The external lengths in Kingman's coalescent

Svante Janson* Götz Kersting†

2 February, 2011

Abstract

In this paper we prove asymptotic normality of the total length of external branches in Kingman's coalescent. The proof uses an embedded Markov chain, which can be described as follows: Take an urn with n *black* balls. Empty it in n steps according to the rule: In each step remove a randomly chosen pair of balls and replace it by one *red* ball. Finally remove the last remaining ball. Then the numbers U_k, $0 \le k \le n$, of red balls after k steps exhibit an unexpected property: $(U_0, \dots, U_n)$ and $(U_n, \dots, U_0)$ are equal in distribution.

MSC 2000 subject classifications. 60K35, 60F05, 60J10

Key words and phrases. coalescent, external branch, time reversibility, urn model

1 Introduction and results

Our main result in this paper is that the total length L_n of all external branches in Kingman's coalescent with n external branches is asymptotically normal for $n \to \infty$.

Kingman's coalescent (1982) consists of two components. First there are the coalescent times $T_1 > T_2 > \cdots > T_n = 0$. They are such that

$$\binom{k}{2}(T_{k-1} - T_k) \ , \quad k = 2, \dots, n$$

are independent, exponential random variables with expectation 1. Second there are partitions $\pi_1 = \big\{\{1, \dots, n\}\big\}, \pi_2, \dots, \pi_n = \big\{\{1\}, \dots, \{n\}\big\}$ of the set $\{1, \dots, n\}$, where the set π_k containes k disjoint subsets of $\{1, \dots, n\}$ and π_{k-1} evolves from π_k by merging two randomly chosen elements of π_k. Moreover, $(T_n, \dots, T_1)$ and $(\pi_n, \dots, \pi_1)$ are independent. For convenience we put $\pi_0 := \emptyset$.

*Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden. `svante.janson@math.uu.se`

†Fachbereich Informatik und Mathematik, Universität Frankfurt, Fach 187, D-60054 Frankfurt am Main, Germany. `kersting@math.uni-frankfurt.de`

As is customary the coalescent can be represented by a tree with n leaves labelled from 1 to n. Each of these leaves corresponds to an external branch of the tree. The other node of the branch with label i is located at level

$$\rho(i) := \max\{k \geq 1 : \{i\} \notin \pi_k\}$$

within the coalescent. The length of this branch is $T_{\rho(i)}$, The total external length of the coalescent is given by

$$L_n := \sum_{i=1}^{n} T_{\rho(i)} \ .$$

This quantity is of a certain statistical interest. Coalescent trees have been introduced by Kingman as a model for the genealogic relationship of n individuals, down to their most recent common ancestor. Mutations can be located everywhere on the branches. Then mutations on external branches affect only single individuals. This fact was used by Fu and Li (1993) in designing their D-statistic and providing a test whether or not data fit to Kingman's coalescent.

Else the total external length has been studied by Möhle (2010) for a kind of coalescent, which differs substantially from Kingman's coalescent. Otherwise single external branches have mainly been investigated in the literature. The asymptotic distribution of $T_{\rho(i)}$ has been obtained by Caliebe et al (2007), using a representation of its Laplace transform due to Blum and François (2005). We address this issue in Section 6 below. Freund and Möhle (2009) studied the external branch length of the Bolthausen-Snitman coalescent, and Gnedin et al (2008) the Λ-coalescent.

Here is our main result.

Theorem 1. *As $n \to \infty$,*

$$\frac{1}{2}\sqrt{\frac{n}{\log n}}\big(L_n - 2\big) \ \stackrel{d}{\to}\ N(0,1) \ .$$

The proof will show that the limiting normal distribution originates from the random partitions and not from the exponential waiting times.

A second glance on this result reveals a peculiarity: The normalization of L_n is carried out using its expectation, but only half of its variance. These two terms have been determined by Fu and Li (1993) (with a correction given by Durrett (2002)). They obtained

$$\mathbf{E}(L_n) = 2 \ , \quad \mathbf{Var}(L_n) = \frac{8nh_n - 16n + 8}{(n-1)(n-2)} \sim \frac{8 \log n}{n}$$

with $h_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n}$, the n-th harmonic number. Below we derive a more general result.

To uncover this peculiarity we shall study the external lengths in more detail. First we look at the point processes η_n on $(0, \infty)$, given by $\eta_n = \sum_{i=1}^{n} \delta_{\sqrt{n}T_{\rho(i)}}$, i.e.

$$\eta_n(B) := \#\{i : \sqrt{n}T_{\rho(i)} \in B\} \tag{1}$$

for Borel sets $B \subseteq (0, \infty)$.

Theorem 2. *As $n \to \infty$ the point process η_n converges in distribution, as point processes on $(0, \infty]$, to a Poisson point process η on $(0, \infty)$ with intensity measure $\lambda(dx) = 8x^{-3}\, dx$.*

We use $(0, \infty]$ in the statement of Theorem 2 instead of $(0, \infty)$ since it is stronger, including for example $\eta_n(a, \infty) \xrightarrow{\mathrm{d}} \eta(a, \infty)$ for every $a > 0$. The significance is that, as $n \to \infty$, there will be points clustering at 0 but not at ∞. (Below in the proof we recall the definition of convergence in distribution of point processes.)

Theorem 2 permits a first orientation. Since $\sqrt{n} L_n = \int x\, \eta_n(dx)$, one is tempted to resort to infinitely divisible distributions. However, the intensity measure $\lambda(dx)$ is slightly outside the range of the Lévy-Chintchin formula. Shortly speaking this means that small points of η_n have a dominant influence on the distribution of L_n and we are within the domain of the normal distribution.

Thus let us look in more detail on the external lengths and focus on

$$L_n^{\alpha,\beta} := \sum_{n^\alpha \leq \rho(i) < n^\beta} T_{\rho(i)} \ , \quad 0 \leq \alpha < \beta \leq 1 \ ,$$

which is the total length of those external branches having their internal nodes between level $\lceil n^\alpha \rceil$ and $\lceil n^\beta \rceil$ within the coalescent. Obviously $L_n = L_n^{0,1}$.

Proposition 3. *For $0 \leq \alpha < \beta \leq 1$*

$$\mathbf{E}(L_n^{\alpha,\beta}) = \frac{2}{n(n-1)} \big(\lceil n^\beta \rceil - \lceil n^\alpha \rceil\big)\big(2n + 1 - \lceil n^\beta \rceil - \lceil n^\alpha \rceil\big)$$

and

$$\mathbf{Var}(L_n^{\alpha,\beta}) \sim 8(\beta - \alpha) \frac{\log n}{n} \ ,$$

as $n \to \infty$.

In particular $\mathbf{E}(L_n^{1-\varepsilon,1}) \sim \mathbf{E}(L_n^{0,1})$, whereas $\mathbf{Var}(L_n^{1-\varepsilon,1}) \sim \varepsilon \mathbf{Var}(L_n^{0,1})$. Thus the proposition indicates that the systematic part of L_n and its fluctuations arise in different regions of the coalescent tree, the former close to the leaves and the latter closer to the root.

Still this proposition gives an inadequate impression.

Theorem 4. *For $0 \leq \alpha < \beta < 1/2$*

$$\mathbf{P}(L_n^{\alpha,\beta} = 0) \ \to \ 1$$

as $n \to \infty$. Moreover

$$\sqrt{n} L_n^{0,\frac{1}{2}} \ \xrightarrow{d} \ \int_2^\infty x\, \eta(dx)$$

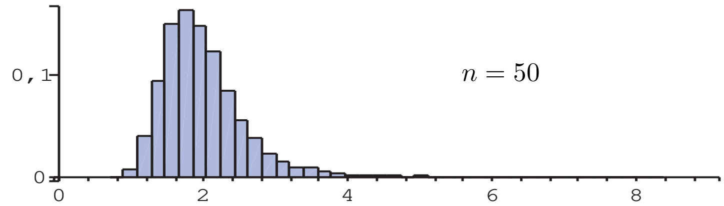
and for $1/2 \leq \alpha < \beta \leq 1$

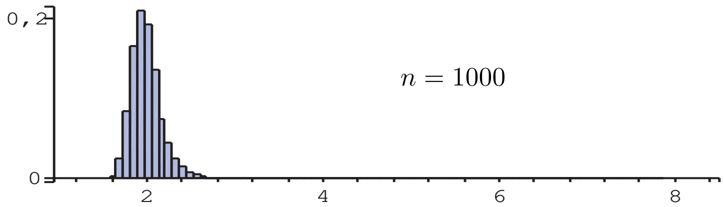
$$\frac{L_n^{\alpha,\beta} - \mathbf{E}(L_n^{\alpha,\beta})}{\sqrt{\mathbf{Var}(L_n^{\alpha,\beta})}} \ \xrightarrow{d} \ N(0,1) \ .$$

In addition $L_n^{\alpha,\beta}$ and $L_n^{\gamma,\delta}$ are asymptotically independent for $\alpha < \beta \leq \gamma < \delta$.

This result implies Theorem 1: In $L_n = L_n^{0,\frac{1}{2}} + L_n^{\frac{1}{2},1}$ the summands are of order $\sqrt{1/n}$ and $\sqrt{\log n/n}$, such that in the limit the second, asymptotically normal component dominates. To this end, however, n has to become exponentially large, otherwise the few long branches, which make up $L_n^{0,\frac{1}{2}}$, cannot be neglected and may produce extraordinary large values of L_n. Thus the normal approximation for the distribution of L_n seems little useful for practical purposes. One expects a fat right tail compared to the normal distribution. Indeed $\int_2^\infty x\,\eta(dx)$ has finite mean but infinite variance.

This is illustrated by the following two histograms from 10000 values of L_n, where the length of the horizontal axis to the right indicates the range of the values.





The heavy tails to the right are clearly visible. Also very large outliers appear: For $n = 50$ the simulated values of L_n range from 0.685 to 8.38, and for $n = 1000$ from 1.57 to 7.87.

Also it turns out that the approximation of the variance in Proposition 3 is good only for very large n. This can be seen already from the formula of Fu and Li. To get an exact formula for the variance we look at a somewhat different quantity, namely

$$\hat{L}_n^{\alpha,\beta} := \sum_{i=1}^{n} (T_{\rho(i)} \wedge T_{\lfloor n^\alpha \rfloor} - T_{\rho(i)} \wedge T_{\lfloor n^\beta \rfloor})$$

with $0 \leq \alpha < \beta \leq 1$, which is the portion of the external length between level $\lfloor n^\alpha \rfloor$ and $\lfloor n^\beta \rfloor$ within the coalescent.

Proposition 5. *For $0 \leq \alpha \leq 1$ with $m := \lfloor n^\alpha \rfloor$*

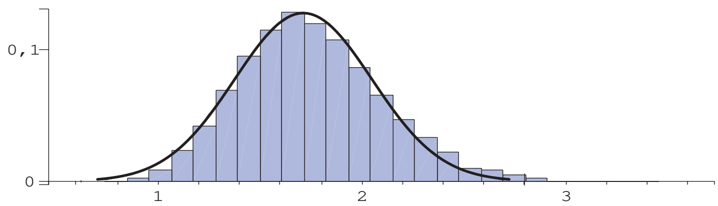
$$\mathbf{E}(\hat{L}_n^{\alpha,1}) = 2\frac{n-m}{n-1}$$

and

$$\mathbf{Var}(\hat{L}_n^{\alpha,1}) = \frac{8(h_{n-1} - h_{m-1})(n + 2m - 2)}{(n-1)(n-2)} - \frac{4(n-m)(4n+m-5)}{(n-1)^2(n-2)} \ .$$

For $\alpha = 0$ we recover the formula of Fu and Li. A similar expression holds for $\hat{L}_n^{\alpha,\beta}$.

Proposition 3 and Theorem 4 carry over to $\hat{L}_n^{\alpha,\beta}$, up to a change in expectation and with the limit $\sqrt{n}\hat{L}_n^{0,\frac{1}{2}} \xrightarrow{\mathrm{d}} \int_2^\infty (x-2)\,\eta(dx)$. The following histogram from a random sample of length 10000 shows that already for $n = 50$ the distribution of $\hat{L}_n^{\frac{1}{2},1}$ fits well to the normal distribution when using the values for expectation and variance, given in Proposition 5.



Our main tool for the proofs is a representation of L_n by means of an imbedded Markov chain $U_0, U_1, \ldots, U_n$, which is of interest of its own. We shall introduce it as an urn model. The relevant fact is that this model possesses an unexpected hidden symmetry, namely it is reversible in time. This is our second main result. For the proof we use another urn model, which allows reversal of time in a simple manner.

The urn models are introduced and studied in Section 2. Proposition 3 is proven in Section 3, Theorems 2 and 4 are derived in Section 4 and Proposition 5 in Section 5. In Section 6 we complete the paper by considering the length of an external branch chosen at random.

2 The urn models

Take an urn with n *black* balls. Empty it in n steps according to the rule: In each step remove a randomly chosen pair of balls and replace it by one *red* ball. In the last step remove the last remaining ball. Let

$$U_k := \text{ number of red balls in the urn after } k \text{ steps .}$$

Obviously $U_0 = U_n = 0$, $U_1 = U_{n-1} = 1$ and $1 \le U_k \le \min(k, n-k)$ for $2 \le k \le n-2$. $U_0, \ldots, U_n$ is a Markov chain with transition probabilities

$$\mathbf{P}(U_{k+1} = u' \mid U_k = u) = \begin{cases} \binom{u}{2}\big/\binom{n-k}{2} \ , & \text{if } u' = u-1 \ , \\ u(n-k-u)\big/\binom{n-k}{2} \ , & \text{if } u' = u \ , \\ \binom{n-k-u}{2}\big/\binom{n-k}{2} \ , & \text{if } u' = u+1 \ . \end{cases}$$

We begin our study of the model by calculating expectations and covariances.

Proposition 6. *For* $0 \le k \le l \le n$

$$\mathbf{E}(U_k) = \frac{k(n-k)}{n-1} \ , \quad \mathbf{Cov}(U_k, U_l) = \frac{k(k-1)(n-l)(n-l-1)}{(n-1)^2(n-2)} \ .$$

Proof. Imagine that the black balls are numbered from 1 to n. Let Z_{ik} be the indicator variable of the event that the black ball with number i is not yet removed after k steps. Then $U_k = n - k - \sum_{i=1}^n Z_{ik}$ and consequently

$$\mathbf{E}(U_k) = n - k - n\mathbf{E}(Z_{1k})$$

and for $k \le l$ in view of $Z_{1l} \le Z_{1k}$

$$\begin{aligned}\mathbf{Cov}(U_k, U_l) &= \sum_{i=1}^n \sum_{j=1}^n \mathbf{Cov}(Z_{ik}, Z_{jl}) \\ &= n(n-1)\mathbf{E}(Z_{1k}Z_{2l}) + n\mathbf{E}(Z_{1l}) - n^2\mathbf{E}(Z_{1k})\mathbf{E}(Z_{1l}) \ .\end{aligned}$$

Also

$$\mathbf{P}(Z_{1k} = 1) = \frac{\binom{n-1}{2}}{\binom{n}{2}} \cdots \frac{\binom{n-k}{2}}{\binom{n-k+1}{2}} = \frac{(n-k)(n-k-1)}{n(n-1)}$$

and for $k \le l$

$$\begin{aligned}\mathbf{P}(Z_{1k} = 1, Z_{2l} = 1) &= \frac{\binom{n-2}{2}}{\binom{n}{2}} \cdots \frac{\binom{n-k-1}{2}}{\binom{n-k+1}{2}} \cdot \frac{\binom{n-k-1}{2}}{\binom{n-k}{2}} \cdots \frac{\binom{n-l}{2}}{\binom{n-l+1}{2}} \\ &= \frac{(n-k-1)(n-k-2)(n-l)(n-l-1)}{n(n-1)^2(n-2)} \ .\end{aligned}$$

Our claim now follows by careful calculation. □

Note that these expressions for expectations and covariances are invariant under the transformation $k \mapsto n - k$, $l \mapsto n - l$. This is not by coincidence:

Theorem 7. *$(U_0, U_1, \ldots, U_n)$ and $(U_n, U_{n-1}, \ldots, U_0)$ are equal in distribution.*

Proof. Leaving aside $U_0 = U_n = 0$ we have $U_k \ge 1$ *a.s.* for the other values of k. Instead we shall look at $U'_k = U_k - 1$ for $1 \le k \le n - 1$. It turns out that for this process one can specify a different dynamics, which is more lucid and amenable to reversing time.

Consider the following alternative box scheme: There are two boxes A and B. At the beginning A contains $n - 1$ black balls whereas B is empty. The balls are converted in $2n - 2$ steps into $n - 1$ red balls lying in B. Namely, in the steps number $1, 3, \ldots, 2n - 3$ a randomly drawn ball from A is shifted to B and in the steps number $2, 4, \ldots, 2n - 2$ a randomly chosen black ball (whether from A or B) is recolored to a red ball. These $2n - 2$ operations are carried out independently.

For $1 \le k \le n - 1$ let

$$U'_k := \text{number of red balls in box } A \text{ after } 2k - 1 \text{ steps,}$$

that is at the moment after the kth move and before the kth recoloring. Obviously the sequence is a Markov chain, also $U'_1 = 0$.

As to the transition probabilities note that after $2k-1$ steps there are $n-k$ black balls in all and $n-k-1$ balls in A. Thus given $U'_k = r$ there are r red and $n-k-r-1$ black balls in A, and the remaining $r+1$ black balls belong to B. Then $U'_{k+1} = r+1$ occurs only, if in the next step the ball recolored from black to red belongs to A and subsequently the ball shifted from A to B is black. Thus

$$\mathbf{P}(U'_{k+1} = r+1 \mid U'_k = r) = \tfrac{n-k-r-1}{n-k} \cdot \tfrac{n-k-r-2}{n-k-1} = \tbinom{n-k-r-1}{2} / \tbinom{n-k}{2} \ .$$

Similarly $U'_{k+1} = r-1$ occurs, if the recolored ball belongs to B and next the ball shifted from A to B is red. The corresponding probability is

$$\mathbf{P}(U'_{k+1} = r-1 \mid U'_k = r) = \tfrac{r+1}{n-k} \cdot \tfrac{r}{n-k-1} = \tbinom{r+1}{2} / \tbinom{n-k}{2} \ .$$

Since $U_1 = 1 = U'_1 + 1$ and in view of the transition probabilities of (U_k) and (U'_k) we see that $(U_1, \dots, U_{n-1})$ and $(U'_1 + 1, \dots, U'_{n-1} + 1)$ indeed coincide in distribution.

Next note that $U'_{n-1} = 0$. Therefore U'_k can be considered as a function not only of the first $2k-1$ but also of the last $2n-2k-1$ shifting and recoloring steps. Since the steps are independent, the process backwards is equally easy to handle. Taking into account that backwards the order of moving and recoloring balls is interchanged, one may just repeat the calculations above to obtain reversibility.

But this repetition can be avoided as well. Let us put our model more formally: Label the balls from 1 to $n-1$ and write the state space as

$$S := \big\{ \big((L_1, c_1), \dots, (L_{n-1}, c_{n-1}) \big) \mid L_i \in \{A, B\}, c_i \in \{b, r\} \big\} \ ,$$

where L_i is the location of ball i and c_i its color. Then in our model the first and second coordinate are changed in turn from A to B and from b to r. This is done completely at random, starting within the first coordinates. Clearly we may interchange the role of the first and second coordinate. Thus our box model is equivalent to the following version:

Again initially A contains $n-1$ black balls whereas B is empty. Now in the steps number $1, 3, \dots, 2n-3$ a randomly chosen black ball is recolored to a red ball and in the steps number $2, 4, \dots, 2n-2$ a randomly drawn ball from A is shifted to B. Again these $2n-2$ operations are carried out independently. Here we consider

$$U''_k := \text{number of black balls in box } B \text{ after } 2k-1 \text{ steps.}$$

Then from the observed symmetry it is clear that $(U'_1, \dots, U'_{n-1})$ and $(U''_1, \dots, U''_{n-1})$ are equal in distribution.

If we finally interchange both colors and boxes as well, then we arrive at the dynamics of the backward process. This finishes the proof. $\square$

There is a variant of our proof, which makes the reversibility of (U'_k) manifest in a different manner. Let again the balls be labelled from 1 to $n-1$. Denote

$$\nu_m := \text{instance between } 1 \text{ and } n-1, \text{ when ball } m \text{ is colored to red,}$$
$$\sigma_m := \text{instance between } 1 \text{ and } n-1, \text{ when ball } m \text{ is shifted to box } B.$$

Then from our construction it is clear that $\nu = (\nu_m)$ and $\sigma = (\sigma_m)$ are two independent random permutations of the numbers $\{1, \dots, n-1\}$. Moreover, at instance k (i.e. after $2k-1$ steps) ball number m is red and belongs to box A, if it was colored before and shifted afterwards, i.e. $\nu_m < k < \sigma_m$. Thus we obtain the formula

$$U'_k = \#\{1 \le m \le n-1 : \nu_m < k < \sigma_m\} \tag{2}$$

and we may conclude the following result.

Corollary 8. *Let ν and σ be two independent random permutations of $\{1, \dots, n-1\}$. Then $(U_1, \dots, U_{n-1})$ is equal in distribution to the process*

$$\big(\#\{1 \le m \le n-1 : \nu_m < k < \sigma_m\} + 1\big)_{1 \le k \le n-1} .$$

Certainly this representation implies Theorem 7 again. Also it contains additional information. For example, it is immediate that $U_k - 1$ has a hypergeometric distribution with parameters $n-1, k-1, n-k-1$.

The next example contains a first application of Theorem 7 to our original urn model.

Example. Let us consider $\tau_n = \max\{k \ge 1 : U_{n-k} = k\}$, the number of red balls in the urn, after the last black ball has been removed. From reversibility τ_n has the same distribution as the moment $\tau'_n = \max\{k \ge 1 : U_k = k\}$, before the first red ball is taken away from the urn. Thus

$$\mathbf{P}(\tau_n \ge k) = \frac{\binom{n-2}{2}}{\binom{n-1}{2}} \frac{\binom{n-4}{2}}{\binom{n-2}{2}} \cdots \frac{\binom{n-2k+2}{2}}{\binom{n-k+1}{2}} = \frac{(n-k)\cdots(n-2k+1)}{(n-1)\cdots(n-k)} .$$

It follows for $t \ge 0$

$$\mathbf{P}\Big(\frac{\tau_n}{\sqrt{n}} \ge t\Big) \to \exp(-t^2) ,$$

as $n \to \infty$. □

More generally the dynamics of our urn looks as follows: Clearly, if n is large, then in the beginning always two black balls are removed from the urn. The rare moments, when red balls are taken away, appear with increasing rate. Indeed it is not difficult to see that in the limit $n \to \infty$ and after a $\sqrt{n}$-scaling of time these instances build up a Poisson process with linearly increasing rate. As we have seen the picture remains the same after reversal of time. This will be made more precise in Section 4.

We conclude this section by imbedding our urn model into the coalescent. Let

$$V_k := k - \#\{i : \rho(i) < k\} , \tag{3}$$

and $U_k := V_{n-k}$, $0 \le k \le n$. Thus V_k is the number of internal branches among the k branches after the $(n-k)$-th coalescing event and U_k is the number of internal branches among the $n-k$ branches after the k-th coalescing event. The coalescing mechanism takes two random branches and combines them into one internal branch. If we code the external branches by black balls and the internal branches by red, this completely conforms to our urn model; thus $(U_0,\dots,U_n)$ is as above. By Theorem 7, $(V_0,\dots,V_n)$ has the same distribution as $(U_0,\dots,U_n)$. In the next sections we make use of the Markov chain $V_0,\dots,V_n$ and its properties.

Remark. For a different interpretation of the process (U_k), suppose that we have $n-1$ pairs of (different) shoes, and that all left shoes are mixed in one pile and all right shoes in another. We sort the shoes by taking first a left shoe (at random), then a right shoe (also at random), then another left shoe, and so on. As soon as we take a shoe that matches one that we already have picked, we put away the pair; otherwise we put the shoe on the table in front of us. If the pairs are numbered and ν_m is the time right shoe m is picked, and σ_m the time left shoe m is picked, then right shoe m is on the table when the k-th left shoe has been picked if and only if $\nu_m < k < \sigma_m$, so by (2), the number of right shoes remaining on the table when the k-th left shoe has been picked is U'_k, $1 \le k \le n-1$. The number of left shoes remaining on the table at the same time is $U'_k + 1 = U_k$, so the total number of shoes on the table is $2U_k - 1$.

This is a variation of the sock-sorting process studied in Steinsaltz (1999) and Janson (2009), Section 8, which is similar except that there is no difference between left and right; we obtain it if we mix all shoes in one pile and pick from it at random. (See Janson (2009) for other interpretations, including *priority queues*, and further references.) It is not surprising that we have the same asymptotical behaviour of U_k and $\max_k U_k$ as for the sock-sorting problem. In particular, we mention the following Gaussian process limit result, cf. Theorem 8.2 in Janson (2009). (This result is not used in the sequel.)

Theorem 9. *As $n \to \infty$, the stochastic process $n^{-1/2}\bigl(U_{\lfloor nt \rfloor} - nt(1-t)\bigr)$ converges in $D[0,1]$ to a continuous Gaussian process $Z(t)$ with mean $\mathbf{E}(Z(t)) = 0$ and covariance function*

$$\mathbf{E}\bigl(Z(s)Z(t)\bigr) = s^2(1-t)^2, \qquad 0 \le s \le t \le 1.$$

Sketch of proof. Note first that $\mathbf{E}(U_{\lfloor nt \rfloor}) = nt(1-t) + O(1)$ by Proposition 6.

It is easily seen that

$$\mathbf{E}(U_{k+1} \mid U_k) = U_k - \frac{2}{n-k}U_k + 1 = \frac{n-k-2}{n-k}U_k + 1$$

and it follows that

$$M_k := \frac{U_k - \mathbf{E}(U_k)}{(n-k)(n-k-1)} = \frac{U_k}{(n-k)(n-k-1)} - \frac{k}{(n-1)(n-k-1)},$$

$k = 0,1,\dots,n-2$, is a martingale.

Consider in the sequel only $k \le (1-\delta)n$ for some fixed $\delta > 0$. Then $\mathbf{Var}(M_k) \le (n-k-1)^{-4}\mathbf{Var}(U_k) = O(n^{-3})$, and it follows from Doob's inequality that

$$\max_k |U_k - \mathbf{E}(U_k)| = O_P(n^{1/2}).$$

(Using Theorem 7 we see that this extends to $0 \le k \le n$.) A straightforward computation of the conditional quadratic variation $\langle M, M\rangle_m := \sum_{k<m} \mathbf{E}\big((M_{k+1}-M_k)^2 \mid U_k\big)$ shows that, uniformly in $0 \le t \le 1-\delta$,

$$n^3 \langle M, M\rangle_{\lfloor nt \rfloor} \xrightarrow{\text{p}} \frac{t^2}{(1-t)^2},$$

which implies, see Theorem VIII.3.11 in Jacod and Shiryaev (1987), that $n^{3/2} M_{\lfloor nt \rfloor} \xrightarrow{\text{d}} \hat{Z}(t)$ in $D[0, 1-\delta]$, where $\hat{Z}(t)$ is a Gaussian martingale given by $\hat{Z}(t) = W(t^2/(1-t)^2)$ for a standard Brownian motion $W(t)$. The result follows, for $t \in [0, 1-\delta]$, with $Z(t) = (1-t)^2 \hat{Z}(t)$.

Since $\delta > 0$ is arbitrary, this yields convergence in $D[0, 1)$. By time-reversal and Theorem 7, we also have convergence in $D(0, 1]$, and together these imply convergence in $D[0, 1]$, see e.g. the proof in Janson (2009). □

3 Proof of Proposition 3

We use the representation

$$L_n^{\alpha,\beta} = \sum_{n^\alpha \le k < n^\beta} T_k X_k \ ,$$

where

$$X_k := \#\{i : \rho(i) = k\} \ ,$$

$1 \le k < n$. In view of the coalescing procedure X_k takes only the values $0, 1, 2$, and from the definition (3) of V_k

$$X_k = 1 + V_k - V_{k+1} \ . \tag{4}$$

From (4), $V_k = U_{n-k}$ and Proposition 6 we obtain after simple calculations

$$\mathbf{E}(X_k) = \frac{2k}{n-1} \ , \quad \mathbf{Var}(X_k) = \frac{2k(n-k-1)(n-3)}{(n-1)^2(n-2)} \tag{5}$$

and for $k < l$

$$\mathbf{Cov}(X_k, X_l) = -\frac{4k(n-l-1)}{(n-1)^2(n-2)} \ . \tag{6}$$

Also from $T_k = \sum_{j=k+1}^n (T_{j-1} - T_j)$ we have $\mathbf{E}(T_k) = 2\sum_{j=k+1}^n \frac{1}{(j-1)j}$ and $\mathbf{Var}(T_k) = 4\sum_{j=k+1}^n \frac{1}{(j-1)^2 j^2}$; thus

$$\mathbf{E}(T_k) = 2\Big(\frac{1}{k} - \frac{1}{n}\Big) \ , \quad \mathbf{Var}(T_k) \le \frac{c}{k^3} \tag{7}$$

for a suitable $c > 0$, independent of n.

Thus from independence

$$\mathbf{E}(L_n^{\alpha,\beta}) = \sum_{n^\alpha \le k < n^\beta} 2\Big(\frac{1}{k} - \frac{1}{n}\Big)\frac{2k}{n-1} \ .$$

Now the first claim follows by simple computation.

Further from independence

$$\mathbf{Var}\Big(\sum_{n^\alpha \le k < n^\beta} (T_k - \mathbf{E}(T_k))X_k\Big) = \sum_{n^\alpha \le k,l < n^\beta} \mathbf{Cov}(T_k, T_l)\mathbf{E}(X_k X_l) \ . \tag{8}$$

Using (5)–(7) we have for $k < l$,

$$\mathbf{Cov}(T_k, T_l)\mathbf{E}(X_k X_l) = \mathbf{Var}(T_l)\mathbf{E}(X_k X_l) \le \mathbf{Var}(T_l)\mathbf{E}(X_k)\mathbf{E}(X_l) \le \frac{c}{l^3} \cdot \frac{4kl}{(n-1)^2},$$

and it follows that

$$\begin{aligned} 0 \le \sum_{n^\alpha \le k < l < n^\beta} \mathbf{Cov}(T_k, T_l)\mathbf{E}(X_k X_l) &\le \sum_{n^\alpha \le k < l < n^\beta} \frac{4ck}{l^2}(n-1)^{-2} \\ &\le \sum_{n^\alpha \le k < n^\beta} 4c(n-1)^{-2} = O(n^{-1}) \ . \end{aligned}$$

Consequently, (8) yields, using again (5)–(7),

$$\begin{aligned} \mathbf{Var}\Big(\sum_{n^\alpha \le k < n^\beta} (T_k - \mathbf{E}(T_k))X_k\Big) &= \sum_{n^\alpha \le k < n^\beta} \mathbf{Var}(T_k)\mathbf{E}(X_k^2) + O(n^{-1}) \\ &\le c \sum_{n^\alpha \le k < n^\beta} \frac{1}{k^3}\Big(\frac{2k}{n-1} + \frac{4k^2}{(n-1)^2}\Big) + O(n^{-1}) \\ &\le \frac{6c}{n-1} \sum_{n^\alpha \le k < n^\beta} \frac{1}{k^2} + O(n^{-1}) = O(n^{-1}) \ . \end{aligned} \tag{9}$$

It remains to show that

$$\mathbf{Var}\Big(\sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)X_k\Big) \sim 8(\beta - \alpha)\frac{\log n}{n} \ .$$

Now

$$\begin{aligned} &\Big|\sum_{n^\alpha \le k < l < n^\beta} \mathbf{E}(T_k)\mathbf{E}(T_l)\mathbf{Cov}(X_k, X_l)\Big| \\ &\qquad \le \sum_{n^\alpha \le k < l < n^\beta} \frac{2}{k} \cdot \frac{2}{l} \cdot \frac{4k}{(n-1)^2} = 16 \sum_{n^\alpha < l < n^\beta} \frac{l - \lceil n^\alpha \rceil}{l(n-1)^2} = O(n^{-1}) \end{aligned}$$

and consequently

$$\begin{aligned} &\mathbf{Var}\Big(\sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)X_k\Big) \\ &\quad = \sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)^2 \mathbf{Var}(X_k) + O(n^{-1}) \\ &\quad = \sum_{n^\alpha \le k < n^\beta} \frac{4}{k^2} \cdot \frac{2k}{n}\Big(1 + O\Big(\frac{k}{n}\Big)\Big) + O(n^{-1}) = 8(\beta - \alpha)\frac{\log n}{n} + O(n^{-1}) \ . \end{aligned}$$

This gives our claim.

4 Proof of Theorems 2 and 4

In this section we use Theorem 7. Namely, $V_0, \dots, V_n$ is a Markov chain with transition probabilities, which can be expressed by means of $X_1, \dots, X_{n-1}$ as follows:

$$\mathbf{P}(X_k = x \mid V_k = v) = \begin{cases} \binom{n-k-v}{2} / \binom{n-k}{2} \ , & \text{if } x = 0 \ , \\ v(n-k-v)/\binom{n-k}{2} \ , & \text{if } x = 1 \ , \\ \binom{v}{2}/\binom{n-k}{2} \ , & \text{if } x = 2 \ . \end{cases}$$

We like to couple these random variables with suitable independent random variables taking values 0 or 1. Note that V_k takes only values $v \le k$, thus for $k \le n/3$

$$\binom{n-k-v}{2} \Big/ \binom{n-k}{2} \ge \binom{n-2k}{2} \Big/ \binom{n-k}{2} \ge \frac{n-3k}{n-k} \ .$$

Therefore we may enlarge our model by means of random variables Y_k, $k \le n/3$, such that

$$\begin{aligned} &\mathbf{P}(X_k = x, \, Y_k = y \mid V_k = v, V_{k-1}, \dots, V_0, Y_{k-1}, \dots, Y_1) \\ &\qquad = \begin{cases} \frac{n-3k}{n-k} \ , & \text{if } x = 0, y = 0 \ , \\ \binom{n-k-v}{2}/\binom{n-k}{2} - \frac{n-3k}{n-k} \ , & \text{if } x = 0, y = 1 \ , \\ v(n-k-v)/\binom{n-k}{2} \ , & \text{if } x = 1, y = 1 \ , \\ \binom{v}{2}/\binom{n-k}{2} \ , & \text{if } x = 2, y = 1 \ . \end{cases} \end{aligned}$$

For $\mathbf{P}(X_k = x \mid V_k = v)$ this gives the above formula, whereas

$$\mathbf{P}(Y_k = y \mid V_k = v, V_{k-1}, \dots, V_0, Y_{k-1}, \dots, Y_1) = \begin{cases} \frac{n-3k}{n-k} \ , & \text{if } y = 0 \ , \\ \frac{2k}{n-k} \ , & \text{if } y = 1 \ . \end{cases}$$

This means that the 0/1-valued random variables Y_k, $k \le n/3$, are independent. For convenience we put $Y_k = 0$ for $k > n/3$. A straightforward computation gives

$$\mathbf{E}(Y_k - X_k \mid V_k = v) = \frac{2(k-v)}{n-k} \ , \tag{10}$$

$$\begin{aligned} \mathbf{E}((Y_k - X_k)^2 \mid V_k = v) &= \frac{2(k-v)}{n-k} + \frac{2v(v-1)}{(n-k)(n-k-1)} \\ &\le \frac{2(k-v)}{n-k} + \frac{2k(k-1)}{(n-k)(n-k-1)} \end{aligned} \tag{11}$$

for $k \le n/3$. Since $k - \mathbf{E}(V_k) = k(k-1)/(n-1)$ from Proposition 6, it follows

$$\mathbf{E}((Y_k - X_k)^2) \le \frac{4k(k-1)}{(n-k)(n-k-1)} \ . \tag{12}$$

Proof of Theorem 2. Recall that, by (1) and (4),

$$\eta_n = \sum_{i=1}^{n} \delta_{\sqrt{n}T_{\rho(i)}} = \sum_{k=1}^{n-1} X_k \delta_{\sqrt{n}T_k}. \tag{13}$$

Recall also that $\eta_n \xrightarrow{\mathrm{d}} \eta$ as point processes on the interval $(0,\infty]$ means that $\int f\, d\eta_n \xrightarrow{\mathrm{d}} \int f\, d\eta$ for every continuous f with compact support in $(0,\infty]$, or equivalently $\eta_n(B) \xrightarrow{\mathrm{d}} \eta(B)$ for every relatively compact Borel subset B of $(0,\infty]$ such that $\eta(\partial B) = 0$ a.s. (Here B is relatively compact, if $B \subseteq [\delta,\infty]$ for some $\delta > 0$.) See, for example, the Appendix in Janson and Spencer (2007) and Chapter 16 (in particular Theorem 16.16) in Kallenberg (2002).

Let us first look at the point process

$$\eta_n' := \sum_{k=1}^{n-1} Y_k \delta_{2\sqrt{n}/k} \ . \tag{14}$$

For $0 < a < b \le \infty$

$$\eta_n'([a,b)) = \sum_{\frac{2\sqrt{n}}{b} < k \le \frac{2\sqrt{n}}{a}} Y_k$$

and

$$\mathbf{E}\big(\eta_n'([a,b))\big) = \sum_{\frac{2\sqrt{n}}{b} < k \le \frac{2\sqrt{n}}{a}} \frac{2k}{n-k} \to 4(a^{-2} - b^{-2}) = 8\int_a^b \frac{dx}{x^3} \ ,$$

thus we obtain from standard results on sums of independent 0/1-valued random variables that $\eta_n'([a,b))$ has asymptotically a Poisson distribution. Also $\eta_n'(B_1), \dots, \eta_n'(B_i)$ are independent for disjoint $B_1, \dots, B_i$. Therefore we obtain from standard results on point processes (for example Kallenberg (2002), Proposition 16.17) weak convergence of η_n' to the Poisson point process η on $(0,\infty]$ with intensity $8x^{-3}\,dx$.

Next we prove that for all $0 < a < b \le \infty$

$$\eta_n([a,b)) - \eta_n'([a,b)) \to 0$$

in probability. To this end note that from (12)

$$\mathbf{E}\Big[\sum_{k \le \frac{2\sqrt{n}}{a}} (Y_k - X_k)^2\Big] = O(n^{-1/2}) \ ,$$

which implies that $\mathbf{P}(X_k = Y_k \text{ for all } k \le \frac{2\sqrt{n}}{a}) \to 1$. Therefore we may well replace Y_k by X_k in $\eta_n'([a,b))$.

Also, by (7), $\sqrt{n}T_k - 2\sqrt{n}/k = \sqrt{n}T_k - \sqrt{n}\mathbf{E}(T_k) - 2/\sqrt{n}$. From (7) and Doob's inequality for any $\varepsilon > 0$

$$\mathbf{P}\big(\max_{k \ge n^{2/5}} \sqrt{n}|T_k - \mathbf{E}(T_k)| \ge \varepsilon\big) \le \frac{n}{\varepsilon^2}\mathbf{Var}(T_{\lceil n^{2/5}\rceil}) = O(n^{-1/5}) \ .$$

Since $\mathbf{P}(Y_k = 0 \text{ for all } k < n^{2/5}) \to 1$, we may as well also replace $2\sqrt{n}/k$ by $\sqrt{n}T_k$ in η_n', which yields η_n by (13) and (14) (use for example Kallenberg (2002), Theorem 16.16). Thus the proof of Theorem 2 is complete. $\square$

Proof of Theorem 4. As to the first claim of Theorem 4 observe that the events $\{L_n^{0,\beta} = 0\} = \{X_k = 0 \text{ for all } k < n^\beta\}$ and $\{V_{\lceil n^\beta \rceil} = \lceil n^\beta \rceil\}$ are equal. Thus

$$\begin{aligned}\mathbf{P}(L_n^{\alpha,\beta} > 0) &\leq \mathbf{P}(L_n^{0,\beta} > 0) = \mathbf{P}(\lceil n^\beta \rceil - V_{\lceil n^\beta \rceil} \geq 1) \\ &\leq \mathbf{E}(\lceil n^\beta \rceil - V_{\lceil n^\beta \rceil}) = \frac{\lceil n^\beta \rceil(\lceil n^\beta \rceil - 1)}{n-1} \ .\end{aligned}$$

For $\beta < 1/2$ this quantity converges to zero, which gives the first claim of the theorem.

For the next claim we use that because of (7) $\sqrt{n}T_{\lceil n^{1/2} \rceil}$ has expectation $2+O(n^{-1/2})$ and variance of order $n^{-1/2}$. Thus $\mathbf{P}(2-\varepsilon < \sqrt{n}T_{\lceil n^{1/2} \rceil} < 2+\varepsilon) \to 1$ for all $\varepsilon > 0$. This implies that the probability of the event

$$\begin{aligned}\int_{[2+\varepsilon,\infty)} x\, \eta_n(dx) &= \sqrt{n}\sum_{k=1}^n T_k X_k I_{\{\sqrt{n}T_k \geq 2+\varepsilon\}} \\ &\leq \sqrt{n} \sum_{k<\sqrt{n}} T_k X_k = \sqrt{n} L_n^{0,\frac{1}{2}} \\ &\leq \sqrt{n}\sum_{k=1}^n T_k X_k I_{\{\sqrt{n}T_k \geq 2-\varepsilon\}} = \int_{[2-\varepsilon,\infty)} x\, \eta_n(dx)\end{aligned}$$

goes to 1. Also for $a > 0$ from Theorem 2 $\int_a^\infty x\, \eta_n(dx) \to \int_a^\infty x\, \eta(dx)$ in distribution. Altogether we obtain, letting $\epsilon \to 0$,

$$\sqrt{n} L_n^{0,\frac{1}{2}} \to \int_2^\infty x\, \eta(dx) \ ,$$

which is our second claim.

As to the last claim of Theorem 4 we note that from (9)

$$L_n^{\alpha,\beta} = \sum_{n^\alpha \leq k < n^\beta} \mathbf{E}(T_k) X_k + O_P(n^{-1/2}) \tag{15}$$

in probability, and also in L^1. In this representation we like to replace X_k by Y_k. We assume first $\beta < 1$. Note that for $\beta < 1$ in view of (7) and (12)

$$\begin{aligned}&\mathbf{Var}\Big(\sum_{n^\alpha \leq k < n^\beta} \mathbf{E}(T_k)(Y_k - X_k - \mathbf{E}(Y_k - X_k \mid V_k))\Big) \\ &\qquad \leq \sum_{n^\alpha \leq k < n^\beta} \frac{4}{k^2} \mathbf{E}((Y_k - X_k)^2) = O(n^{\beta-2})\end{aligned}$$

and from (10), (7) and Proposition 6

$$\begin{aligned}
\mathbf{Var}\Big(\sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)\mathbf{E}(Y_k - X_k \mid V_k)\Big) &= \mathbf{Var}\Big(\sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)\frac{2V_k}{n-k}\Big) \\
&\le 2 \sum_{n^\alpha \le k \le l < n^\beta} 4\frac{\mathbf{E}(T_k)\mathbf{E}(T_l)}{(n-k)(n-l)}\mathbf{Cov}(V_k, V_l) \\
&\le 32 \sum_{n^\alpha \le k \le l < n^\beta} \frac{k}{l} \cdot \frac{(n-l)}{(n-k)(n-1)^2(n-2)} = O(n^{2\beta-3}) \ .
\end{aligned}$$

Thus $\sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)\big(Y_k - X_k) - \mathbf{E}(Y_k - X_k)\big) = O_P(n^{-1/2})$ and (15) yields

$$L_n^{\alpha,\beta} - \mathbf{E}(L_n^{\alpha,\beta}) = \sum_{n^\alpha \le k < n^\beta} \mathbf{E}(T_k)(Y_k - \mathbf{E}(Y_k)) + O_P(n^{-1/2}) \ .$$

Also $\mathbf{Var}(\frac{1}{n}\sum_{n^\alpha \le k < n^\beta} Y_k) \le n^{-2}\sum_{n^\alpha \le k < n^\beta} 2k/(n-k) = O(n^{-1})$, and because of (7) we end up with

$$L_n^{\alpha,\beta} - \mathbf{E}(L_n^{\alpha,\beta}) = 2 \sum_{n^\alpha \le k < n^\beta} \frac{Y_k - \mathbf{E}(Y_k)}{k} + O_P(n^{-1/2}) \ . \tag{16}$$

This is a representation of the external length by a sum of independent random variables.

Now $\mathbf{Var}(Y_k) = \frac{2k}{n-k} - \frac{4k^2}{(n-k)^2}$, thus for $\beta < 1$

$$\begin{aligned}
\mathbf{Var}\Big(2 \sum_{n^\alpha \le k < n^\beta} \frac{Y_k - \mathbf{E}(Y_k)}{k}\Big) &= 4 \sum_{n^\alpha \le k < n^\beta} \Big(\frac{2}{k(n-k)} - \frac{4}{(n-k)^2}\Big) \\
&\sim 8(\beta - \alpha)\frac{\log n}{n} \ .
\end{aligned}$$

Moreover for $\delta > 0$ we have $\mathbf{E}(|Y_k - \mathbf{E}(Y_k)|^{2+\delta}) \le \frac{2k}{n-k} + (\frac{2k}{n-k})^{2+\delta} \le \frac{4k}{n-k}$ for $k \le n/3$, thus

$$\sum_{n^\alpha \le k < n^\beta} \frac{1}{k^{2+\delta}}\mathbf{E}(|Y_k - \mathbf{E}(Y_k)|^{2+\delta}) \le 4 \sum_{n^\alpha \le k < n^\beta} \frac{1}{k^{1+\delta}(n-k)} \le \frac{8}{\delta n}\frac{1}{(n^\alpha - 1)^\delta} \ .$$

Thus for $\alpha \ge 1/2$ we get

$$\sum_{n^\alpha \le k < n^\beta} \frac{1}{k^{2+\delta}}\mathbf{E}(|Y_k - \mathbf{E}(Y_k)|^{2+\delta}) = o\Big(\frac{(\log n)^{1+\delta/2}}{n^{1+\delta/2}}\Big) \ ,$$

and we may use Lyapunov's criterion for the central limit theorem. Consequently, (16) implies

$$\frac{L_n^{\alpha,\beta} - \mathbf{E}(L_n^{\alpha,\beta})}{\sqrt{8(\beta-\alpha)\log n/n}} \xrightarrow{\mathrm{d}} N(0,1).$$

This finishes the proof in the case $\beta < 1$, using Proposition 3.

The case $\beta = 1$ then follows from $L_n^{\alpha,1} = L_n^{\alpha,1-\varepsilon} + L_n^{1-\varepsilon,1}$ using Proposition 3.

The last claim on asymptotic independence follows from (16), too. □

5 Proof of Proposition 5

Let $0 \leq \alpha \leq 1$ and $m = \lfloor n^\alpha \rfloor$. Since $k - V_k = \#\{i : \rho(i) < k\}$ is the number of external branches, which are found between level $k-1$ and k,

$$\hat{L}_n^{\alpha,1} = \sum_{m<k\leq n} (T_{k-1} - T_k)(k - V_k) \ .$$

From independence

$$\mathbf{E}(\hat{L}_n^{\alpha,1}) = \sum_{m<k\leq n} \frac{2}{k(k-1)} \cdot \frac{k(k-1)}{n-1} \ .$$

This gives the first claim. Next, letting

$$E_n := \mathbf{E}(L_n^{\alpha,1} \mid V_0, \ldots, V_n) = \sum_{m<k\leq n} \frac{k - V_k}{\binom{k}{2}} \ ,$$

we have

$$\mathbf{Var}(L_n^{\alpha,1}) = \mathbf{Var}(L_n^{\alpha,1} - E_n) + \mathbf{Var}(E_n) \ .$$

Now, using Proposition 6,

$$\begin{aligned}
\mathbf{Var}(L_n^{\alpha,1} - E_n) &= \sum_{m<k\leq n} \mathbf{E}\Big(\Big(T_{k-1} - T_k - \frac{1}{\binom{k}{2}}\Big)^2\Big)\mathbf{E}((k-V_k)^2) \\
&= \sum_{m<k\leq n} \frac{1}{\binom{k}{2}^2}\Big(\frac{k^2(k-1)^2}{(n-1)^2} + \frac{k(k-1)(n-k)(n-k-1)}{(n-1)^2(n-2)}\Big) \\
&= 4\frac{n-m}{(n-1)^2} + 4\sum_{m<k\leq n} \frac{(n-k)(n-k-1)}{k(k-1)(n-1)^2(n-2)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{Var}(E_n) = \sum_{m<k,l\leq n} \frac{1}{\binom{k}{2}\binom{l}{2}} \mathbf{Cov}(V_k, V_l) \\
&= 4\sum_{m<k\leq n} \frac{(n-k)(n-k-1)}{k(k-1)(n-1)^2(n-2)} + 8\sum_{m<k<l\leq n} \frac{(n-l)(n-l-1)}{l(l-1)(n-1)^2(n-2)} \\
&= 4\sum_{m<k\leq n} \frac{(n-k)(n-k-1)}{k(k-1)(n-1)^2(n-2)} + 8\sum_{m<l\leq n} \frac{(l-m-1)(n-l)(n-l-1)}{l(l-1)(n-1)^2(n-2)} \ .
\end{aligned}$$

Thus

$$\mathbf{Var}(L_n^{\alpha,1}) = 4\frac{n-m}{(n-1)^2} + 8\sum_{m<k\leq n} \frac{(k-m)(n-k)(n-k-1)}{k(k-1)(n-1)^2(n-2)} \ .$$

Now

$$\begin{aligned}
&(k-m)(n-k)(n-k-1)\\
&\qquad=\big(k-1-(m-1)\big)\big(k(k-1)-2(n-1)k+n(n-1)\big)\\
&\qquad=k(k-1)^2-(2n+m-3)k(k-1)\\
&\qquad\qquad+(n+2m-2)(n-1)k-mn(n-1),
\end{aligned}$$

thus

$$\begin{aligned}
&\tfrac12(n-m)(n-2)+\sum_{m<k\le n}\frac{(k-m)(n-k)(n-k-1)}{k(k-1)}\\
&\quad=\tfrac12(n-m)(n-2)+\tfrac12(n-m)(n+m-1)-(n-m)(2n+m-3)\\
&\qquad+(h_{n-1}-h_{m-1})(n+2m-2)(n-1)-\Big(\frac1m-\frac1n\Big)mn(n-1)\\
&\quad=(h_{n-1}-h_{m-1})(n+2m-2)(n-1)-\tfrac12(n-m)(4n+m-5)\ .
\end{aligned}$$

Combining our formulas the result follows. □

6 The length of a random external branch

Finally we look at the distribution of the length of an external branch chosen at random. Equivalently, letting $\rho := \rho(1)$, we may consider

$$R_n := T_\rho\ ,$$

the length of the branch ending in the leaf with label 1. Its asymptotic distribution can be obtained in an elementary manner and without recourse to the results of the preceding sections. Recall $\rho := \max\{k \ge 1 : \{1\} \notin \pi_k\}$, thus

$$\mathbf{P}(\rho<k)=\frac{\binom{n-1}{2}}{\binom{n}{2}}\cdots\frac{\binom{k}{2}}{\binom{k+1}{2}}=\frac{k(k-1)}{n(n-1)}\ .$$

Letting

$$R_n' := \sum_{k=\rho+1}^{n}\frac{1}{\binom{k}{2}}=2\Big(\frac1\rho-\frac1n\Big)\ ,$$

$\{R_n' > r\} = \{\rho < 2n/(nr+2)\}$ and for $x > 0$

$$\mathbf{P}(nR_n'>x)=\mathbf{P}(\rho<2n/(x+2))\sim\frac{4}{(x+2)^2}\ .$$

We show that this limiting result carries over to R_n. From

$$R_n-R_n'=\sum_{k=\rho+1}^{n}\Big(T_{k-1}-T_k-\frac{1}{\binom{k}{2}}\Big)=\sum_{k=2}^{n}\Big(T_{k-1}-T_k-\frac{1}{\binom{k}{2}}\Big)I_{\{\rho<k\}}$$

it follows that

$$\mathbf{P}\Big(R_n - R'_n \neq \sum_{\sqrt{n}<k\le n} \Big(T_{k-1} - T_k - \frac{1}{\binom{k}{2}}\Big) I_{\{\rho<k\}}\Big) \le \mathbf{P}(\rho < \sqrt{n}) = o(1).$$

Also from independence

$$\begin{aligned}\mathbf{E}\Big[\Big(\sum_{\sqrt{n}<k\le n} \Big(T_{k-1} - T_k - \frac{1}{\binom{k}{2}}\Big) I_{\{\rho<k\}}\Big)^2\Big] \\ &= \sum_{\sqrt{n}<k\le n} \mathbf{E}\Big[\Big(T_{k-1} - T_k - \frac{1}{\binom{k}{2}}\Big)^2\Big] \mathbf{P}(\rho < k) \\ &= \sum_{\sqrt{n}<k\le n} \frac{1}{\binom{k}{2}^2} \frac{\binom{k}{2}}{\binom{n}{2}} = o(n^{-2}) \ .\end{aligned}$$

Consequently $R_n = R'_n + o(n^{-1})$ in probability. Thus we end up with the following result, which was obtained by Caliebe et al (2007) by means of Laplace transform methods.

Proposition 10. *nR_n converges in distribution to the law μ on $\mathbb{R}^+$ with density $\mu(dx) = 8(x+2)^{-3}\,dx$.*

References

[1] Blum, M.G.B. and François, O. (2005) Minimal clade size and external branch length under the neutral coalescent. *Adv. Appl. Prob.* **37**, 647–662.

[2] Caliebe, A., Neininger, R., Krawczak, M. and Rösler, U. (2007) On the length distribution of external branches in coalescent trees: Genetic diversity within species. *Theor. Population Biology* **72**, 245–252.

[3] Durrett, R. (2002) Probability models for DNA sequence evolution. Probability and its Applications (New York). *Springer-Verlag, New York.*

[4] Freund F. and Möhle, M. (2009) On the time back to the most recent common ancestor and the external branch length of the Bothausen-Snitman coalescent. *Markov Proc. Rel. Fields* **15**, 387–416.

[5] Fu, Y.X and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

[6] Gnedin, A., Iksanov, A. and Möhle, M. (2008) On the asymptotics of exchangeable coalescents with multiple collisions. *J. Appl. Proobab.* **45**, 1186–1195.

[7] Jacod, J. and Shiryaev, A.N. (1987) *Limit Theorems for Stochastic Processes.* Springer-Verlag, Berlin.

[8] Janson, S. (2009) Sorting using complete subintervals and the maximum number of runs in a randomly evolving sequence. *Annals of Combinatorics* **12**, 417–447.

[9] Janson S. and Spencer J. (2007) A point process describing the component sizes in the critical window of the random graph evolution. *Combin. Probab. Comput.* **16**, 631–658.

[10] Kallenberg, O. (2002) *Foundations of Modern Probability.* 2nd ed., Springer, New York.

[11] Kingman, J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.* **13**, 235–248.

[12] Möhle, M. (2010) Asymptotic results for coalescent processes without proper frequencies and applications to the two-parameter Poisson-Dirichlet coalescent. *Stoch. Process. Appl.* **120**, 2159–2173.

[13] Steinsaltz, D. (1999) Random time changes for sock-sorting and other stochastic process limit theorems. *Electron. J. Probab.* **4**, no. 14, 25 pp.