Site Frequency Spectrum of the Bolthausen-Sznitman Coalescent

Götz Kersting[1], Arno Siri-Jégousse[2], and Alejandro H. Wences[2]

[1]Goethe Universität, Institut für Mathematik, Frankfurt am Main, Germany.
[2]UNAM, IIMAS, Departamento de Probabilidad y Estadística, Mexico.

Abstract

We derive explicit formulas for the two first moments of he site frequency spectrum $(SFS_{n,b})_{1\leq b\leq n-1}$ of the Bolthausen-Sznitman coalescent along with some precise and efficient approximations, even for small sample sizes n. These results provide new L_2-asymptotics for some values of $b = o(n)$. We also study the length of internal branches carrying $b > n/2$ individuals. In this case we obtain the distribution function and a convergence in law. Our results rely on the random recursive tree construction of the Bolthausen-Sznitman coalescent.

Keywords: Bolthausen-Sznitman coalescent, Site frequency spectrum, Random recursive tree, Branch lengths.
MSC2010: 60J75 (primary), 60C05, 60G09, 60F25, 92D10.

1 Introduction

The Bolthausen-Sznitman coalescent is an exchangeable coalescent with multiple collisions that has recently gained attention in the theoretical population genetics literature. It has been described as the limit process of the genealogies of different population evolution models, including models that contemplate the effect of natural selection

[15, 16]. It has also been proposed as a new null model for the genealogies of rapidly adapting populations, such as pathogen microbial populations, and other populations that show departures from Kingman's null model [1, 13].

A measure of the genetic diversity in a present day sample of a population is often used in population genetics in order to infer its evolutionary past and the forces at play in its dynamics. The *Site Frequency Spectrum* (SFS) is a well known theoretical model of the genetic diversity present in a population, it assumes that neutral mutations arrive to the population as a Poisson Process and that each arriving mutation falls in a different site of the genome (infinite sites model), in contrast to the *Allele Frequency Spectrum* in which mutations are assumed to fall on the same site but create a new allele every time (infinite alleles model). Given the close relation between the Site Frequency Spectrum and the whole structure of the underlying genealogical tree, it can be used as a model selection tool for the evolutionary dynamics of a population [3, 10, 4].

In this work we give explicit expressions of the first and second moments for the whole Site Frequency Spectrum $(SFS_{n,b})_{1\leq b<n}$ of the Bolthausen-Sznitman coalescent, which to our knowledge were only known for Kingman's coalescent until now [5]. Here $SFS_{n,b}$ denotes the number of mutations shared by b individuals in the sample of size n. For the expectation we obtain the formula

$$\mathbb{E}\left[SFS_{n,b}\right] = \theta n \int_0^1 \frac{\Gamma(b-p)}{\Gamma(b+1)} \frac{\Gamma(n-b+p)}{\Gamma(n-b+1)} \frac{dp}{\Gamma(1-p)\Gamma(1+p)},$$

where θ denotes the mutation rate. For larger values of n there might occur problems in the calculation of this integral due to the exorbitant growth of the Gamma function. Also this formula allows no insight into the shape of the expected site frequency spectrum. For this purpose approximations are helpful. A first approximation, resting on Stirling's formula, reads for $2 \leq b \leq n-1$

$$\mathbb{E}[SFS_{n,b}] \approx \frac{\theta}{n-1} \frac{b-1}{b} f_1\left(\frac{b-1}{n-1}\right) \tag{1}$$

where f_1 is a convex, non-monotone function on $(0,1)$ defined by

$$f_1(u) := \int_0^1 u^{-p-1}(1-u)^{p-1} \frac{\sin(\pi p)}{\pi p}\, dp\ . \tag{2}$$

We remark that this integral may be reduced to the (complex) exponential integral $Ei(\cdot)$. These formulas show that the shape of the Site Frequency Spectrum, restricted to the range $2 \leq b < n$, is explained essentially by one function not depending on the population size n. Also our approximations update those given in [13] for the case of families with frequencies close to 0 and 1, since we have $f_1(u) \sim (u \log u)^{-2}$ close to 0 and $f_1(u) \sim ((u-1)\log(1-u))^{-1}$ close to 1, see equations (30) and (31) below. The case $b = 1$ is not covered by (1), it has to be treated separately, which reflects the dominance of external branches in the Bolthausen-Sznitman coalescent. See Theorem 3.4 for a complete summary.

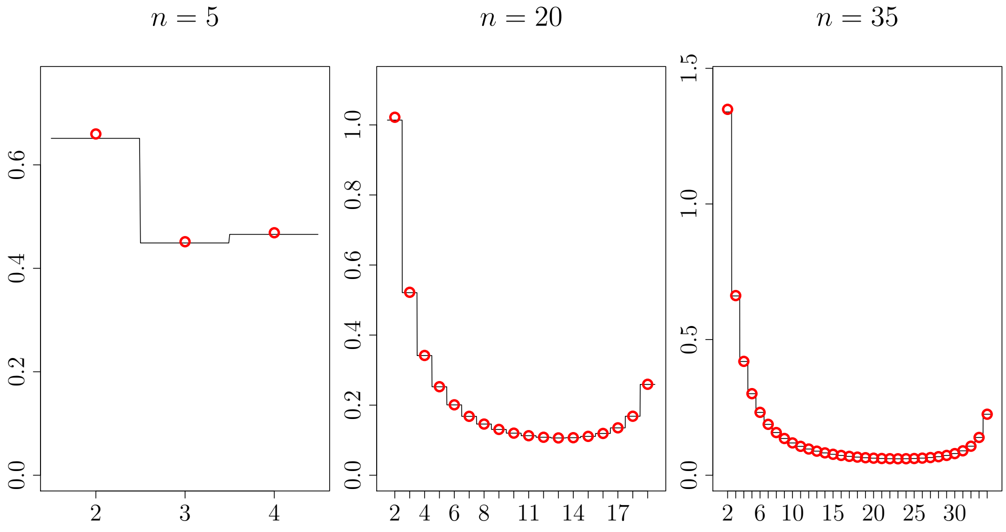

Figure 1: Comparison of exact and approximated values of $\mathbb{E}[SFS_{n,b}]$, red circles present the exact values for $b = 2$ to $n - 1$, and the black lines their refined approximations (3).

The above approximation is accurate also from a numerical point of view. Only for $b = 2$ we encounter an enlarged relative error which anyhow remains less than 10 percent for $n \geq 8$. If a more precise result is desired then the following refined approximation may be applied for $2 \leq b \leq n$:

(3)

$$\mathbb{E}\left[SFS_{n,b}\right] \approx \theta n \frac{b-1}{b}\left(\frac{1}{(n-1)^2} f_1\left(\frac{b-1}{n-1}\right) - \frac{1}{(n-1)^3} g_1\left(\frac{b-1}{n-1}\right)\right),$$

with a positive function g_1 on $(0,1)$ given by

$$(4) \qquad g_1(u) := \frac{1}{2u^2(1-u)^2} \frac{\pi^2 + \log^2 \frac{1-u}{u} + \frac{2}{u} \log \frac{1-u}{u}}{\left(\pi^2 + \log^2 \frac{1-u}{u}\right)^2}.$$

With this formula we have a relative error remaining below 1 percent for $b = 2$ and $n \geq 10$, below 0.5 percent for $b = 2$ and $n \geq 150$, and below 0.3 percent for $b \geq 3$ and $n \geq 10$. Thus this approximation appears well-suited for practical purpose. Figure 1 illustrates its precision in the cases $n = 5$, 20, 35 and $\theta = 1$.

For $b = 1$ the approximation formula corresponding to (1) reads

$$\begin{aligned} \mathbb{E}\left[SFS_{n,1}\right] &= \theta n \int_0^1 \frac{\Gamma(n-1+p)}{\Gamma(n)} \frac{dp}{\Gamma(1+p)} \\ &\approx \theta n \int_0^1 (n-1)^{p-1} \frac{dp}{\Gamma(1+p)}, \end{aligned}$$

which is an immediate consequence of Stirling's approximation. It is precise for small n and requires no further correction as in the case $b \geq 2$.

We also study the asymptotic behavior of the second moments which, together with the above asymptotics for the first moment, leads to the following L^2 convergences:

$$\frac{\log n}{n} SFS_{n,1} \to \theta,$$

and, whenever $b \geq 2$ and $b = o\left(\sqrt{n}/\log n\right)$,

$$\frac{b(b-1)\log^2(n/b)}{n} SFS_{n,b} \to \theta.$$

These generalize and strengthen the results in [2] for the Bolthausen-Sznitman coalescent.

We also provide the joint distribution function of the branch lengths of large families, i.e families of size at least half the total population size, and their marginal distribution function. These results are useful to obtain the marginal distribution function of the Site Frequency Spectrum and a sampling formula for the half of the vector corresponding to large family sizes, although we do not present such tedious computations here.

Asymptotic results for related functionals on the Bolthausen-Sznitman coalescent have been derived by studying the block count chain of the coalescent through a coupling with a random walk as in [8] and [9], where asymptotics for the total number of jumps, and the total, internal, and external branch lengths of the Bolthausen-Sznitman coalescent are described; these results give the asymptotic behaviour of the total number of mutations present in the population, the number of mutations present in a single individual, and the number of mutations present in at least 2 individuals. Also, a Markov chain approximation of the initial steps of the process was developed in [2] where asymptotics for the total tree length and the Site Frequency Spectrum of small families were derived for a class of Λ-coalescents containing the Bolthausen-Sznitman coalescnet.

Progress has also been made for the finite coalescent even for the general coalescent process. The finite Bolthausen-Sznitman coalescent has been studied through the spectral decomposition of its jump rate matrix described in [11] where the authors used it to derive explicit expressions for the transition probabilities and the Green's matrix of this coalescent, and also the Kingman coalescent. The spectral decomposition of the jump rate matrix of a general coalescent, including coalescents with multiple mergers, is also used in [17] where an expression for the expected Site Frequency Spectrum is given in terms of matrix operations which in the case of the Bolthausen-Sznitman coalescent result in an algorithm requiring on the order of n^2 computations. In [7] another expression in terms of matrix operations is given for this and other functionals on general coalescent processes, both in expected value (and higher moments) and in distribution; these expressions however are deduced from the theory of phase-type distributions, in particular distributions of rewards constructed on top of coalescent processes, and also require vast computations for large population sizes.

Our method, mainly based on the Random Recursive Tree construction of the Bolthausen-Sznitman coalescent given in [6], gives easy-to-compute expressions for the first and second moments of the Site Frequency Spectrum of this particular coalescent. This combinatorial construction not only allows us to study the bottom but also the top of the tree thus providing an additional insight into the past of the population and large families, both asymptotically and for any fixed population size.

In Section 2 we layout the basic intuitions that compose the bulk of our method, including the Random Recursive Tree construction of the Bolthausen-Sznitman coalescent and the derivation of the first moment of the Site Frequency Spectrum for the infinite coalescent as a first application (Corollary 2.2). In Section 3 we present our results on the first and second moments of the branch lengths (Theorem 3.1) and of the Site Frequency Spectrum (Corollary 3.2) for any fixed family size and initial population. We then use these expressions to obtain asymptotic approximations of these moments as the initial population goes to infinity (Theorems 3.4 and 3.5) which lead to L^2 convergence results on the SFS (Corollary 3.6). In Section 4 we restrict ourselves to the case of large family sizes and present the joint and marginal distribution functions of their branch lengths (Theorems 4.1 and 4.3), along with a limit in law result (Corollary 4.2). Section 5 provides explanations for approximations (1) and (3). Finally, in Sections 6 and 7 we provide detailed proofs of our results.

2 Preliminaries

Consider the Bolthausen-Sznitman coalescent $(\Pi^\infty(t))_{t\geq 0}$ with values in $\mathscr{P}_\infty$, the space of partitions of $\mathbb{N}$, and the ranked coalescent $(|\Pi^\infty(t)|^\downarrow)_{t\geq 0}$, with values in the space of mass partitions $\mathscr{P}_{[0,1]}$, made of the asymptotic frequencies of $\Pi^\infty(t)$ reordered in a non-increasing way. In what follows we present the Random Recursive Tree (RRT) construction of the Bolthausen-Sznitman coalescent given by Goldschmidt and Martin in [6]; then we follow the argument given in the same paper to establish that

$$|\Pi^\infty(t)|^\downarrow \overset{d}{=} PD(e^{-t}, 0), \tag{5}$$

where $PD(\alpha, \theta)$ is the $(\alpha, \theta)-$Poisson-Dirichlet distribution.

Briefly, the construction of the Bolthausen-Sznitman coalescent in terms of Random Recursive Trees proceeds as follows. We work on the set of recursive trees whose labeled nodes form a partition π of $[n] := \{1, \dots, n\}$, where the ordering of the nodes that confers the term “recursive” is given by ordering the blocks of π according to their least elements. A cutting-merge procedure is defined on the set of recursive trees of this form with a marked edge, this procedure consists of cutting the marked edge and merging all the labels in the subtree below with the node above, thus creating a new recursive tree whose labels form a new (coarser) partition of $[n]$ (see Figure 2). With this operation in mind we consider a RRT with labels $\{1\}, \cdots, \{n\}$, say T, to which we also attach independent standard exponential variables to each edge. Then, for each time $t > 0$ we retrieve the partition of $[n]$ obtained by performing a cutting-merge procedure on all the edges of T whose exponential variable is less than t. This gives a stochastic process $(\Pi^n(t))_{t\geq 0}$ with values on the set of partitions of $[n]$ that can be proven to be the n-Bolthausen-Sznitman coalescent.

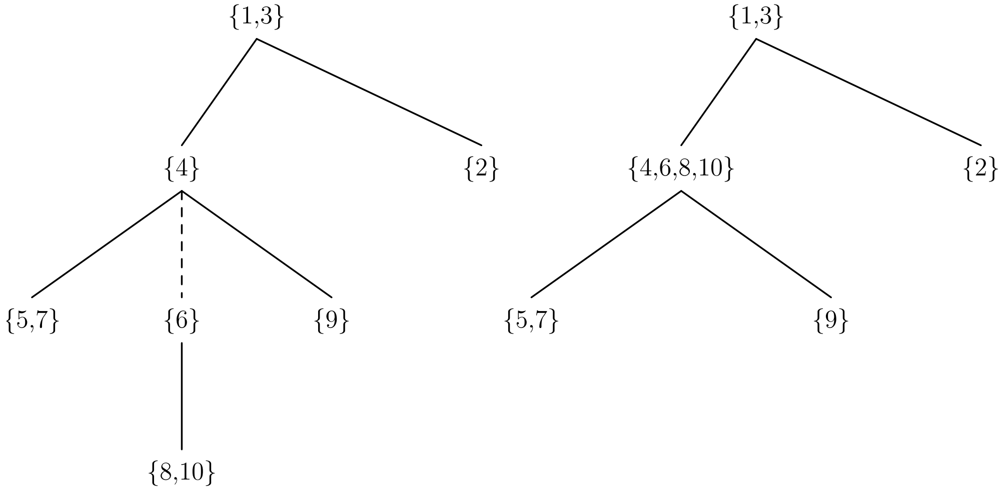

Figure 2: On the left, an example of a recursive tree whose labels constitute a partition of $\{1, \cdots, 10\}$. On the right, the resulting recursive tree after a cutting-merge procedure performed on the marked edge (dashed line) of the first tree.

The fact that $|\Pi^\infty(t)|^{\downarrow} \overset{d}{=} PD(e^{-t}, 0)$ now follows readily. To see this, consider the construction of T where nodes arrive sequentially and each arriving node attaches to any of the previous nodes with equal probability. Considering also their exponential edges and having in mind the cutting-merge procedure we see that for any fixed time t, and assuming that $b-1$ nodes have arrived and formed k blocks of sizes $s_1, \ldots, s_k$ in $\Pi^{b-1}(t)$, the next arriving node, node $\{b\}$, will form a new block in $\Pi^b(t)$ if and only if it attaches to any of the roots of the sub-trees of T that form the said k blocks and if, furthermore, its exponential edge is greater than t; this occurs with probability $\frac{ke^{-t}}{b-1}$. On the other hand, in order for $\{b\}$ to join the jth block of size s_j it must either attach to the root of the sub-tree of T that builds this block and its exponential edge must be less than t, which happens with probability $\frac{1-e^{-t}}{b-1}$, or it must attach to any other node of the said sub-tree, which happens with probability $\frac{s_j-1}{b-1}$; thus, the probability of attaching to the jth block is $\frac{s_j-e^{-t}}{b-1}$. We recognize in these expressions the probabilities that define the Chinese Restaurant Process with parameters $\alpha = e^{-t}$ and $\theta = 0$.

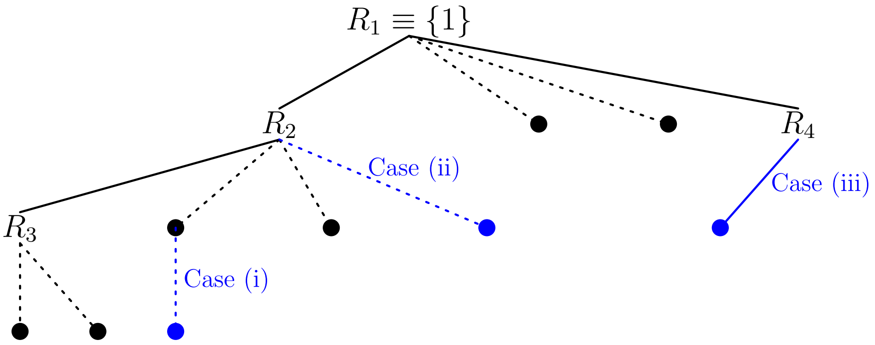

Figure 3: Schematic representation of passing from $\Pi^n(t)$ to $\Pi^{n+1}(t)$ for fixed t, by adding a new node (blue) to a RRT. Solid lines and dotted lines represent edges whose exponential variables are greater than t and less than or equal to t, respectively. In this case at time t there are four subtrees rooted at R_1, R_2, R_3, and R_4, which form the blocks that constitute $\Pi^n(t)$; these blocks are also the tables of a Chinese Restaurant Process. In case (i) the new node will be included in the block formed by R_2 at time t, irrespective of whether its exponential edge is greater than t or not. In case (ii) the new node forms part of the block rooted at R_4 because its exponential edge is less than t. Finally, in case (iii) the new node is a new root of a subtree that will form an additional block of $\Pi^{n+1}(t)$ (i.e. the new node opens a new table in the Chinese Restaurant Process).

We now provide two straightforward applications of the RRT construction described above which nonetheless contain the essential intuitions underlying the forthcoming proofs.

2.1 Site Frequency Spectrum in the infinite coalescent

For the first application consider a subset $I \subset (0,1)$ and define $(C_I(t))_{t\geq 0}$ to be the process of the number of blocks in $\Pi^\infty(t)$ with asymptotic frequencies in I. Then

$$\ell_I := \int_0^\infty C_I(t)\, dt \tag{6}$$

gives the total branch length of families with size frequencies in I in the infinite coalescent.

Our first theorem is a simple corollary of the equality in law (5).

Theorem 2.1. For $I \subset (0,1)$, we have

$$\mathbb{E}[\ell_I] = \int_I \int_0^1 u^{-p-1}(1-u)^{p-1} \frac{\sin(\pi p)}{\pi p}\, dp\, du.$$

In particular, note that if in the infinite sites model with mutation rate θ we define SFS_I to be the number of mutations shared by a proportion u of individuals with u ranging in I, then by conditioning on ℓ_I we get

Corollary 2.2. For $I \subset (0,1)$, we have

$$\mathbb{E}\left[SFS_I\right] = \theta \int_I \int_0^1 u^{-p-1}(1-u)^{p-1} \frac{\sin(\pi p)}{\pi p}\, dp\, du. \tag{7}$$

Proof of Theorem 2.1. Since

$$\mathbb{E}\left[\ell_I\right] = \int_0^\infty \mathbb{E}\left[C_I(t)\right]\, dt$$

it only remains to compute $\mathbb{E}\left[C_I(t)\right]$ and simplify the expressions, but this is a straightforward consequence of Equation (6) in [14] which states that if $\varrho = (a_1, \cdots)$ is $PD(\alpha, \theta)$ distributed, and $f : \mathbb{R} \to \mathbb{R}$ is a function, then

$$\mathbb{E}\left[\sum_{i=1}^{\infty} f(a_i)\right] = \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)\Gamma(1-\alpha)} \int_0^1 f(u) \frac{(1-u)^{\alpha+\theta-1}}{u^{\alpha+1}}\, du. \tag{8}$$

Taking $f(u) = \mathbb{1}_I(u)$we get

$$\mathbb{E}[C_I(t)] = \frac{1}{\Gamma(e^{-t})\Gamma(1-e^{-t})} \int_0^1 \mathbb{1}_I(u) \frac{(1-u)^{e^{-t}-1}}{u^{e^{-t}+1}}\, du.$$

Using Euler's reflection formula, making $p = e^{-t}$ on the above expression and integrating on $[0, \infty)$ we finish the proof. □

2.2 Time to the absorption

In this section we prove a useful lemma for the upcoming proofs, but a first consequence of this lemma gives the distribution function of the time to absorption, A_n, in the n-coalescent, a result already proved in [12].

Here Be stands for the Beta function

$$Be(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

and Ψ for the digamma function

$$\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} = -\gamma - \sum_{n=1}^{\infty} \left(\frac{1}{z+n-1} - \frac{1}{n} \right)$$

where γ stands for the Euler-Mascheroni constant.

Lemma 2.3. Let T be a RRT on a set of n labels and with exponential edges. Define the two functionals $m(T)$ and $M(T)$ that give the minimum and the maximum of the exponential edges attached to the root of T. Then

$$\mathbb{P}(m(T) > s) = \frac{1}{(n-1)Be(n-1, e^{-s})}, \tag{9}$$

and

$$\mathbb{P}(M(T) \leq s) = \frac{1}{(n-1)Be(n-1, 1-e^{-s})}. \tag{10}$$

Also, for independent trees T_1 and T_2 of respective size n_1 and n_2, we have

$$\begin{aligned} &\mathbb{P}(m(T_2) - M(T_1) > s) \\ =&\frac{1}{(n_1-1)(n_2-1)} \int_0^1 \frac{\Psi(n_1-p) - \Psi(1-p)}{Be(n_2-1, e^{-s}p)Be(n_1-1, 1-p)}\, dp. \end{aligned} \tag{11}$$

The proof of (10) follows the same lines as in [12] where the law of the time to absorption of the Bolthausen-Sznitman coalescent is derived, since this time is the maximum of the exponential edges attached to the root of a RRT. That is,

$$\mathbb{P}(A_n \leq s) = \frac{1}{(n-1)Be(n-1, 1-e^{-s})}, \tag{12}$$

and, as $n \to \infty$,

$$A_n - \log\log n \xrightarrow{d} -\log E \tag{13}$$

where E is a standard exponential random variable. The latter convergence in distribution was elegantly proved in [6] using a construction of random recursive trees in continuous time, whereas in this case it follows from Stirling's approximation to the Gamma functions appearing in (12).

On the other hand, the equality (11) will be used in the computation of the distribution function of branch lengths with large family sizes presented in Section 4.

Proof of Lemma 2.3. Let $E_2, \cdots, E_n$ be the exponential edges associated to the nodes of T. For the proof of (9) we consider the event $\{m(T) > s\}$. This event occurs when, in the recursive construction of T along with the exponential edges, the ith node ($2 \le i \le n$) does not attach to $\{1\}$ whenever $E_i < s$; this happens with probability $1 - \frac{1-e^{-s}}{i-1}$. Thus, considering the n nodes, we obtain

$$\begin{aligned}\mathbb{P}(m(T) > s) &= e^{-s}\left(\frac{1+e^{-s}}{2}\right)\cdots\left(\frac{n-2+e^{-s}}{n-1}\right)\\ &= \frac{1}{(n-1)Be(n-1,e^{-s})}.\end{aligned}$$

For (10) we instead build the tree such that the ith node does not attach to $\{1\}$ whenever $E_i > s$; this happens with probability $1 - \frac{e^{-s}}{i-1}$. Thus we obtain

$$\begin{aligned}\mathbb{P}(M(T) \le s) &= (1-e^{-s})\left(\frac{2-e^{-s}}{2}\right)\cdots\left(\frac{n-1-e^{-s}}{n-1}\right)\\ &= \frac{1}{(n-1)Be(n-1,1-e^{-s})}.\end{aligned}$$

Finally we compute

$$\begin{aligned}&\mathbb{P}(m(T_2) - M(T_1) > s)\\ =&\frac{1}{(n_1-1)(n_2-1)}\int_0^\infty \frac{1}{Be(n_2-1,e^{-(s+t)})}\frac{d}{dt}\left(\frac{1}{Be(n_1-1,1-e^{-t})}\right)dt\end{aligned}$$

and by changing the variable $p = e^{-x}$ we obtain (11). □

3 Moments of the Site Frequency Spectrum

By a simple adaptation of our previous notation for branch lengths in the infinite coalescent (C_I and ℓ_I), in the finite case we also define for $1 \leq b \leq n-1$ the process $(C_{n,b}(t))_{t\geq 0}$ and the random variables $(\ell_{n,b})$, where $C_{n,b}(t)$ is the number of blocks of size b in $\Pi^n(t)$, and

$$\ell_{n,b} := \int_0^\infty C_{n,b}(t)\, dt. \tag{14}$$

We now provide explicit expressions for $\mathbb{E}\left[\ell_{n,b}\right]$ and $\mathbb{E}\left[\ell_{n,b_1}\ell_{n,b_2}\right]$; for this we define the functions

$$L_1(n,b) = \int_0^1 \frac{\Gamma(b-p)}{\Gamma(b+1)}\frac{\Gamma(n-b+p)}{\Gamma(n-b+1)}\frac{dp}{\Gamma(1-p)\Gamma(1+p)},$$

$$\begin{aligned} L_2(n,b_1,b_2) = \int_0^1\int_0^{p_1} &\frac{\Gamma(b_1-p_1)}{\Gamma(b_1+1)}\frac{\Gamma(b_2-b_1+p_1-p_2)}{\Gamma(b_2-b_1+1)} \\ &\times\frac{\Gamma(n-b_2+p_2)}{\Gamma(n-b_2+1)}\frac{dp_2\, dp_1}{p_1\Gamma(1-p_1)\Gamma(p_1-p_2)\Gamma(p_2+1)} \end{aligned}$$

and

$$\begin{aligned} L_3(n,b_1,b_2) = \int_0^1\int_0^1 &\frac{\Gamma(b_1-p_1)}{\Gamma(b_1+1)}\frac{\Gamma(b_2-p_2)}{\Gamma(b_2+1)} \\ &\times\frac{\Gamma(n-b_1-b_2+p_1+p_2)}{\Gamma(n-b_1-b_2+1)}\frac{dp_2\, dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1\vee p_2)\Gamma(p_1+p_2)}. \end{aligned}$$

Theorem 3.1. For any pair of integers n, b such that $1 \leq b \leq n-1$, we have

$$\mathbb{E}[\ell_{n,b}] = nL_1(n,b) \tag{15}$$

Also, for any triple of integers n, b_1, b_2, with $1 \leq b_1 \leq b_2 \leq n-1$, we have

$$\mathbb{E}\left[\ell_{n,b_1}\ell_{n,b_2}\right] = nL_2(n,b_1,b_2) + nL_3(n,b_1,b_2)\mathbb{1}_{\{b_1+b_2\leq n\}} \tag{16}$$

As before, we may define $SFS_{n,b}$ as the number of mutations shared by b individuals in the n-coalescent. By conditioning on the value of the associated branch lengths we get

Corollary 3.2. For $1 \leq b \leq n-1$,

$$\mathbb{E}[SFS_{n,b}] = \theta n L_1(n,b)$$

and, for $1 \leq b_1 \leq b_2 \leq n-1$, we have,

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(SFS_{n,b_1}, SFS_{n,b_2}\right) =& \theta^2 n L_2(n,b_1,b_2) + \theta^2 n L_3(n,b_1,b_2)\mathbb{1}_{b_1+b_2\leq n} \\ & - \theta^2 n^2 L_1(n,b_1)L_1(n,b_2) + \theta n L_1(n,b)\mathbb{1}_{b_1=b=b_2}.\end{aligned}$$

We also characterize the asymptotic behavior of the functions L_1, L_2 and L_3 as $n \to \infty$, which in turn give asymptotic approximations for the first and second moments of the branch lengths and of SFS. For this we recall the function f_1 defined in (2) and also define for $0 < u_1 < u_2 < 1$,
(17)

$$f_2(u_1,u_2) := \int_0^1 \int_0^{p_1} \frac{u_1^{-p_1-1}(u_2-u_1)^{p_1-p_2-1}(1-u_2)^{p_2-1}}{p_1\Gamma(1-p_1)\Gamma(p_1-p_2)\Gamma(p_2+1)}\, dp_2\, dp_1,$$

and, for $u_1, u_2 > 0, u_1 + u_2 < 1$,
(18)

$$f_3(u_1,u_2) := \int_0^1 \int_0^1 \frac{u_1^{-p_1-1}u_2^{-p_2-1}(1-u_1-u_2)^{p_1+p_2-1}}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1 \vee p_2)\Gamma(p_1+p_2)}\, dp_2\, dp_1\ .$$

Lemma 3.3. We have as $n \to \infty$,

$$\max_{2\leq b\leq n-1}\left|\frac{n^2 L_1(n,b)}{f_1\left(\frac{b-1}{n-1}\right)} - \frac{b-1}{b}\right| \to 0, \tag{19}$$

whereas for $b = 1$,

$$\frac{n^2}{(\log n) f_1\left(\frac{1}{n-1}\right)} L_1(n,1) \to 1. \tag{20}$$

Similarly

$$\max_{2\leq b_1<b_2\leq n-1}\left|\frac{n^3 L_2(n,b_1,b_2)}{f_2\left(\frac{b_1-1}{n-1},\frac{b_2-1}{n-1}\right)} - \frac{b_1-1}{b_1}\right| \to 0, \tag{21}$$

and if also $b_1 \vee (n-b_2) \to \infty$ then

$$\max_{\substack{2\leq b_1\leq b_2\leq n-1\\ b_1+b_2<n}}\left|\frac{n^3 L_3(n,b_1,b_2)}{f_3\left(\frac{b_1-1}{n-2},\frac{b_2-1}{n-2}\right)} - \left(\frac{b_1-1}{b_1}\right)\left(\frac{b_2-1}{b_2}\right)\right| \to 0. \tag{22}$$

Remark. The above lemma does not cover the cases $b_1 = 1$ or $b_1 = b_2$ for L_2, nor the cases $b_1 = 1$, $b_2 = 1$, $n = b_1 + b_2$ or $b_1 \vee (n - b_2) \not\to \infty$ for L_3. However, using the same techniques we also obtain asymptotics in these cases which are used in Theorem 3.5 below.

The proof of the above lemma also gives asymptotic expressions for the functions f_1, f_2 and f_3, leading to straightforward asymptotics for the expectation and covariance of SFS. The complete picture for the first moment is given in the next result.

Theorem 3.4. As n goes to infinity,
(i) The expected number of external mutations ($b = 1$) has the following asymptotics

$$\frac{\log n}{n}\mathbb{E}[SFS_{n,1}] \to \theta.$$

(ii) If $b \geq 2$ and $\frac{b}{n} \to 0$, then

$$\frac{b(b-1)}{n}\log^2\left(\frac{n}{b}\right)\mathbb{E}[SFS_{n,b}] \to \theta.$$

(iii) If $\frac{b}{n} \to u \in (0,1)$, then

$$n\mathbb{E}[SFS_{n,b}] \to \theta f_1(u) = \theta\int_0^1 u^{-1-p}(1-u)^{p-1}\frac{\sin(\pi p)}{\pi p}\,dp.$$

(iv) If $\frac{n-b}{n} \to 0$, then

$$(n-b)\log\left(\frac{n}{n-b}\right)\mathbb{E}[SFS_{n,b}] \to \theta.$$

(v) Let $I = (x, y)$ with $0 < x < y < 1$ and define

$$SFS_{n,I} := \sum_{b=\lceil nx \rceil}^{\lfloor ny \rfloor} SFS_{n,b}.$$

Then

$$\mathbb{E}\left[SFS_{n,I}\right] \to \mathbb{E}[SFS_I]$$

as it is defined in (7).

Case (i) and case (ii) for fixed b also follow from Theorem 4 in [2]. Cases (ii) and (iv) give an update to the approximation of the SFS for small and large families made in [13].

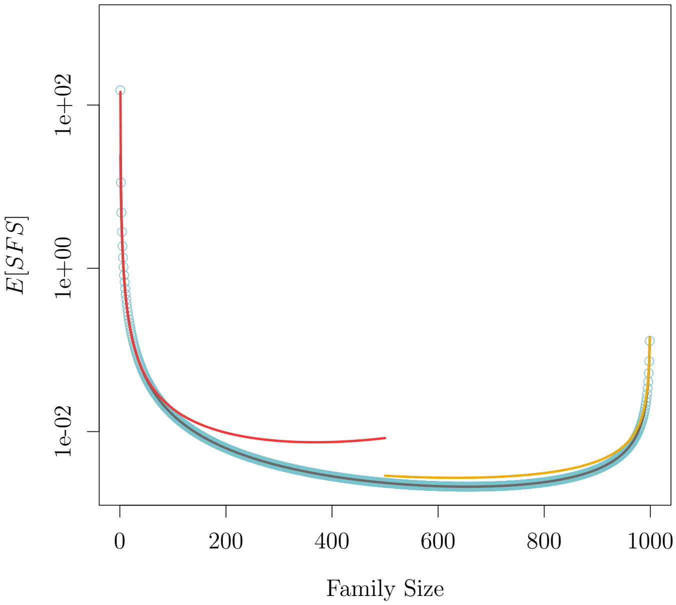

Figure 4: Exact and asymptotic approximations for $\mathbb{E}[SFS]$ in a population of size 1000: The blue circles give the exact value as given in Corollary 3.2. The gray line is the asymptotic approximation as given in Theorem 3.4 (iii). Red (resp. yellow) line is given by Theorem 3.4 (ii) (resp. (iv)).

In the same spirit and using the same techniques we now provide the complete picture for the second moments. In what follows we use the notation $f(n) \sim g(n)$ to denote that

$$\frac{f(n)}{g(n)} \to 1$$

as $n \to \infty$.

Theorem 3.5. The covariance function has the following asymptotics as n goes to infinity, in each of the following cases:

b_1	$b_2 - b_1$	$n - b_2$	$\mathbb{Cov}(SFS_{n,b_1}, SFS_{n,b_2})$
> 1	> 0	$\sim n$	$\frac{\theta^2}{b_1(b_1-1)b_2(b_2-1)} \mathcal{O}\left(\frac{n^2}{\log^5 n}\right)$
$\sim n$	> 0	> 0	$\frac{\theta^2}{(b_2-b_1)(n-b_1)} \frac{1}{\log^2 n}$
$\sim n$	0	> 0	$\frac{\theta^2+\theta}{n-b_2} \frac{1}{\log n}$
> 1	$\sim n$	$= b_1$	$\theta^2 \mathcal{O}\left(\frac{n}{\log^4 n}\right)$
> 1	$\sim n$	$= b_1 + const^+$	$\theta^2 L_1(n - b_2, b_1) \frac{n}{\log n}$
1	0	$\sim n$	$\theta^2 \mathcal{O}\left(\frac{n^2}{\log^3 n}\right)$
1	> 0	$\sim n$	$\theta^2 \mathcal{O}\left(\frac{n^2}{\log^4 n}\right)$
1	$\sim nu$	$\sim n(1-u)$	$\theta^2 \mathcal{O}\left(\frac{1}{\log^2 n}\right)$
1	$\sim n$	> 1	$\theta^2 \mathcal{O}\left(\frac{n}{\log^3 n}\right)$
1	$\sim n$	1	$\theta^2 \mathcal{O}\left(\frac{n}{\log^3 n}\right)$
> 1	0	$\sim n$	$\theta^2 \mathcal{O}\left(\frac{n^2}{\log^5 n}\right)$
$\sim nu$	> 0	$\sim n(1-u)$	$\frac{\theta^2}{(1-u)(b_2-b_1)} \frac{1}{n \log^2 n}$
$\sim nu$	0	$\sim n(1-u)$	$\frac{\theta f_1(u)}{n}$
> 1	$\sim nu$	$\sim n(1-u)$	$\theta^2 \mathcal{O}\left(\frac{1}{\log^3 n}\right)$
$\sim nu$	$\sim n(1-u)$	> 0	$-\frac{\theta^2 f_1(u)}{n-b_2} \frac{1}{\log n}$
$\sim nu_1$	$\sim nu_2$	$\sim n(1-u_1-u_2)$	$\frac{\theta^2\left(f_2(u_1,u_1+u_2)+f_3(u_1,u_1+u_2)\mathbb{1}_{2u_1+u_2\leq 1}-f_1(u_1)f_1(u_1+u_2)\right)}{n^2}$
$\sim nu$	$\sim n(1-2u)$	$= b_1$	$\frac{\theta^2 \int_0^\infty \int_0^\infty \frac{e^{-y_1 e^{-y_2}}}{y_1 \vee y_2}\, dy_1\, dy_2}{u(1-u)} \frac{1}{n \log n}$
$\sim nu$	$\sim n(1-2u)$	$= b_1 + const^+$	$\frac{\theta^2 \int_0^\infty \int_0^\infty \frac{e^{-y_1 e^{-y_2 (y_1+y_2)}}}{y_1 \vee y_2}\, dy_1\, dy_2}{u(1-u)(n-b_2-b_1)} \frac{1}{n \log^2 n}$

Also for $I, \widehat{I} \subset (0,1)$, and $SFS_{n,I}, SFS_{n,\widehat{I}}$ as defined in Theorem 3.4 (V), we have

(23)

$$\mathbb{Cov}\left(SFS_{n,I}, SFS_{n,\widehat{I}}\right) \to$$

$$\theta^2 \int_I \int_{\widehat{I}} f_2(u_1,u_2) + f_3(u_1,u_2)\mathbb{1}_{u_1+u_2<1} - f_1(u_1)f_1(u_2)\, du_2\, du_1 + \theta \int_{I\cap\widehat{I}} f_1(u)\, du.$$

These approximations follow from the asymptotics for L_1, L_2, and L_3 substituted in the covariance formula given in Corollary 3.2. For the sake of simplicity we do not provide the explicit computations. We only treat the case where the expected value $\mathbb{E}[SFS_{n,b}]$ diverges, then an application of Chebyshev's inequality allows us to prove the following weak law of large numbers with L^2-convergence, which generalizes and strengthens results on the Bolthausen-Sznitman coalescent derived in [2].

Corollary 3.6. Suppose that $b/n \to 0$ in such a way that $\mathbb{E}[SFS_{n,b}] \to \infty$, or equivalently that $b = o\left(\sqrt{n}/\log n\right)$. Then we have the following L^2-convergence:

$$\frac{SFS_{n,b}}{\mathbb{E}[SFS_{n,b}]} \to \theta.$$

In view of Theorem 3.4 this means that for $b = 1$

$$\frac{\log n}{n}\, SFS_{n,1} \to \theta,$$

and for $b \geq 2$, $b = o\left(\sqrt{n}/\log n\right)$

$$\frac{b(b-1)\log^2(n/b)}{n}\, SFS_{n,b} \to \theta.$$

4 Distribution of the Family-Sized Branch Lengths

In this section we discuss the particular case of $\ell_{n,b}$ when $b > n/2$. In this case we are able to provide an explicit formula for the distribution function of the length of the coalescent of order b. This leads to convergence in law results, but also to the law of $SFS_{n,b}$. Observe that in this case, for all $t \geq 0$, $C_{n,b}(t) \in \{0,1\}$ and $\ell_{n,b}$ is just the time during which the block of size b survives before coalescing with other blocks (if it ever exists, otherwise obviously $\ell_{n,b} = 0$). We first find an expression for the distribution function of $\ell_{n,b}$.

Theorem 4.1. Suppose that $n/2 < b < n$. For any $s \geq 0$,
(24)

$$\mathbb{P}(\ell_{n,b} > s) = \frac{n}{(n-b)b(b-1)} \int_0^1 \frac{\Psi(b-p) - \Psi(1-p)}{Be(n-b, e^{-s}p)Be(b-1, 1-p)}\, dp.$$

From the derived distribution of $\ell_{n,b}$ in Theorem 4.1 we obtain that, conditioned on $\ell_{n,b} > 0$, the variable $(\log n)\,\ell_{n,b}$ has a limiting distribution.

Corollary 4.2. Suppose that $b/n \to u \in [1/2, 1)$ as $n \to \infty$, then letting $\alpha = \log(1-u) - \log u$, we have

$$\frac{n}{\log n}\mathbb{P}(\ell_{n,b} > 0) \to \frac{G(\alpha)}{u(1-u)}$$

where

$$G(x) = \int_0^1 e^{px} \frac{\sin \pi p}{\pi}\, dp = \frac{1 + e^x}{\pi^2 + x^2}.$$

Furthermore,

$$\mathbb{P}((\log n)\,\ell_{n,b} > s | \ell_{n,b} > 0) \to \frac{G(\alpha - s)}{G(\alpha)}.$$

We now give the joint distribution of the branch lengths for large families, i.e. the joint distribution of the vector $(\ell_{n,b})_{b>n/2}$. For this we introduce the following events: for any collection of integers $\mathbf{b} = (b_1, \cdots, b_m)$ such that $n/2 < b_1 < b_2 < \cdots < b_m < n$, and any collection of nonnegative numbers $\mathbf{s} = (s_1, \cdots, s_m)$, define the event

$$\Lambda_{\mathbf{b},\mathbf{s}} := \left(\bigcap_{i=1}^{m} \{\ell_{b_i} > s_i\} \right) \bigcap \left(\bigcap_{\substack{b > b_1 \\ b \notin \mathbf{b}}} \{\ell_b = 0\} \right),$$

that is, the event that a block of size b_1 exists for a time larger than s_1, that this block then merges with some other blocks of total size exactly $b_2 - b_1$, that this new block exists for a time larger than s_2, and so on, until the last merge of the growing block occurs with the remaining blocks of total size exactly $n - b_m$.

Theorem 4.3. For $\mathbf{b} = (b_1, \cdots, b_m)$ and $\mathbf{s} = (s_1, \cdots, s_m)$ as above, we have

(25)

$$\mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}\right) = \frac{n}{b_1(b_2-b_1)\cdots(n-b_m)} \frac{\exp\{-\langle (m:1), \mathbf{s}\rangle\}}{m!} \int_0^1 p^m \frac{\Psi(b_1-p)-\Psi(1-p)}{Be(b_1-1,1-p)}\, dp$$

and

(26)

$$\mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}, \bigcap_{n/2<b<b_1} \{\ell_{n,b}=0\}\right)$$
$$= \frac{n}{(b_2-b_1)\cdots(n-b_m)} \frac{\exp\{-\langle (m:1), \mathbf{s}\rangle\}}{m!} \times$$
$$\left(\int_0^1 \frac{p^m}{b_1} \frac{\Psi(b_1-p)-\Psi(1-p)}{Be(b_1-1,1-p)} - \frac{p^{m+1}}{m+1} \sum_{n/2<b<b_1} \frac{1}{b(b_1-b)} \frac{\Psi(b-p)-\Psi(1-p)}{Be(b-1,1-p)}\, dp\right),$$

where

$$(m:1) \coloneqq (m, m-1, \ldots, 1).$$

and $\langle \cdot, \cdot \rangle$ is the usual inner product in Euclidean space.

By conditioning on $(\ell_{n,b})_{b>n/2}$ and using equation (26) one can obtain a sampling formula for the vector $(SFS_{n,b})_{b>n/2}$, although the computations are rather convoluted and we do not present them here.

5 The approximations

Here we derive the approximations given above in the Introduction. From Stirling's approximation we have the well-known formula $\Gamma(m+c)/\Gamma(m) \approx m^c$. Its application requires some care, since we shall apply this approximation also for small values of m down to $m = 1$. It is known and easily confirmed by computer that the approximation is particularly accurate within the range $0 \le c \le 1$. Thus we use for $p \in (0,1)$ and $b \ge 2$ the approximations

$$\frac{\Gamma(b-p)}{\Gamma(b+1)} = \frac{1}{b(b-1)} \frac{\Gamma(b-1+(1-p))}{\Gamma(b-1)} \approx \frac{1}{b(b-1)}(b-1)^{1-p} = \frac{(b-1)^{-p}}{b}$$

and

$$\frac{\Gamma(n-b+p)}{\Gamma(n-b+1)} = \frac{1}{n-b} \frac{\Gamma(n-b+p)}{\Gamma(n-b)} \approx (n-b)^{p-1}.$$

Also by Euler's reflection formula $\Gamma(1-p)\Gamma(1+p) = \pi p/\sin(\pi p)$. Inserting these formulas into the expression (15) for the expected SFS we obtain

$$\begin{aligned}\mathbb{E}[SFS_{n,b}] &\approx \theta n \frac{b-1}{b}\int_0^1 (b-1)^{-p-1}(n-b)^{p-1}\frac{\sin(\pi p)}{\pi p}\,dp\\ &= \theta\frac{n}{(n-1)^2}\frac{b-1}{b}f_1\Big(\frac{b-1}{n-1}\Big).\end{aligned}$$

It turns out that this approximation overestimates the expected SFS, which can be somewhat counterbalanced by replacing the scaling factor $n/(n-1)^2$ by $1/(n-1)$. This yields our first approximation (1).

For the second approximation (3) we apply the expansion

$$\frac{\Gamma(m+c)}{\Gamma(m)} = m^c\Big(1-\frac{c(1-c)}{2m}+O(m^{-2})\Big),$$

see [18]. Again this approximation is particularly accurate for $0 \le c \le 1$ leading for $p \in (0,1)$ and $b \ge 2$ to

$$\begin{aligned}\frac{\Gamma(b-p)}{\Gamma(b+1)}\frac{\Gamma(n-b+p)}{\Gamma(n-b+1)} &\approx \frac{(b-1)^{-p}}{b}(n-b)^{p-1}\Big(1-\frac{(1-p)p}{2(b-1)}\Big)\Big(1-\frac{p(1-p)}{2(n-b)}\Big)\\ &\approx \frac{(b-1)^{-p}}{b}(n-b)^{p-1}\Big(1-(n-1)\frac{p(1-p)}{2(b-1)(n-b)}\Big).\end{aligned}$$

Using this approximation in the expression for the expected SFS we get for $b \ge 2$

$$\begin{aligned}\mathbb{E}[SFS_{n,b}] \approx \theta n\frac{b-1}{b}\Bigg(&\frac{1}{(n-1)^2}f_1\Big(\frac{b-1}{n-1}\Big)\\ &-\frac{n-1}{2}\int_0^1(b-1)^{-p-2}(n-b)^{p-2}\frac{\sin(\pi p)}{\pi}(1-p)\,dp\Bigg)\\ = \theta n\frac{b-1}{b}\Big(&\frac{1}{(n-1)^2}f_1\Big(\frac{b-1}{n-1}\Big)-\frac{1}{(n-1)^3}g_1\Big(\frac{b-1}{n-1}\Big)\Big)\end{aligned}$$

with the function g_1 as defined in (4). This integral can be evaluated by elementary means yielding formula (3).

6 Proofs of Section 3

As in the infinite coalescent case, the proof of Theorem 3.1 begins with the definition (14) and by noting that

$$\mathbb{E}\left[\ell_{n,b}\right] = \mathbb{E}\left[\int_0^\infty C_{n.b}(t)\ dt\right] = \int_0^\infty \mathbb{E}\left[C_{n,b}(t)\right]\ dt,$$

and similarly

$$\mathbb{E}\left[\ell_{n,b_1}\ell_{n,b_2}\right] = \int_0^\infty \int_0^\infty \mathbb{E}\left[C_{n,b_1}(t_1)C_{n,b_2}(t_2)\right]\ dt_1\ dt_2,$$

so it only remains to compute $\mathbb{E}\left[C_{n,b}(t)\right]$ and $E\left[C_{n,b_1}(t)C_{n,b_2}(t)\right]$ in each case and simplify the expressions.

Proof of Theorem 3.1 *(first moment).* Let $\mathcal{B}$ be the collection of all possible blocks of size b in a partition of $[n]$. Then

$$\mathbb{E}\left[C_{n,b}(t)\right] = \mathbb{E}\left[\sum_{B\in\mathcal{B}} \mathbb{1}_{B\in\Pi^n(t)}\right] = \sum_{B\in\mathcal{B}} \mathbb{P}\left(B\in\Pi^n(t)\right),$$

and by exchangeability of $\Pi^n(t)$,

$$\mathbb{E}\left[C_{n,b}(t)\right] = \binom{n}{b}\mathbb{P}\left(\{1,\cdots,b\}\in\Pi^n(t)\right).$$

Thus, using (8), the fact that $|\Pi^\infty(t)|^\downarrow =: (A_1, A_2, \dots) \overset{d}{=} PD(e^{-t}, 0)$, and writing Π^n as $\Pi^\infty_{|n}$, we obtain

$$\begin{aligned}\mathbb{E}[C_{n,b}(t)] &= \binom{n}{b}\mathbb{E}\left[\sum_{i=1}^\infty A_i^b(1-A_i)^{n-b}\right]\\ &= \binom{n}{b}\int_0^1 u^{b-1}(1-u)^{n-b}\frac{u^{-e^{-t}}(1-u)^{e^{-t}-1}}{\Gamma(1-e^{-t})\Gamma(e^{-t})}\ du\\ &= \frac{n\Gamma(n)}{\Gamma(n-b+1)\Gamma(b+1)}\frac{Be(b-e^{-t}, n-b+e^{-t})}{\Gamma(1-e^{-t})\Gamma(1+e^{-t})}.\end{aligned}$$

Finally, by changing the variable $p = e^{-t}$, we obtain (15). □

Now we use the random tree construction of the n-Bolthausen-Sznitman coalescent in order to compute the second moments of $\ell_{n,b}$.

Proof of Theorem 3.1 *(second moments).* Let $1 \leq b_1 \leq b_2 \leq n-1$, and $\mathcal{B}_1, \mathcal{B}_2$ be the collection of all possible blocks of sizes b_1 and b_2 respectively in a partition of $[n]$. Then

$$\begin{aligned}\mathbb{E}\left[\ell_{n,b_1}\ell_{n,b_2}\right] &= \int_0^\infty \int_0^\infty \mathbb{E}\left[C_{n,b_1}(t_1)C_{n,b_2}(t_2)\right]\, dt_2\, dt_1 \\ &= \int_0^\infty \int_0^\infty \sum_{B_1\in\mathcal{B}_1}\sum_{B_2\in\mathcal{B}_2} \mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right)\, dt_2\, dt_1. \qquad (27)\end{aligned}$$

We now compute $\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right)$ by cases.

i) Suppose that $B_1 \cap B_2 = \emptyset$. By exchangeability we have

$$\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right) = \mathbb{P}(\{1,\cdots,b_1\} \in \Pi^n(t_1), \{b_1+1,\cdots,b_1+b_2\} \in \Pi^n(t_2))$$

where this probability is of course 0 if $b_1+b_2 > n$. Now suppose that $t_1 \leq t_2$. In terms of the RRT construction of the Bolthausen-Sznitman coalescent, the event

$$\{\{1,\cdots,b_1\} \in \Pi^n(t_1), \{b_1+1,\cdots,b_1+b_2\} \in \Pi^n(t_2)\}$$

is characterized by a RRT with exponential edges, say $E_2,\cdots,E_n$, constructed as follows: for $i \in \{1,\cdots,b_1-1\}$ the node $\{i+1\}$ along with E_{i+1} arrive to the tree but with the imposed restriction that it may not attach to $\{1\}$ and have $E_{i+1} > t_1$ at the same time, which occurs with probability e^{-t_1}/i; this ensures that $\{i+1\}$ coalesces with $\{1\}$ before time t_1 for all $i < b_1$, thus creating the block $\{1,\cdots,b_1\}$ up to time t_1. After $\{1\},\cdots,\{b_1\}$ have arrived, the node $\{b_1+1\}$ must attach to $\{1\}$ and E_{b_1+1} must be greater than t_2, which occurs with probability e^{-t_2}/b_1; the node $\{b_1+1\}$ will be the root of a sub-tree formed with the nodes $\{b_1+2\},\cdots,\{b_1+b_2\}$ which will build the block $\{b_1+1,\cdots,b_1+b_2\}$ at time t_2. Thus, for each $i \in \{1,\cdots,b_2-1\}$ the node $\{b_1+i+1\}$ must arrive and attach to any of $\{b_1+1\},\cdots,\{b_1+i\}$, which occurs with probability $\frac{i}{b_1+i}$, and, furthermore, conditional on this event, it may not attach to $\{b_1+1\}$ and have $E_{b_1+i+1} > t_2$ at the same time, which occurs with probability $\frac{e^{-t_2}}{i}$. Finally, if $n-b_1-b_2 > 0$, for $i \in \{0,\cdots,n-b_1-b_2-1\}$ the node $\{b_1+b_2+i+1\}$ must either attach to any of $\{b_1+b_2+j\}$, $1 \leq j \leq i$, or attach to $\{1\}$ or $\{b_1+1\}$ and have $E_{b_1+b_2+i+1} > t_1$ or $E_{b_1+b_2+i+1} > t_2$ respectively; this occurs with probability $\frac{e^{-t_1}+e^{-t_2}+i}{b_1+b_2+i}$. Putting all together we obtain

$$
\begin{aligned}
&\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right)\\
=&\left[\prod_{i=1}^{b_1-1}\left(1-\frac{e^{-t_1}}{i}\right)\right]\left[\frac{e^{-t_2}}{b_1}\prod_{i=1}^{b_2-1}\left(1-\frac{e^{-t_2}}{i}\right)\frac{i}{b_1+i}\right]\left[\prod_{i=0}^{n-b_1-b_2-1}\frac{e^{-t_1}+e^{-t_2}+i}{b_1+b_2+i}\right]\\
=&\frac{1}{(n-1)!}\frac{\Gamma(b_1-e^{-t_1})}{\Gamma(1-e^{-t_1})}e^{-t_2}\frac{\Gamma(b_2-e^{-t_2})}{\Gamma(1-e^{-t_2})}\frac{\Gamma(n-b_1-b_2+e^{-t_1}+e^{-t_2})}{\Gamma(e^{-t_1}+e^{-t_2})},
\end{aligned}
$$

where the last product is set to 1 if $n-b_2-b_1=0$. On the other hand, if $t_2 < t_1$, by exchangeability we may instead compute

$$\mathbb{P}(\{1,\cdots,b_2\} \in \Pi^n(t_2), \{b_2+1,\cdots,b_2+b_1\} \in \Pi^n(t_1))$$

obtaining

$$
\begin{aligned}
&\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right)\\
=&\frac{1}{(n-1)!}\frac{\Gamma(b_2-e^{-t_2})}{\Gamma(1-e^{-t_2})}e^{-t_1}\frac{\Gamma(b_1-e^{-t_1})}{\Gamma(1-e^{-t_1})}\frac{\Gamma(n-b_2-b_1+e^{-t_2}+e^{-t_1})}{\Gamma(e^{-t_2}+e^{-t_1})}.
\end{aligned}
$$

ii) Suppose that $B_1 \subset B_2$. Of course if $t_1 > t_2$ we have $\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right) = 0$ whenever B_1 is strictly contained in B_2. Assuming that $t_1 \leq t_2$ and using the same rationale as before we obtain

$$
\begin{aligned}
&\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right)\\
=&\left[\prod_{i=1}^{b_1-1}\frac{i-e^{-t_1}}{i}\right]\left[\prod_{i=0}^{b_2-b_1-1}\frac{i+e^{-t_1}-e^{-t_2}}{b_1+i}\right]\left[\prod_{i=0}^{n-b_2-1}\frac{e^{-t_2}+i}{b_2+i}\right]\\
=&\frac{1}{(n-1)!}\frac{\Gamma(b_1-e^{-t_1})}{\Gamma(1-e^{-t_1})}\frac{\Gamma(b_2-b_1+e^{-t_1}-e^{-t_2})}{\Gamma(e^{-t_1}-e^{-t_2})}\frac{\Gamma(n-b_2+e^{-t_2})}{\Gamma(e^{-t_2})},
\end{aligned}
$$

where the product in the middle is set to 1 if $B_1 = B_2$.

iii) If $B_1 \cap B_2 \neq \emptyset$ and $B_1 \not\subset B_2$, we clearly have $\mathbb{P}\left(B_1 \in \Pi^n(t_1), B_2 \in \Pi^n(t_2)\right) = 0$.

From the previous computations, and summing over the corresponding cases, we see that if $b_1 + b_2 \leq n$ then, changing the variable

$p = e^{-t}$, the integral in (27) is given by

$$\begin{aligned}\mathbb{E}\left[\ell_{n,b_1}\ell_{n,b_2}\right] =& \frac{n}{b_1!b_2!(n-b_1-b_2)!}\\ & \int_0^1\int_0^1 \frac{\Gamma(b_1-p_1)}{\Gamma(1-p_1)}\frac{\Gamma(b_2-p_2)}{\Gamma(1-p_2)}\frac{\Gamma(n-b_1-b_2+p_1+p_2)}{\Gamma(p_1+p_2)}\frac{dp_1\,dp_2}{p_1\vee p_2}\\ +& \frac{n}{b_1!(b_2-b_1)!(n-b_2)!}\\ & \int_0^1\int_0^{p_1} \frac{\Gamma(b_1-p_1)}{\Gamma(1-p_1)}\frac{\Gamma(b_2-b_1+p_1-p_2)}{\Gamma(p_1-p_2)}\frac{\Gamma(n-b_2+p_2)}{\Gamma(p_2+1)}\frac{dp_2\,dp_1}{p_1}\end{aligned}$$

whereas if $b_1 + b_2 > n$ the first summand in the above expression is set to zero. Rearranging terms we obtain (16). □

Proof of Lemma 3.3 *(asymptotics for L_1).* Again, we have from Stirling's formula that $\Gamma(m+c)/\Gamma(m+d) = m^{c-d}(1+\mathcal{O}(1/m))$ for any real numbers c and d, where the $\mathcal{O}(1/m)$ term holds uniformly for $0 \le c, d \le 1$. Letting $m = b-1$ and $n-b$ leads to the following equality:

$$\begin{aligned}&\frac{n}{b(n-b)}\frac{\Gamma(n-b+p)}{\Gamma(n-b)}\frac{\Gamma(b-p)}{\Gamma(b)}\\ &\qquad = \frac{n}{b(n-b)}(n-b)^p(b-1)^{-p}\left(1+\mathcal{O}\left(\frac{1}{b}\right)+\mathcal{O}\left(\frac{1}{n-b}\right)\right).\end{aligned}$$

Thus, using Euler's reflection formula to write $\Gamma(1-p)\Gamma(1+p)$ as $\pi p/\sin(\pi p)$ in the definition of L_1, we get

$$\begin{aligned}L_1(n,b) &= \left(1+\mathcal{O}\left(\frac{1}{b}\right)+\mathcal{O}\left(\frac{1}{n-b}\right)\right)\frac{1}{b(n-b)}\int_0^1 \frac{\sin(\pi p)}{\pi p}\left(\frac{n-b}{b-1}\right)^p dp\\ &= \left(1+\mathcal{O}\left(\frac{1}{b}\right)+\mathcal{O}\left(\frac{1}{n-b}\right)\right)\frac{b-1}{b(n-1)^2}f_1\left(\frac{b-1}{n-1}\right)\end{aligned}$$

Thus, for every $\epsilon > 0$ there is a $b_0 \in \mathbb{N}$ such that for large enough $n \in \mathbb{N}$ we have

$$\max_{b_0\le b\le n-b_0}\left|\frac{n^2L_1(n,b)}{f_1\left(\frac{b-1}{n-1}\right)} - \frac{b-1}{b}\right| < \epsilon. \tag{28}$$

It remains to study the approximation as $n \to \infty$ in the cases where $n-b$ or b remain constant. In the first case, when $n-b = c$, we have

$b \to \infty$ as $n \to \infty$ and, by Stirling's approximation and dominated convergence and substituting $p = y/\log b$ on the one hand

$$\begin{aligned}
L_1(n,b) \sim & \int_0^1 \frac{\sin(\pi p)}{\pi p} b^{-p-1} \frac{\Gamma(c+p)}{\Gamma(c+1)}\, dp \\
=& \frac{1}{bc} \int_0^{\log b} \frac{\sin(\pi y/\log b)}{\pi y/\log b} e^{-y} \frac{\Gamma(c+y/\log b)}{\Gamma(c)} \frac{dy}{\log b} \\
\sim & \frac{1}{bc\log b} \int_0^\infty e^{-y}\, dy.
\end{aligned}$$

and on the other hand because of $b \to \infty$

$$\begin{aligned}
\frac{1}{n^2} f_1\left(\frac{b-1}{n-1}\right) \sim & \frac{1}{bc} \int_0^1 \frac{\sin(\pi p)}{\pi p} b^{-p} c^p\, dp \\
=& \frac{1}{bc} \int_0^{\log b} \frac{\sin(\pi y/\log b)}{\pi y/\log b} e^{-y} c^{y/\log b} \frac{dy}{\log b} \\
\sim & \frac{1}{bc\log b} \int_0^\infty e^{-y}\, dy.
\end{aligned}$$

Thus $L_1(n,b) \sim n^{-2} f_1((b-1)/(n-1))$ which extends (28) for $b > n - b_0$.

Similarly for the second case, if $b \geq 2$ is fixed, we have $n - b \to \infty$ as $n \to \infty$. Thus, with $1 - p = y/\log n$

$$\begin{aligned}
L_1(n,b) \sim & \int_0^1 \frac{\sin(\pi p)}{\pi p} \frac{\Gamma(b-p)}{\Gamma(b+1)} n^{p-1}\, dp \\
=& \frac{1}{\log^2(n)} \int_0^{\log n} \frac{\sin(\pi - \pi y/\log n)}{(1-y/\log n)\pi y/\log n} \frac{\Gamma\left(b-1+\frac{y}{\log n}\right)}{\Gamma(b+1)} y e^{-y}\, dy \\
\sim & \frac{1}{b(b-1)\log^2 n} \int_0^\infty y e^{-y}\, dy
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n^2} f_1\left(\frac{b-1}{n-1}\right) \sim & \frac{1}{(b-1)^2} \int_0^1 \frac{\sin \pi p}{\pi p} (b-1)^{1-p} n^{p-1}\, dp \\
=& \frac{1}{(b-1)^2 \log^2 n} \int_0^{\log n} \frac{\sin(\pi - \pi y/\log n)}{(1-y/\log n)\,\pi y/\log n} (b-1)^{y/\log n} y e^{-y}\, dy \\
\text{(29)} & \\
\sim & \frac{1}{(b-1)^2 \log^2 n} \int_0^\infty y e^{-y}\, dy.
\end{aligned}$$

Thus $L_1(n,b) \sim (b-1)n^{-2}f_1((b-1)/(n-1))/b$, which extends (28) for $b < b_0$. This extends (28) for $b < b_0$. Thus we proved (19).

For the proof of (20), we substitute b by 1 and perform similar computations:

$$\begin{aligned} L_1(n,1) &= \int_0^1 \frac{\Gamma(1-p)}{\Gamma(2)} \frac{\Gamma(n-1+p)}{\Gamma(n)} \frac{dp}{\Gamma(1-p)\Gamma(1+p)} \\ &\sim \int_0^1 n^{p-1} \frac{dp}{\Gamma(1+p)} \\ &= \int_0^{\log n} e^{-y} \frac{dy}{(\log n)\Gamma(2-y/\log n)} \\ &\sim \frac{1}{\log n} \int_0^\infty e^{-y}\, dy, \end{aligned}$$

and from (29) with choosing $b = 2$

$$\frac{1}{n^2} f_1\left(\frac{1}{n-1}\right) \sim \frac{1}{\log^2 n} \int_0^\infty y e^{-y}\, dy.$$

This proves (20). □

Proof of Lemma 3.3 *(asymptotics for L_2 and L_3)* . The arguments here are similar to the arguments in the proof of the asymptotics for L_1, but we avoid repeating similar and tedious computations. We only layout the first steps of the proof. By Stirling's approximation applied to the integrands appearing in L_2 and L_3, we obtain, for $b_2 - b_1 > 0$,

$$\begin{aligned} &\frac{\Gamma(b_1-p_1)}{\Gamma(b_1+1)} \frac{\Gamma(b_2-b_1+p_1-p_2)}{\Gamma(b_2-b_1+1)} \frac{\Gamma(n-b_2+p_2)}{\Gamma(n-b_2+1)} = \\ &\frac{1}{(n-1)^3} \left(\frac{b_1-1}{n-1}\right)^{-p_1-1} \left(\frac{b_2-b_1}{n-1}\right)^{p_1-p_2-1} \left(\frac{n-b_2}{n-1}\right)^{p_2-1} \times \\ &\left(1 + \mathcal{O}\left(\frac{1}{b_1}\right) + \mathcal{O}\left(\frac{1}{b_2-b_1}\right) + \mathcal{O}\left(\frac{1}{n-b_2}\right)\right), \end{aligned}$$

and, for $n - b_2 - b_1 > 0$,

$$\begin{aligned} &\frac{\Gamma(b_1-p_1)}{\Gamma(b_1+1)} \frac{\Gamma(b_2-p_2)}{\Gamma(b_2+1)} \frac{\Gamma(n-b_1-b_2+p_1+p_2)}{\Gamma(n-b_1-b_2+1)} = \\ &\frac{1}{(n-2)^3} \left(\frac{b_1-1}{n-2}\right)^{-p_1-1} \left(\frac{b_2-1}{n-2}\right)^{-p_2-1} \left(1 - \frac{b_1+b_2}{n-2}\right)^{p_1+p_2-1} \times \\ &\left(1 + \mathcal{O}\left(\frac{1}{b_1}\right) + \mathcal{O}\left(\frac{1}{b_2}\right) + \mathcal{O}\left(\frac{1}{n-b_1-b_2}\right)\right); \end{aligned}$$

thus

$$L_2(n,b_1,b_2) = \frac{1}{(n-1)^3} f_2\left(\frac{b_1-1}{n-1}, \frac{b_2-1}{n-1}\right)\left(1 + \mathcal{O}\left(\frac{1}{b_1}\right) + \mathcal{O}\left(\frac{1}{b_2-b_1}\right) + \mathcal{O}\left(\frac{1}{n-b_2}\right)\right),$$

and

$$L_3(n,b_1,b_2) = \frac{1}{(n-2)^3} f_3\left(\frac{b_1-1}{n-2}, \frac{b_2-1}{n-2}\right)\left(1 + \mathcal{O}\left(\frac{1}{b_1}\right) + \mathcal{O}\left(\frac{1}{b_2}\right) + \mathcal{O}\left(\frac{1}{n-b_1-b_2}\right)\right).$$

Similar to the analysis in the proof of (19), to obtain (21) it remains to study the cases where at least one of $b_1, b_2 - b_1$, or $n - b_2$ remains constant, whereas for (22) the cases of interest are where one of b_1, b_2, or $n - b_2 - b_1$ remain constant. □

Proof of Theorem 3.4. We first derive the asymptotic behavior of the function f_1. We have

$$f_1(u) \sim \frac{1}{u^2 \log^2 u} \qquad \text{as } u \downarrow 0. \tag{30}$$

For the proof note that for $u < 1/2$ we have $(1-u)^{p-1} \le 2$. Therefore dominated convergence implies for $u \downarrow 0$

$$\begin{aligned} f_1(u) &= \frac{1}{u^2}\int_0^1 u^{1-p}(1-u)^{p-1}\frac{dp}{\Gamma(1-p)\Gamma(1+p)} \\ &= \frac{1}{u^2}\int_0^1 e^{-(p-1)\log u}(1-u)^{p-1}(1-p)\frac{dp}{\Gamma(2-p)\Gamma(1+p)} \\ &= \frac{1}{u^2}\int_0^{-\log u} e^{-y}(1-u)^{y/\log\frac{1}{u}}\frac{y}{\log\frac{1}{u}}\cdot\frac{dy}{\log\frac{1}{u}\Gamma(1-\frac{y}{\log u})\Gamma(2+\frac{y}{\log u})} \\ &\sim \frac{1}{u^2\log^2 u}\int_0^\infty ye^{-y}\,dy \end{aligned}$$

implying (30). Also

$$f_1(u) \sim -\frac{1}{(1-u)\log(1-u)} \qquad \text{as } u \uparrow 1, \tag{31}$$

which we obtain again by means of dominated convergence in the limit $u \uparrow 1$ as follows:

$$\begin{aligned} f_1(u) &= \frac{1}{u(1-u)}\int_0^1 e^{p\log(1-u)}u^{-p}\frac{dp}{\Gamma(1-p)\Gamma(1+p)} \\ &= \frac{1}{u(1-u)}\int_0^{-\log(1-u)}\frac{e^{-y}u^{y/\log(1-u)}\,dy}{(-\log(1-u))\Gamma(1+\frac{y}{\log(1-u)})\Gamma(1-\frac{y}{\log(1-u)})} \\ &\sim -\frac{1}{(1-u)\log(1-u)}\int_0^\infty e^{-y}\,dy. \end{aligned}$$

These asymptotics together with Lemma 3.3 imply our claims. Without loss of generality let $\theta = 1$. From (20) we obtain

$$\mathbb{E}\left[SFS_{n,1}\right] = nL_1(n,1) \sim \frac{1}{n} f_1\left(\frac{1}{n-1}\right) \sim \frac{\log n}{n} \frac{(n-1)^2}{\log^2(n-1)}$$

which yields claim (i).

Similary from (19) we get for $b \geq 2$ and $b/n \to 0$

$$\mathbb{E}\left[SFS_{n,b}\right] = nL_1(n,b) \sim \frac{b-1}{nb} f_1\left(\frac{b-1}{n-1}\right) \sim \frac{b-1}{nb} \frac{(n-1)^2}{(b-1)^2 \log^2 \frac{b-1}{n-1}}$$

which in view of $b/n \to 0$ yields assertion (ii).

Claim (iii) is an immediate consequence of formula (19), since here we have $(b-1)/b \to 1$.

Next under the condition $(n-b)/n \to 0$ we get from (19) and (31)

$$\mathbb{E}\left[SFS_{n,b}\right] \sim \frac{b-1}{nb} f_1\left(\frac{b-1}{n-1}\right) \sim -\frac{b-1}{nb} \frac{n-1}{(n-b)\log \frac{n-b}{n-1}} \sim \frac{1}{(n-b)\log \frac{n}{n-b}}$$

which confirms assertion (iv).

Finally, we have from (19)

$$\mathbb{E}\left[SFS_{n,I}\right] \sim \frac{1}{n} \sum_{\frac{b}{n} \in I} f_1\left(\frac{b}{n}\right) \sim \int_I f_1(u)\, du,$$

which is claim (v). This finishes the proof. □

Proof of Theorem 3.5. The approximations follow from the asymptotics for L_1, L_2, and L_3 substituted in the covariance formula given in Corollary 3.2. □

Proof of Corollary 3.6. We have to prove that

$$\mathbb{V}ar(SFS_{n,b}) = o\big(\mathbb{E}[SFS_{n,b}]^2\big).$$

From the monotonicity properties of the gamma function we have for

$1 \le b \le n-1$

$$
\begin{aligned}
L_2(n,b,b) &= \int_0^1 \int_0^{p_1} \frac{\Gamma(b-p_1)}{\Gamma(b+1)} \frac{\Gamma(n-b+p_2)}{\Gamma(n-b+1)} \frac{dp_2\, dp_1}{p_1 \Gamma(1-p_1)\Gamma(p_2+1)} \\
&\le \int_0^1 \frac{\Gamma(b-p_1)}{\Gamma(b+1)} \frac{\Gamma(n-b+p_1)}{\Gamma(n-b+1)} \frac{1}{\Gamma(1-p_1)p_1} \int_0^{p_1} \frac{\Gamma(1+p_1)}{\Gamma(1+p_1)\Gamma(1+p_2)}\, dp_2\, dp_1 \\
&\le \sup_{1 \le x \le y \le 2} \frac{\Gamma(y)}{\Gamma(x)} \int_0^1 \frac{\Gamma(b-p_1)}{\Gamma(b+1)} \frac{\Gamma(n-b+p_1)}{\Gamma(n-b+1)} \frac{dp_1}{\Gamma(1-p_1)\Gamma(p_1+1)} \\
&= \sup_{1 \le x \le y \le 2} \frac{\Gamma(y)}{\Gamma(x)} L_1(n,b). \tag{32}
\end{aligned}
$$

Concerning $L_3(n,b,b)$ we have for $b = o(n)$ by Stirling's approximation uniformly in $0 \le p_1, p_2 \le 1$

$$
\frac{\Gamma(n-2b+p_1+p_2)}{\Gamma(n-2b+1)} \sim n \frac{\Gamma(n-b+p_1)}{\Gamma(n-b+1)} \frac{\Gamma(n-b+p_2)}{\Gamma(n-b+1)},
$$

hence, with $1 < \eta < 2$

$$
\begin{aligned}
&\iint\limits_{\substack{0 \le p_1, p_2 \le 1 \\ \eta < p_1 + p_2 \le 2}} \frac{\Gamma(b-p_1)}{\Gamma(b+1)} \frac{\Gamma(b-p_2)}{\Gamma(b+1)} \frac{\Gamma(n-2b+p_1+p_2)}{\Gamma(n-2b+1)} \\
&\qquad \times \frac{dp_2\, dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1 \vee p_2)\Gamma(p_1+p_2)} \\
&\sim n \iint\limits_{\substack{0 \le p_1, p_2 \le 1 \\ \eta < p_1 + p_2 \le 2}} \frac{\Gamma(b-p_1)}{\Gamma(b+1)} \frac{\Gamma(b-p_2)}{\Gamma(b+1)} \frac{\Gamma(n-b+p_1)}{\Gamma(n-b+1)} \frac{\Gamma(n-b+p_2)}{\Gamma(n-b+1)} \\
&\qquad \times \frac{dp_2\, dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1 \vee p_2)\Gamma(p_1+p_2)} \\
&\le \frac{n}{\eta - 1} \sup_{\eta \le x \le 2} \frac{1}{\Gamma(x)} \int_0^1 \int_0^1 \frac{\Gamma(b-p_1)}{\Gamma(b+1)} \frac{\Gamma(b-p_2)}{\Gamma(b+1)} \frac{\Gamma(n-b+p_1)}{\Gamma(n-b+1)} \\
&\qquad \times \frac{\Gamma(n-b+p_2)}{\Gamma(n-b+1)} \frac{dp_2\, dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)\Gamma(1+p_1)\Gamma(1+p_2)} \\
&= \frac{n}{\eta - 1} \sup_{\eta \le x \le 2} \frac{1}{\Gamma(x)} L_1(n,b)^2. \tag{33}
\end{aligned}
$$

Also, by another application of Stirling's approximation and for $b =$

$$o(n)$$

$$\iint\limits_{\substack{0\le p_1,p_2\le 1\\ 0<p_1+p_2\le\eta}} \frac{\Gamma(b-p_1)}{\Gamma(b+1)}\frac{\Gamma(b-p_2)}{\Gamma(b+1)}\frac{\Gamma(n-2b+p_1+p_2)}{\Gamma(n-2b+1)}$$

$$\times\frac{dp_2\,dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1\vee p_2)\Gamma(p_1+p_2)}$$

$$=O\Big(\iint\limits_{\substack{0\le p_1,p_2\le 1\\ 0<p_1+p_2\le\eta}} b^{-p_1-p_2-2}(n-2b)^{p_1+p_2-1}$$

$$\times\frac{dp_2\,dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1\vee p_2)\Gamma(p_1+p_2)}\Big)$$

$$=O\Big(b^{-\eta-2}(n-2b)^{\eta-1}\iint\limits_{0\le p_1 p_2\le 1}\frac{dp_2\,dp_1}{\Gamma(1-p_1)\Gamma(1-p_2)(p_1\vee p_2)\Gamma(p_1+p_2)}\Big)$$

$$=o\Big(\frac{n}{b^4\log^4 n}\Big) \tag{34}$$

Combining (33) and (34) with Theorem 3.4 (i) and (ii) and letting $\eta\to 2$ we obtain

$$L_3(n,b,b)=nL_1(n,b)^2(1+o(1))+o(n^{-1}\mathbb{E}[SFS_{n,b}]^2).$$

Using this estimate together with (32) and with Theorem 3.1, Corollary 3.2 yields

$$\mathbb{V}ar(SFS_{n,b})=O(\mathbb{E}[SFS_{n,b}])+o(\mathbb{E}[SFS_{n,b}]^2)$$

Because of our assumption $\mathbb{E}[SFS_{n,b}]\to\infty$ our claim is proved. □

7 Proofs of Section 4

Proof of Theorem 4.1. Note that since $b>n/2$, and by the exchangeability of Π^n, we have:

$$\mathbb{P}(\ell_{n,b}>s)=\binom{n}{b}\mathbb{P}\left(\mathcal{L}(\{t:\{1,\cdots,b\}\in\Pi^n(t)\})>s\right),$$

where $\mathcal{L}$ is the Lebesgue measure, and $\mathcal{L}(\{t:\{1,\cdots,b\}\in\Pi^n(t)\})$ gives the time that the block $\{1,\cdots,b\}$ exists in the Bolthausen-Sznitman coalescent starting with n individuals.

We now describe the event $\{\mathcal{L}(\{t : \{1, \cdots, b\}\} \in \Pi^n(t)) > s\}$ in terms of the RRT construction of the Bolthausen-Sznitman coalescent. Let $\mathcal{G}$ be the event that the nodes $\{1\}, \{2\}, \cdots, \{b\}$ and $\{1\}, \{b+1\}, \cdots, \{n\}$ form two sub-trees, say T_1 and T_2 rooted at $\{1\}$; i.e.

$$\mathcal{G} := \{T \colon \{j\} \text{ does not attach to } \{i\}, \text{ for all } 2 \leq i \leq b \text{ and } b < j \leq n\}.$$

Then

$$\mathcal{L}(\{t : \{1, \cdots, b\}\} \in \Pi^n(t)) = \begin{cases} 0 & \text{if } T \notin \mathcal{G} \\ (m(T_2) - M(T_1)) \vee 0 & \text{if } T \in \mathcal{G}. \end{cases}$$

Indeed, observe that by the cutting-merge procedure $T \notin \mathcal{G}$ if and only if any block of Π^n that contains all of $\{1, \cdots, b\}$ also contains some $j \in \{b+1, \cdots, n\}$. On the other hand, on the event $\{T \in \mathcal{G}\}$, the random variable $M(T_1)$ is just the time at which the block $\{1, \cdots, b\}$ appears in Π^n, while $m(T_2)$ is the time at which it coalesces with some other block in T_2. Furthermore, observe that conditioned on $\{T \in \mathcal{G}\}$, T_1 and T_2 are two independent RRTs of sizes b and $n-b+1$ respectively. Thus, by Lemma 2.3 we have

$$\begin{aligned} &\mathbb{P}(\ell_{n,b} > s) \\ &= \binom{n}{b} \mathbb{P}(T \in \mathcal{G}) \mathbb{P}(m(T_2) - M(T_1) > s) \\ &= \binom{n}{b} \prod_{i=0}^{n-b-1} \left(\frac{1+i}{b+i}\right) \frac{1}{(b-1)(n-b)} \int_0^1 \frac{\Psi(b-p) - \Psi(1-p)}{Be(n-b, e^{-s}p) Be(b-1, 1-p)} \, dp \\ &= \frac{n}{(n-b)b(b-1)} \int_0^1 \frac{\Psi(b-p) - \Psi(1-p)}{Be(n-b, e^{-s}p) Be(b-1, 1-p)} \, dp. \end{aligned}$$

□

Proof of Corollary 4.2. Observe that, uniformly for $p \in (0,1)$, we have

$$\Psi(b-p) - \Psi(1-p) = \sum_{k=1}^{b-1} \frac{1}{k-p} = \frac{1}{1-p} + \log b + \mathcal{O}(1),$$

thus, substituting in (24) and also using Stirling's approximation and

Euler's reflection formula, we obtain

$$\begin{aligned}\mathbb{P}(\ell_{n,b} > 0) \sim & \frac{1}{u(1-u)n} \int_0^1 \left(\frac{1-u}{u}\right)^p \frac{\sin \pi p}{\pi} \left(\frac{1}{1-p} + \log n + \mathcal{O}(1)\right) dp \\ \sim & \frac{\log n}{u(1-u)n} \int_0^1 e^{p\alpha} \frac{\sin \pi p}{\pi}\, dp \\ = & \frac{\log n}{u(1-u)n} G(\alpha).\end{aligned}$$

On the other hand, for any $s > 0$ we have

$$\begin{aligned}\mathbb{P}\left((\log n)\, \ell_{n,b} > s\right) \sim & \frac{1}{u(1-u)n} \int_0^1 \frac{b^{-p}\,(n-b)^{pe^{-s/\log n}}}{\Gamma(1-p)\Gamma(pe^{-s/\log n})} \left(\frac{1}{1-p} + \log b + \mathcal{O}(1)\right) dp \\ \sim & \frac{\log n}{u(1-u)n} \int_0^1 e^{p\alpha} (n-b)^{p(e^{-s/\log n}-1)} \frac{1}{\Gamma(1-p)\Gamma(p)}\, dp \\ \sim & \frac{\log n}{u(1-u)n} \int_0^1 e^{p\alpha} (n-b)^{-ps/\log n} \frac{1}{\Gamma(1-p)\Gamma(p)}\, dp \\ \sim & \frac{\log n}{u(1-u)n} \int_0^1 e^{p(\alpha-s)} \frac{\sin \pi p}{\pi}\, dp \\ = & \frac{\log n}{u(1-u)n} G(\alpha - s).\end{aligned}$$

□

Proof of Theorem 4.3. Letting $\ell_\pi := \mathcal{L}\left(t : \pi \in \Pi^n(t)\right)$ for any subset $\pi \subset [n]$, by exchangeability of $\Pi^n(t)$ we have

$$\mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}\right) = \frac{n!}{b_1!(b_2-b_1)!\cdots(n-b_m)!} \mathbb{P}\left(\bigcap_{1\le i\le m} A_{b_i,s_i}, \bigcap_{\substack{b>b_1 \\ b\notin \mathbf{b}}} \bar{A}_{b,0}\right)$$

where

$$A_{b,s} = \{\ell_{\{1,\ldots,b\}} > s\}$$

and

$$\bar{A}_{b,0} = \{\ell_{\{1,\ldots,b\}} = 0\}.$$

Recall that $M\left(T|_{b_1}\right)$ is defined as the maximum of the exponential edges associated to the root of $T|_{b_1}$. Letting $b_{m+1} := n$, and also letting E_b, $1 \leq b \leq n$, be the exponential variable associated to b, we have

$$\mathbb{P}\left(\bigcap_{1\leq i\leq m} A_{b_i,s_i}, \bigcap_{\substack{b>b_1 \\ b\notin \mathbf{b}}} \bar{A}_{b,0} \right)$$
$$= \left(\prod_{i=1}^{m+1} \frac{1\cdot 2 \cdots (b_{i+1}-b_i)}{b_i(b_i+1)\cdots(b_{i+1}-1)} \right) \mathbb{P}\left(E_{b_1+1} - M\left(T\big|_{b_1}\right) > s_1, \bigcap_{i=2}^{m} E_{b_i+1} - E_{b_{i-1}+1} > s_i \right),$$

where the product above is the probability that T is structured in such a way that $\{b_1+1\}$ attaches to $\{1\}$ and is the root of a subtree formed with $\{b_1+1,\ldots,b_2\}$, that $\{b_2+1\}$ attaches to $\{1\}$ and is the root of a subtree formed with $\{b_2+1,\ldots,b_3\}$, and so forth. Using the independence of the exponential variables we obtain

$$\mathbb{P}\left(E_{b_1+1} - M\left(T\big|_{b_1}\right) > s_1, \bigcap_{i=2}^{m}\{E_{b_i+1} - E_{b_{i-1}+1} > s_i\} \right)$$
$$= \int_0^\infty dt_1 \int_{t_1+s_1}^\infty dt_2 \ldots \int_{t_m+s_m}^\infty dt_{m+1} \left(\frac{d}{dt_1}\mathbb{P}\left(M(T\big|_{b_1}) \leq t_1 \right) \right) e^{-t_2}\ldots e^{-t_{m+1}}$$
$$= \int_0^\infty dt_1 \int_{t_1+s_1}^\infty dt_2 \ldots \int_{t_{m-1}+s_{m-1}}^\infty dt_m \left(\frac{d}{dt_1}\mathbb{P}\left(M(T\big|_{b_1}) \leq t_1 \right) \right) e^{-t_2}\ldots e^{-2t_m}e^{-s_m}$$
$$\vdots$$
$$= \frac{\exp\{-\langle (m:1), \mathbf{s}\rangle\}}{m!} \int_0^\infty e^{-mt_1} \frac{d}{dt_1}\mathbb{P}\left(M(T\big|_{b_1}) \leq t \right) \, dt_1.$$

From (10) and making $p = e^{-t}$ in the above integral, and putting all together we obtain (25). Finally (26) follows from

$$\mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}, \ell_{n,b_1-1} = 0\right) = \mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}\right) - \mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}, \ell_{n,b_1-1} > 0\right)$$

and, recursively,

$$\mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}, \bigcap_{n/2<b<b_1} \{\ell_{n,b} = 0\} \right) = \mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}\right) - \sum_{n/2<b<b_1} \mathbb{P}\left(\Lambda_{\mathbf{b},\mathbf{s}}, \ell_{n,b} > 0, \bigcap_{i=1}^{b_1-b-1} \{\ell_{n,b+i} = 0\} \right).$$

Substituting (25) in the above expression, we obtain (26). □

Aknowledgements Alejandro and Arno would like to thank Geronimo Uribe Bravo for his fruitful suggestions on the second moment method.

References

[1] M. Desai, A. Walczak and D. Fisher, Genetic diversity and the structure of genealogies in rapidly adapting populations, *Genetics* **193** (2013), 565–585.

[2] C.S. Diehl and G. Kersting, Tree lengths for general Λ-coalescents and the asymptotic site frequency spectrum around the Bolthausen-Sznitman coalescent, to appear in *Ann. Appl. Probab.*, *Preprint on Arxiv.*

[3] B. Eldon, M. Birkner, J. Blath and F. Freund, Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?, *Genetics* **199** (2015), 841–856.

[4] F. Freund and A. Siri-Jégousse, Distinguishimg coalescent models - which statistics matter most? *Preprint on Biorxiv.*

[5] Y. Fu, Statistical properties of segregating sites, *Theor. Pop. Biol.* **48** (1995), 172–197.

[6] C. Goldschmidt and J.B. Martin, Random recursive trees and the Bolthausen-Sznitman coalescent, *Electron. J. Probab.* **10** (2005), 718–745.

[7] A. Hobolth, A. Siri-Jégousse and M. Bladt, Phase-type distributions in population genetics, *Theor. Pop. Biol.* **127** (2019), 16–32.

[8] A. Iksanov and M. Möhle, A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree, *Electron. Commun. Probab.* **12** (2007), 28–35.

[9] G. Kersting, J.C. Pardo and A. Siri-Jégousse, Total internal and external lengths of the Bolthausen-Sznitman coalescent, *J. Appl. Probab.* **51A** (2014), 73–86.

[10] J. Koskela, Multi-locus data distinguishes between population growth and multiple merger coalescents. *Stat. Appl. Genet. Mol. Biol.* **17** (2018).

[11] J. Kukla and H.H. Pitters, A spectral decomposition for the Bolthausen-Sznitman coalescent and the Kingman coalescent, *Electron. Commun. Probab.* **20** (2015), paper no. 87.

[12] M. Möhle andd H.H. Pitters, A spectral decomposition for the block counting process of the Bolthausen-Sznitman coalescent, *Electron. Commun. Probab.* **19** (2014), paper no. 47.

[13] R.H. Neher and O. Hallatschek, Genealogies of rapidly adapting populations, *Proc. Nat. Acad. Sci. USA* **110** (2013), 437–442.

[14] J. Pitman and M. Yor, The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator, *Ann. Probab.* **25** (1997), 855–900.

[15] J. Schweinsberg, Coalescent processes obtained from supercritical Galton-Watson processes, *Stoch. Proc. Appl.* **106** (2003), no. 1, 107–139.

[16] J. Schweinsberg, Rigorous results for a population model with selection II: genealogy of the population, *Electron. J. Probab.* **22** (2017), paper no. 38.

[17] J.P. Spence, J.A. Kamm and Y.S. Song, The site frequency spectrum for general coalescents, *Genetics* **202** (2016), 1549–1561.

[18] F.G. Tricomi and A. Erdélyi, The asymptotic expansion of a ratio of Gamma functions, *Pacific J. Math.* **1** (1951), 133–142.