}.$$

This converges to 0 as N goes to infinity. Similarly, by Lemma 10.2, we have

$$\mathbb{E}_{\text{diff}}\left[N^{-2}(\tau^{(N,2)} - \tau^O_{\mathcal{O}})\right] = \frac{1}{N(N-1)}\frac{\alpha_N}{2}\frac{2N}{\alpha_N(2-\alpha_N)} = \frac{1}{2-\alpha_N}\frac{1}{N-1}.$$

This similarly goes to 0 as N goes to infinity. The $\{\tau^O_i - \tau^D_{i-1}\}$ are geometric with parameter $2N^{-2}$ under $\mathbb{P}_{\text{diff}}$. The N^{-2} time-rescaling gives an exponential random variable with rate 2.

The result holds equivalently under $\mathbb{P}_{\text{same}}(\,\cdot\,|\mathcal{O} \geq 1]$ as the distributions given by Lemma 10.2 are the same under both laws. □

In particular, as $\lambda_N Geom(\lambda_N)$ converges to a unit rate exponential random variable when $\lambda_N \to 0$, and as the splitting times coalesce instantaneously with the time rescaling, we expect that $N^{-2}\tau^{(N,2)}$ should converge to a geometric sum of exponential random variables, which is an exponential random variable. Indeed this is the case, as one can see in Theorem 3.1.

Proof. [Proof of Theorem 3.1] The distribution of $\tau^{(N,2)}$ is determined in the decomposition (67) and the Lemma 10.2.

Let $\varphi_N, \psi_N, \sigma_N$ denote the characteristic functions of the $\tau_i^O - \tau_{i-1}^D, \tau_i^D - \tau_i^O$, and $\tau^{(N,2)} - \tau_{\mathcal{O}}^O$, respectively. Then

$$\mathbb{E}_{\text{diff}}\left[e^{itN^{-2}\tau}\right] = \sigma_N(tN^{-2})\sum_{j=0}^{\infty}\mathbb{P}_{\text{diff}}(\mathcal{O}=j)\varphi_N(tN^{-2})^j\psi_N(tN^{-2})^j$$

This is, where $r_N = \frac{1}{2-\alpha_N}$,

$$\sigma_N(tN^{-2})\sum_{j=1}^{\infty} r_N(1-r_N)^{j-1}\varphi_N(tN^{-2})^j\psi_N(tN^{-2})^j = \sigma_N(tN^{-2})r_N\frac{\varphi_N(tN^{-2})\psi_N(tN^{-2})}{1-(1-r_N)\varphi_N(tN^{-2})\psi_N(tN^{-2})}. \tag{68}$$

By Lemma 10.3 we have $\psi_N(tN^{-2})$ and $\sigma_N(tN^{-2})$ converge to 1 and $\varphi_N(tN^{-2})$ converges $(1+it2^{-1})^{-1}$. As α_N converges to α as N goes to infinity, r_N converges to $r = \frac{1}{2-\alpha}$. Therefore

$$\sigma_N(tN^{-2})r_N\frac{\varphi_N(tN^{-2})\psi_N(tN^{-2})}{1-(1-r_N)\varphi_N(tN^{-2})\psi_N(tN^{-2})} \to r\frac{\varphi(t)}{1-(1-r)\varphi(t)} = (1+it(\frac{2}{2-\alpha})^{-1})^{-1}. \tag{69}$$

This is the claim for $\mathbb{P}_{\text{diff}}$. The result holds equivalently under $\mathbb{P}_{\text{same}}(\cdot|\mathcal{O}\geq 1)$ as the distributions given by Lemma 10.3 are the same under both laws. □

Corollary 10.4. *If $\alpha_N \to \alpha \in [0,1]$, then $N^{-2}\tau^{(N,2)}$ converges, under the law $\mathbb{P}_{\text{same}}(\cdot)$ to a random variable that is 0 when conditioned on $\{\mathcal{O}=0\}$, and that conditioned on $\{\mathcal{O}\geq 1\}$ is exponential with rate $\frac{2}{2-\alpha}$.*

Proof. By (6), the two lineages coalesce instantaneously with probability $\frac{\alpha_N}{2-\alpha_N}$, else it splits instantaneously to two lineages in two distinct individuals. The result thus follows from the Markov property, $\alpha_N \to \alpha$, and Theorem 3.1. □