Combinatorics, Probability and Computing (2008) 17, 727–740.
doi:10.1017/S0963548308009115 Printed in the United Kingdom

Coloured Loop-Erased Random Walk on the Complete Graph

JOMY ALAPPATTU[1] and JIM PITMAN[2†]

[1]Department of Mathematics, University of California at Berkeley,
970 Evans Hall, Berkeley, CA 94720-3840, USA
(e-mail: jomy.alappattu@gmail.com)

[2]Department of Statistics, University of California at Berkeley,
367 Evans Hall, Berkeley, CA 94720-3860, USA
(e-mail: pitman@stat.berkeley.edu)

Received 30 November 2006; revised 9 April 2008; first published online 8 September 2008

Starting from a sequence regarded as a walk through some set of values, we consider the associated loop-erased walk as a sequence of directed edges, with an edge from i to j if the loop-erased walk makes a step from i to j. We introduce a colouring of these edges by painting edges with a fixed colour as long as the walk does not loop back on itself, then switching to a new colour whenever a loop is erased, with each new colour distinct from all previous colours. The pattern of colours along the edges of the loop-erased walk then displays stretches of consecutive steps of the walk left untouched by the loop-erasure process. Assuming that the underlying sequence generating the loop-erased walk is a sequence of independent random variables, each uniform on $[N] := \{1, 2, \dots, N\}$, we condition the walk to start at N and stop the walk when it first reaches the subset $[k]$, for some $1 \leqslant k \leqslant N-1$. We relate the distribution of the random length of this loop-erased walk to the distribution of the length of the first loop of the walk, via Cayley's enumerations of trees, and via Wilson's algorithm. For fixed N and k, and $i = 1, 2, \dots$, let B_i denote the events that the loop-erased walk from N to $[k]$ has $i+1$ or more edges, and the ith and $(i+1)$th of these edges are coloured differently. We show that, given that the loop-erased random walk has j edges for some $1 \leqslant j \leqslant N-k$, the events B_i for $1 \leqslant i \leqslant j-1$ are independent, with the probability of B_i equal to $1/(k+i+1)$. This determines the distribution of the sequence of random lengths of differently coloured segments of the loop-erased walk, and yields asymptotic descriptions of these random lengths as $N \to \infty$.

1. Introduction

The *loop-erased walk* derived from a sequence $(X_n, n = 0, 1, \dots)$ is a sequence $(\mathcal{L}_n, n = 0, 1, \dots)$ of finite subsequences of $(X_n, n = 0, 1, \dots)$ defined as follows. Let $\mathcal{L}_0 = (Y_{0,0}) = (X_0)$, and inductively, if X_n is not in $\mathcal{L}_{n-1} = (Y_{n-1,0}, \dots, Y_{n-1,L_{n-1}})$, then form $\mathcal{L}_n$ by appending X_n to the

† Research supported in part by NSF grant DMS-0405779.

end of $\mathcal{L}_{n-1}$, *i.e.*, $\mathcal{L}_n = (Y_{n,0}, \dots, Y_{n,L_n})$ with $L_n = L_{n-1} + 1$, $Y_{n,i} = Y_{n-1,i}$ for $0 \leqslant i \leqslant L_{n-1}$, and $Y_{n,L_n} = X_n$. On the other hand, if $X_n = Y_{n-1,j}$ for some $1 \leqslant j \leqslant L_{n-1}$, then construct $\mathcal{L}_n$ by truncating the part of $\mathcal{L}_{n-1}$ beyond $Y_{n-1,j}$, *i.e.*, let $L_n = j$, and define $\mathcal{L}_n = (Y_{n,0}, \dots, Y_{n,L_n})$ by $Y_{n,i} = Y_{n-1,i}$ for $0 \leqslant i \leqslant j$; in this case, we say that a *loop* has occurred. For each $n = 0, 1, \dots$, we call L_n the *length of the loop-erased walk at time n*, with the understanding that if $\mathcal{L}_n$ is a single point (X_i) for some i, the length is zero. So the length measures the number of steps of the path, or the number of *edges* of the path, with each edge representing some one-step transition (X_n, X_{n+1}) of the original sequence $(X_n, n = 0, 1, \dots)$. See [6] and [7] for equivalent alternative definitions of loop-erased walk, and discussions of some basic results on the loop-erasure of random sequences.

There is a natural way to 'colour' the loop-erasure of the walk as follows. Assume that we have some infinite palette $\{\mathcal{C}_1, \mathcal{C}_2, \dots\}$ of colours. Run the walk, and until the first loop occurs, colour the edges of the walk with the colour $\mathcal{C}_1$. When the first loop occurs, erase the coloured edges as the definition of loop-erasure requires, and continue the walk, now colouring the subsequent edges with the colour $\mathcal{C}_2$ until the next loop occurs. More generally, keep colouring the edges of the walk with a fixed colour $\mathcal{C}_i$ until a loop occurs, at which point we change to a new fixed colour $\mathcal{C}_{i+1}$.

Fix a positive integer $N \geqslant 2$, and let $(X_n, n = 0, 1, \dots)$ be a sequence of independent random variables, with $X_0 = N$ for convenience, and $X_1, X_2, \dots$ independent and uniformly distributed on the set $[N] := \{1, 2, \dots, N\}$. We use this random sequence to construct a *loop-erased random walk* on $[N]$, following the definition for loop-erased walk above. Apart from some delay due to self-loops when $X_{n+1} = X_n$ for some n, the sequence of steps of the loop-erased walk is the same as if it were derived from a random walk on the complete graph. In particular, we are interested in the random colouring of the loop-erased walk when it first reaches the subset $[k]$ for some $1 \leqslant k \leqslant N - 1$, and this random colouring has the same distribution for a sequence of independent random variables $(X_n, n = 0, 1, \dots)$ as for a random walk on the complete graph. Note that we are not so interested in the particular sequence of colours being used, as in the relative lengths of the coloured segments.

Let R_N denote the *first repeat time* for $X_0, X_1, \dots$, *i.e.*, the first index i such that $X_i \in \{X_0, \dots, X_{i-1}\}$, which is the length of the first loop that is erased in the process of loop-erasure. The distribution of R_N is determined by the well-known solution of the classical birthday problem, that is,

$$\mathbb{P}(R_N > j) = \prod_{i=1}^{j} \frac{N - i}{N}, \tag{1.1}$$

$$\mathbb{P}(R_N = j) = \frac{j}{N} \prod_{i=1}^{j-1} \frac{N - i}{N}. \tag{1.2}$$

Our main result relates the distribution of R_N to the distribution of the length of the loop-erased walk stopped when it hits the target set $[k]$.

Let $\zeta_{N,k}$ be the first time i such that $X_i \in [k]$, and note that $\zeta_{N,k}$ is a geometric random variable with parameter k/N. We use the notation $X \overset{d}{=} Y$ to mean that the random variables X and Y have the same distribution, and $X \overset{d}{=} (Y \mid A)$ to indicate that the distribution of X is the same as the conditional distribution of Y given the event A. Also let

$$(a)_b := a(a-1)(a-2)\cdots(a-b+1)$$

be the usual falling factorial for $b = 1, 2, \dots$ with $(a)_0 = 1$.

Theorem 1.1. *Let $\lambda_{N,k}$ be the length of the loop-erased random walk derived from $X_0 = N$ and a sequence of independent variables $X_1, X_2, \dots$ with uniform distribution on $[N]$, stopped at time $\zeta_{N,k}$ when the sequence first hits $[k]$, and let R_N be the first repeat time derived from the same sequence. Then*

$$\lambda_{N,k} \stackrel{d}{=} (R_N - k | R_N > k);$$

that is, for every $1 \leqslant j \leqslant N-k$,

$$\mathbb{P}(\lambda_{N,k} = j) = \mathbb{P}(R_N - k = j | R_N > k) = \frac{(k+j)(N-k-1)_{j-1}}{N^j}. \tag{1.3}$$

Moreover, in the coloured loop-erased walk of length $\lambda_{N,k}$ obtained by stopping at time $\zeta_{N,k}$, let B_i denote the event that the ith and $(i+1)$th edges of the loop-erased walk are coloured differently. Then given $\lambda_{N,k} = j$, the events B_i for $1 \leqslant i \leqslant j-1$ are independent with

$$\mathbb{P}(B_i) = \frac{1}{k+i+1}.$$

We prove this result in Section 2, then show in Section 3 how the simple formula (1.3) is closely related both to Wilson's loop-erased random walk algorithm to generate spanning trees of a graph, and to Cayley's formula for the number of forests with a fixed number of vertices and a fixed set of roots. In Section 4, we go closer into the motivations behind Theorem 1.1, by showing how colouring loop-erased segments of a random walk is related to other known results about randomly breaking up discretized segments, which naturally induces random compositions on integers. Such random compositions are closely connected to results on Aldous's Brownian continuum random tree [1] and to stick-breaking schemes. In Section 5, we discuss some open questions which arise naturally from our analysis.

2. Proof of Theorem 1.1

Most of the work for this proof is done by the following lemma, where we use the notation of Theorem 1.1.

Lemma 2.1. *Let $v_{N,k}$ denote the number of different colours of segments of the coloured loop-erased walk started at $X_0 = N$ and stopped on first reaching $[k]$, let $\lambda_{N,k}$ be the total length of the loop-erased walk at this time, and given $\lambda_{N,k} = j$, let Y_i denote the indicator of the event B_i that the ith and $(i+1)$th edges in the loop-erased random walk from N to $[k]$ are coloured differently. Then, for each choice of positive integers s and j with $s \leqslant j$ and each choice of positive integers*

$i_1, \dots, i_{s-1}$ *with* $0 < i_1 < \cdots < i_{s-1} < j$,

$$\begin{aligned}\mathbb{P}(\lambda_{N,k} = j, \nu_{N,k} = s, Y_{i_1} = 1, Y_{i_2} = 1, \dots, Y_{i_{s-1}} = 1, Y_i = 0 \\ \textit{for } i \in [j-1] - \{i_1, \dots, i_{s-1}\}) \\ = \frac{(k+1)(N-k-1)_{j-1}}{N^j} \prod_{r=1}^{s-1} \frac{1}{k+i_r} \end{aligned} \tag{2.1}$$

where, if $s = 1$, *the product is understood to equal* 1.

Proof. Consider first the case when all edges are painted the same colour, *i.e.*, $\nu_{N,k} = s = 1$, and $Y_i = 0$ for all i. Then the last loop-erasure of the walk can only occur at vertex N, and thus the random walk takes some number n of steps to vertices in $[N] - [k]$, then it hits N one final time, and then it hits $(j-1)$ distinct vertices excluding N and $[k]$ before finally hitting some element of $[k]$. It thus follows that

$$\begin{aligned}\mathbb{P}(\lambda_{N,k} = j, \nu_{N,k} = 1) &= \left(1 + \sum_{n=0}^{\infty} \left(1 - \frac{k}{N}\right)^n \frac{1}{N}\right) \frac{(N-k-1)_{j-1}}{N^{j-1}} \frac{k}{N} \\ &= \frac{(k+1)(N-k-1)_{j-1}}{N^j},\end{aligned}$$

as desired.

Next, consider the case of two colours, with a colour change at the hth vertex in the path from N to $[k]$ for some $1 \leqslant h \leqslant j-1$. Again, the random walk takes n steps outside $[k]$ before hitting vertex N for the last time, and then the random walk takes h steps without looping, to distinct vertices. At the hth vertex, the random walk takes some number m of steps to vertices outside $[k]$ and the 0th, 1st, $\dots, (h-1)$th vertices before hitting the hth vertex one final time, and then it takes $j-h$ steps without looping until it hits something in $[k]$. So, again appealing to independence and the uniform distribution of the variables X_i for $i \geqslant 1$, we see that

$$\begin{aligned}&\mathbb{P}(\lambda_{N,k} = j, \nu_{N,k} = 2, Y_h = 1, Y_i = 0 \quad \text{for } i \in [j-1] - \{h\}) \\ &= \left(1 + \sum_{n=0}^{\infty} \left(1 - \frac{k}{N}\right)^n \frac{1}{N}\right) \left(\sum_{m=0}^{\infty} \left(1 - \frac{k+h}{N}\right)^m \frac{1}{N}\right) \frac{(N-k-1)_{j-1}}{N^{j-1}} \frac{k}{N} \\ &= \frac{(k+1)(N-k-1)_{j-1}}{N^j} \cdot \frac{1}{k+h},\end{aligned}$$

as desired.

Extending this argument to three or more colours is straightforward. □

Proof of formula (1.3). The equality of the second and third expressions is evident from (1.1) and (1.2). To prove the equality of the first and third expressions in this equation, we sum up equation (2.1) over all s between 1 and $N-k$ and all possible sequences $(i_1, \dots, i_{s-1})$, corresponding to all possible subsets of $[j-1]$, including the empty subset, and to all possible sequences of values of the colour change indicators (Y_i). Then (1.3) is seen to amount to

$$\sum_{I \subset [j-1]} \prod_{i \in I} \frac{1}{k+i} = \frac{k+j}{k+1}, \tag{2.2}$$

and this equality is clear from the fact that both sides are equal to

$$\prod_{i\in[j-1]}\left(1+\frac{1}{k+i}\right),$$

which finishes the proof of (1.3). □

For the second part of the theorem, we need a lemma about Bernoulli trials.

Lemma 2.2. *A sequence $Y_1,\dots,Y_{j-1}$ is an independent sequence of Bernoulli random variables, such that Y_i has parameter $\frac{1}{k+i+1}$ if and only if, for each choice of positive integers $i_1,\dots,i_{s-1}$ with $0<i_1<\cdots<i_{s-1}<j$,*

$$P(Y_{i_1}=\cdots=Y_{i_{s-1}}=1, Y_i=0 \quad \textit{for } \ i\in[j-1]-\{i_1,\dots,i_{s-1}\}) = \frac{k+1}{k+j}\prod_{r=1}^{s-1}\frac{1}{k+i_r}.$$

Proof. Suppose that the sequence $Y_1,\dots,Y_{j-1}$ is independent Bernoulli, such that Y_i has parameter $p_i := \frac{1}{k+i+1}$. Then

$$\begin{aligned} P(Y_{i_1}=\cdots=Y_{i_{s-1}}=1, Y_i=0 \quad &\text{for } \ i\in[j-1]-\{i_1,\dots,i_{s-1}\}) \\ &= \prod_{r=1}^{s-1}p_{i_r}\prod_{i\in[j-1]-\{i_1,\dots,i_{s-1}\}}(1-p_i). \end{aligned}$$

Using the fact that $p_i = p_{i-1}(1-p_i)$ for $i=1,\dots,j-1$ (where we let $p_0=\frac{1}{k+1}$) and the fact that $(1-p_1)\cdots(1-p_{j-1}) = \frac{k+1}{k+j}$, we see that this last product becomes

$$\prod_{r=1}^{s-1}p_{i_r-1}\prod_{i=1}^{j-1}(1-p_i) = \frac{k+1}{k+j}\prod_{r=1}^{s-1}p_{i_r-1} = \frac{k+1}{k+j}\prod_{r=1}^{s-1}\frac{1}{k+i_r},$$

as desired. The converse is obvious by just reversing the sequence of equalities. □

Proof of Theorem 1.1. Formula (1.3) has already been established. From Lemma 2.1 and (1.3) we see that

$$\begin{aligned} \mathbb{P}(v_{N,k}=s, Y_{i_1}=1, Y_{i_2}=1,\dots,Y_{i_{s-1}}=1, Y_i=0 \\ \text{for } \ i\in[j-1]-\{i_1,\dots,i_{s-1}\}|\lambda_N=j) \\ = \frac{k+1}{k+j}\prod_{r=1}^{s-1}\frac{1}{k+i_r} \end{aligned} \tag{2.3}$$

and now the conclusion follows from Lemma 2.2. □

3. The length of the loop-erased random walk

3.1. The Markov property of the length of the loop-erasure

An alternative method of proving formula (1.3) begins with the following observation. If L_n denotes the length of the loop-erasure of the i.i.d. sequence $(X_0,\dots,X_n)$ of random variables uniform on $[N]$, then $(L_n, n = 0, 1, \dots)$ has the same dynamics as a Markov chain with the following transition probabilities, started at $L_0 = 0$:

$$Q_N(i,j) = \begin{cases} 1/N & 0 \leqslant j \leqslant i, \\ (N-i-1)/N & j = i+1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

In fact, for $n = 1, 2, \dots, L_n \overset{d}{=} \min(L_{n-1} + 1, X_n - 1)$. Although by definition $L_0 = 0$ throughout this article, we could just as well start with some arbitrary loopless path (and a vertex at the end of the path from which to step) whose length L_0 is a random variable taking values in $\{0, 1, \dots, N-1\}$ and then run loop-erased random walk; if we then choose X_0 independent of $X_1, X_2, \dots$ so that $X_0 - 1 \overset{d}{=} L_0$, it follows by induction that

$$L_n \overset{d}{=} \min_{0 \leqslant j \leqslant n} (X_{n-j} + j - 1).$$

Using the independence of the $(X_i, i = 0, 1, \dots)$, it follows that

$$\begin{aligned} Q_N^n(i, [m, \infty)) &= \mathbb{P}(L_n \geqslant m | L_0 = i) \\ &= \mathbb{P}(\min_{0 \leqslant j \leqslant n} (X_{n-j} + j - 1) \geqslant m | X_0 = i) \\ &= 1(i + n - 1 \geqslant m) f_{N,m-n+1} f_{N,m-n+2} \cdots f_{N,m}, \end{aligned} \tag{3.2}$$

where for X a random variable uniformly distributed on $[N]$, $f_{N,m} := \mathbb{P}(X > m)$. Note that $f_{N,m} = 1$ if $m \leqslant 0$ and $f_{N,m} = 0$ if $m \geqslant N$. From this, we obtain arbitrary entries of powers of the transition matrix:

$$\begin{aligned} Q_N^n(i, m) &= Q_N^n(i, [m, \infty)) - Q_N^n(i, [m+1, \infty)) \\ &= f_{N,m-n+2} \cdots f_{N,m} (1(i + n - 1 \geqslant m) f_{N,m-n+1} \\ &\qquad - 1(i + n - 1 \geqslant m + 1) f_{N,m+1}). \end{aligned} \tag{3.3}$$

Moreover, letting $n \to \infty$ in (3.2) and using the fact that $X_1, X_2, \dots$ are non-negative random variables, it follows that

$$\lim_{n \to \infty} Q^n(i, [m, \infty)) = f_{N,1} \cdots f_{N,m},$$
$$\lim_{n \to \infty} Q^n(i, m) = f_{N,1} \cdots f_{N,m} (1 - f_{N,m+1}),$$

where, if $m \leqslant 0$, the product $f_{N,1} \cdots f_{N,m}$ is understood to be 1. Note that the first limit is the probability $\mathbb{P}(R_N > m)$, and the second limit is the probability $\mathbb{P}(R_N = m + 1)$. Thus, applying the convergence theorem for irreducible aperiodic Markov chains [3, p. 314], we obtain the following result.

Proposition 3.1. *The stationary distribution of the Markov chain $(L_n, n = 0, 1, \dots)$ is the distribution of the random variable $R_N - 1$, where R_N is the index of the first repeat in an i.i.d. sequence of random variables uniform on $[N]$.* □

As an aside, the exact same reasoning can be applied to a non-uniform random variable X on the positive integers, to obtain the following.

Corollary 3.2. *Let X be a positive-integer-valued random variable, and define an independent sequence $X_0, X_1, \dots$, where X_0 has some distribution on the positive integers, and $X_1, X_2, \dots$ is an i.i.d. sequence of variables with the same distribution as X. Define a transition matrix Q on the non-negative integers as follows:*

$$Q(i,j) = \begin{cases} \mathbb{P}(X = j+1) & 0 \leqslant j \leqslant i, \\ \mathbb{P}(X > i+1) & j = i+1. \end{cases}$$

Then if L_n is the loop-erasure of the path $(X_0, \dots, X_n)$, $(L_n, n = 0, 1, \dots)$ has the same transition dynamics as a Markov chain with transition matrix Q, and if $g_m := \mathbb{P}(X > m)$, then powers of the transition matrix are given by

$$(Q)^n(i,m) = g_{m-n+2} \cdots g_m (1(i+n-1 \geqslant m) g_{m-n+1} - 1(i+n-1 \geqslant m+1) g_{m+1}).$$

Moreover, if $\mathbb{P}(X = 1) > 0$, then the Markov chain is irreducible and positive recurrent on the non-negative integers, with a stationary distribution determined by either of the following formulas for $m = 1, 2, \dots$:

$$\pi([m,\infty)) = \prod_{i=1}^{m} \mathbb{P}(X > i),$$

$$\pi(m) = \mathbb{P}(X \leqslant m+1) \prod_{i=1}^{m} \mathbb{P}(X > i).$$

We now use equation (3.3) to provide a second proof of formula (1.3). Conditioning on the value of $\zeta_{N,k}$, which is a geometric random variable with parameter k/N, we see that for $1 \leqslant j \leqslant N-k$,

$$\begin{aligned} \mathbb{P}(\lambda_{N,k} = j) = \mathbb{P}(L_{\zeta_{N,k}} = j) &= \sum_{n=1}^{\infty} \mathbb{P}(L_{n-1} = j-1 | \zeta_{N,k} = n) \mathbb{P}(\zeta_{N,k} = n) \\ &= \sum_{n=1}^{\infty} Q_{N-k}^{n-1}(0, j-1) \left(1 - \frac{k}{N}\right)^{n-1} \frac{k}{N}. \end{aligned} \tag{3.4}$$

To calculate $Q_{N-k}^{n-1}(0, j-1)$, we use equation (3.3), to see that

$$Q_{N-k}^{n-1}(0, j-1) = \begin{cases} f_{N-k,1} \cdots f_{N-k,j-1}(1 - f_{N-k,j}) = \mathbb{P}(R_{N-k} = j) & \text{if } 0 \leqslant j \leqslant n-2, \\ f_{N-k,1} \cdots f_{N-k,j-1} = \mathbb{P}(R_{N-k} > j-1) & \text{if } j = n-1, \\ 0 & \text{if } j \geqslant n. \end{cases} \tag{3.5}$$

Now returning to equation (3.4), we see that

$$\mathbb{P}(\lambda_{N,k} = j) = \mathbb{P}(R_{N-k} > j-1)\left(1 - \frac{k}{N}\right)^{j-1} \frac{k}{N} + \mathbb{P}(R_{N-k} = j) \sum_{n=j+2}^{\infty} \left(1 - \frac{k}{N}\right)^{n-1} \frac{k}{N} \tag{3.6}$$

$$= \frac{(N-k-1)_{j-1}}{(N-k)^{j-1}} \cdot \frac{(N-k)^{j-1}}{N^{j-1}} \cdot \frac{k}{N} + \frac{j}{N-k} \cdot \frac{(N-k-1)_{j-1}}{(N-k)^{j-1}} \cdot \frac{(N-k)^j}{N^j} \tag{3.7}$$

$$= \frac{(j+k)(N-k-1)_{j-1}}{N^j}, \tag{3.8}$$

as desired.

3.2. Relation to using Wilson's algorithm

Yet another derivation of formula (1.3) is provided by Wilson's algorithm [10]; to explore this connection, we will need to introduce some preliminaries on trees. A *rooted tree* $T = (V, E, r)$ consists of a vertex set V, an edge set $E \subset V \times V$, and a distinguished vertex $r \in V$ called the *root*, such that for any non-root vertex $v \in V$, there is a unique directed sequence of edges that leads from v to r, and such that there are no undirected loops, *i.e.*, from any vertex there does not exist a sequence of distinct undirected edges which leads back to that vertex.

Wilson's algorithm can make use of the random walk on the complete graph to generate a random tree with N vertices labelled by $[N]$ as follows. Let T_0 be the one-point tree with root and vertex labelled 1. Suppose that $T_0, \dots, T_{n-1}$ have been defined, with respective vertex sets $V_0 = \{1\}, V_1, \dots, V_{n-1}$. If T_{n-1} has N vertices, stop the algorithm and output T_{n-1}. Otherwise, pick some vertex $v \in [N] - V_{n-1}$ (it does not matter how one chooses v), and run a random walk on the complete graph from v until it hits some vertex in V_{n-1}. Loop-erase this random walk, and add the loop-erased path to T_{n-1} to form the tree T_n, still with root labelled 1, and vertex set V_n. Also, call this loop-erased path from w to V_{n-1} a *macrostep* of the algorithm.

According to [10], the random tree generated by Wilson's algorithm applied to the complete graph with N vertices is uniformly distributed among all rooted trees labelled by $[N]$ with root 1; call this the *uniform spanning tree with root* 1 (where the word *spanning* is used to signify that the tree has the full set of N vertices). Suppose that we start Wilson's algorithm at a vertex N. The first macrostep is just the loop-erased walk from N to 1, and thus the length of this macrostep has the same distribution as $\lambda_{N,1}$. On the other hand, Wilson's algorithm implies that this macrostep is also the path from N to 1 in the uniform spanning tree with root 1. If $H_{N,1}$ is the length of this path, *i.e.*, the number of edges in the path, then Wilson's algorithm clearly implies that $H_{N,1} \overset{d}{=} \lambda_{N,1}$. Moreover, Meir and Moon [8] proved that $H_{N,1} \overset{d}{=} (R_N - 1 | R_N > 1)$, and thus Wilson's algorithm coupled with this result in [8] yields an alternative proof of formula (1.3) in the case $k = 1$. In fact, the methods of Wilson's algorithm and Meir and Moon can be applied to a random *rooted forest labelled by* $[N]$ (*i.e.*, a collection of trees with N total vertices labelled by $[N]$) with a fixed set of roots labelled by $[k]$, which is uniform among all such rooted forests labelled by $[N]$ with the same root set $[k]$, to prove formula (1.3) for arbitrary k.

The result of Meir and Moon made use of Cayley's formula for the enumeration of forests. Thus our derivation of formula (1.3) yields a novel proof of Cayley's formula.

Corollary 3.3 (Cayley's formula). *The number $t_{N,k}$ of forests with N vertices labelled by $[N]$ and k rooted trees with root set $[k]$ is given by $t_{N,k} = kN^{N-k-1}$.*

Proof. As mentioned above, applying Wilson's algorithm with root 1, started at some other vertex v, proves that $\lambda_{N,1}$ (the length of the loop-erased random walk from v to 1) and $H_{N,1}$ (the edge-distance between v and 1 in a spanning tree labelled by $[N]$ which is uniform among all spanning trees with root r) have the same distribution. For any fixed $1 \leqslant j \leqslant N - 1$, we want to count the number of spanning trees with root r such that the edge-distance from v to r equals j. We have to choose the $j - 1$ vertices in the path from v to r, and then each vertex on the path (including v and r) may be considered the root of a tree; thus, there are $(N-2)_{j-1} t_{N,j+1}$ such spanning trees. Thus, it follows that

$$\frac{(j+1)(N-2)_{j-1}}{N^j} = \mathbb{P}(\lambda_{N,1} = j) = \mathbb{P}(H_{N,1} = j) = \frac{(N-2)_{j-1} t_{N,j+1}}{t_{N,1}}. \tag{3.9}$$

In the particular case $j = N-1$, it is obvious that $t_{N,j+1} = 1$, so this equality implies that $t_{N,1} = N^{N-2}$. But now substituting this into equation (3.9) yields Cayley's formula for arbitrary j. $\square$

Lyons and Peres [7] use very similar reasoning in applying Wilson's algorithm to calculate $t_{N,1}$, by computing the probability of a particular tree, namely a path of length $N-1$.

4. Some scaling limit results

4.1. Connections to Poisson and Rayleigh processes

See Pittel's paper [9] for a discussion of related results where a similar distribution involving the lengths of gaps between 1s of an independent Bernoulli sequence arises. Pittel shows that the sequence of lengths of macrosteps obtained when applying Wilson's algorithm to the complete graph with N vertices has the same distribution as the sequence of spacings between successes of independent Bernoulli(j/N) variables, $2 \leqslant j \leqslant N$. For each $N = 1, 2, \dots$, let $(Y_{N,j}, j = 1, 2, \dots)$ be such an independent sequence of Bernoulli random variables, where for $1 \leqslant j \leqslant n$, $Y_{N,j}$ has parameter j/N, and for $j > N$, $Y_{N,j} = 1$. Let $Z_{N,1}$ be the smallest index j such that $Y_{N,j} = 1$, let $Z_{N,2}$ be the second-smallest index j such that $Y_{N,j} = 1$, and define $Z_{N,i}$ similarly for larger i. Then for any positive integer m, as $N \to \infty$,

$$\frac{1}{\sqrt{N}}(Z_{N,1}, \dots, Z_{N,m}) \xrightarrow{d} (P_1, \dots, P_m),$$

where $P_1 < P_2 < \cdots$ are the successive points of an inhomogeneous Poisson point process on $[0, \infty)$ with rate t at time t; that is, at every continuity point of the distribution function of $(P_1, \dots, P_m)$, the distribution function of $(Z_{N,1}, \dots, Z_{N,m})$ converges to the distribution function of $(P_1, \dots, P_m)$. This follows from, *e.g.*, results of [3, §2.6].

A similar type of scaling limit comes up when analysing repeat values. Let $R_{N,1} := R_N$ be the index of first repeat in an i.i.d. sequence of random variables uniform on $[N]$, and let $R_{N,2}$ be the second-smallest index i such that $X_i \in \{X_0, \dots, X_{i-1}\}$, let $R_{N,3}$ be the third-smallest such index, and so on. Then (see [2] and work cited there) the sequence $(R_{N,i}, i = 1, 2, \dots)$ has the same finite-dimensional scaling limits as the sequence $(Z_{N,j}, j = 1, 2, \dots)$: for any positive integer m, as $N \to \infty$,

$$\frac{1}{\sqrt{N}}(R_{N,1}, \dots, R_{N,m}) \xrightarrow{d} (P_1, \dots, P_m).$$

This latter scaling limit result is an integral part of one of Aldous's constructions of the Brownian continuum random tree [1].

But now, using the fact that $\lambda_{N,k} \overset{d}{=} (R_N - k | R_N > k)$ from Theorem 1.1, we obtain the following.

Corollary 4.1. *For a fixed $\mu > 0$, as $N \to \infty$, $(\lambda_{N,\lfloor \mu\sqrt{N}\rfloor})/\sqrt{N}$ converges in distribution to $(P_1 - \mu | P_1 > \mu)$, where P_1 is the first point of an inhomogeneous Poisson point process on $[0,\infty)$ with rate t at time t.*

This result can also be used to provide an estimate for $v_{N,k}$, the number of colours in the loop-erased random walk from N to $[k]$. It is clear that conditional on $\lambda_{N,k} = j$, the expected value of $v_{N,k}$ is given by

$$E(v_{N,k} | \lambda_{N,k} = j) = 1 + \sum_{i=1}^{j-1} \frac{1}{k+i+1}.$$

It thus follows that

$$E(v_{N,\lfloor \mu\sqrt{N}\rfloor}) = \sum_{j=1}^{N-\lfloor \mu\sqrt{N}\rfloor} \left(1 + \frac{1}{\lfloor \mu\sqrt{N}\rfloor + 2} + \cdots + \frac{1}{\lfloor \mu\sqrt{N}\rfloor + j}\right) \mathbb{P}(\lambda_{N,\lfloor \mu\sqrt{N}\rfloor} = j).$$

But as $N \to \infty$, the term $1 + \frac{1}{\lfloor \mu\sqrt{N}\rfloor+2} + \cdots + \frac{1}{\lfloor \mu\sqrt{N}\rfloor+j}$ approaches

$$1 + \log\left(1 + \frac{j}{\mu\sqrt{N}}\right),$$

and by Corollary 4.1, the term $\mathbb{P}(\lambda_{N,\lfloor \mu\sqrt{N}\rfloor} = j)$ approaches

$$\frac{1}{\sqrt{N}}\left(\frac{j}{\sqrt{N}} + \mu\right)\exp\left(-j^2/2N - \mu j/\sqrt{N}\right).$$

Therefore, as $N \to \infty$,

$$E(v_{N,\lfloor \mu\sqrt{N}\rfloor}) \to \int_0^\infty \left(1 + \log\left(1 + \frac{x}{\mu}\right)\right)(\mu + x)\exp\left(-\frac{x^2}{2} - \mu x\right)dx \tag{4.1}$$

$$= 1 - \log\mu + \exp\left(\frac{\mu^2}{2}\right)\int_\mu^\infty t\log t \exp\left(-\frac{t^2}{2}\right)dt \tag{4.2}$$

after some simplification and the substitution $t = x + \mu$.

An alternative derivation of this scaling limit comes from the fact that the length of the loop-erased random walk increases at unit speed until a length j when a loop occurs, after which its new length is uniformly distributed among $\{0, 1, \ldots, j\}$. As such, it is closely related to the *standard Rayleigh process* $(R_t, t \geqslant 0)$ [4], defined as follows: let $R_0 = 0$, and for $P_{i-1} < t < P_i$ (with the convention that $P_0 = 0$), let R_t grow at unit speed; at each time P_i, let R_{P_i} be selected uniformly within the interval $(0, R_{P_i -})$. If we also make note of the basic fact that $\zeta_{N,k}$, the geometric time at which the walk from N hits $[k]$, satisfies the scaling limit $\zeta_{N,\lfloor \mu\sqrt{N}\rfloor}/\sqrt{N} \xrightarrow{d} X_\mu$, where X_μ has an exponential distribution with parameter μ, then the following is clear.

Corollary 4.2. *As $N \to \infty$, $\lambda_{N,\lfloor \mu\sqrt{N}\rfloor}/\sqrt{N}$ converges in distribution to R_{X_μ}, where $(R_t, t \geqslant 0)$ is the standard Rayleigh process, and X_μ is independent of $(R_t, t \geqslant 0)$ and has an exponential distribution with parameter μ.*

Using the fact from [4] that $R_t \stackrel{d}{=} P_1 \wedge t$, where P_1 has the standard Rayleigh distribution with $\mathbb{P}(P_1 > \lambda) = e^{-\lambda^2/2}$, shows that the scaling limits in Corollaries 4.1 and 4.2 have the same distribution.

Consider a finite k, condition on the length of the loop-erased random walk from N to $[k]$ equalling j, and let $C_1 < \cdots < C_{v_{N,k}-1}$ be the colour-changing indices, *i.e.*, those indices i in $[j-1]$ such that $B_i = 1$. By Theorem 1.1, it follows that

$$\mathbb{P}(C_1 > i) = \left(1 - \frac{1}{k+2}\right)\left(1 - \frac{1}{k+3}\right)\cdots\left(1 - \frac{1}{k+i+1}\right) = \frac{k+1}{k+i+1},$$

$$\mathbb{P}(C_m > i+j | C_{m-1} = i) = \frac{k+i+1}{k+i+j+1}.$$

If we now let $k = \lfloor \mu\sqrt{N} \rfloor$, and let $N \to \infty$, then it follows that $C_1/\sqrt{N} \xrightarrow{d} D_1$, where $P(D_1 > \lambda) = \frac{\mu}{\mu+\lambda}$. More generally, we see the following.

Proposition 4.3. *As $N \to \infty$,*

$$\frac{1}{\sqrt{N}}(C_1, \ldots, C_{v_{N,\lfloor \mu\sqrt{N}\rfloor}-1}) \xrightarrow{d} (D_1, \ldots, D_{v(\mu)}),$$

where $D_1 < D_2 < \cdots$ are the points of an inhomogeneous Poisson point process with rate $\frac{1}{\mu+t}$ at time t, and $v(\mu) := \sup\{i : D_i < R_{X_\mu}\}$. □

Sequence $(D_1, \ldots, D_{v(\mu)})$ is called the sequence of *ladder indices* of the standard Rayleigh process up to time X_μ: if, at a jump time P_i, the Rayleigh process jumps down to some Q_i uniformly chosen in $(0, R_{P_i-})$, then $(D_1, \ldots, D_{v(\mu)})$ is the subsequence $(Q_{i_1}, \ldots, Q_{i_{v(\mu)}})$ of $(Q_1, Q_2, \ldots)$ up to time X_μ such that, for each index i_c, $Q_{i_c} < Q_{i_c+m}$ for all $m = 1, 2, \ldots$.

The close connection between loop-erased random walks and Rayleigh processes is also highlighted in recent work of Schweinsberg [11], in which it is shown that the loop-erased random walk on a wide variety of sequences of graphs converges to the standard Rayleigh process after suitable normalization.

As a check, note that conditional on $R_{X_\mu} = x$, the number of colours in the scaled walk is one plus a Poisson random variable with parameter $\log(1 + x/\mu)$, and thus the expected value of this number of colours is

$$1 + \int_0^\infty \log\left(1 + \frac{x}{\mu}\right)(x+\mu)\exp\left(-\frac{x^2}{2} - \mu x\right)dx,$$

which is the same as derived in equation (4.1).

4.2. Stick-breaking

In the previous subsection, letting $k = \lfloor \mu\sqrt{N} \rfloor$ led to scaling limit results which were closely related to particular Poisson and Rayleigh processes on $[0, \infty)$. Fixing k while letting $N \to \infty$ leads to a very different result. To explore this, first note that Theorem 1.1 also relates to some other basic results on random compositions. See [5, §I.5] and [3, pp. 52–53] for discussions of how the record times of a sequence of i.i.d. random variables $(W_i, i = 1, 2, \ldots)$, *i.e.*, the times j

such that $W_j > W_i$ for all $1 \leqslant i \leqslant j-1$, is distributed like the occurrences of 1s in a sequence of independent Bernoulli($\frac{1}{i+1}$) random variables.

In the setting of Theorem 1.1, let $(\sigma_{N,k,1}, \dots, \sigma_{N,k,\nu_{N,k}})$ denote the random composition of $\lambda_{N,k}$ representing the numbers of edges of each colour, working along the coloured loop-erased path from N to $[k]$, and consider the reversed sequence of segment lengths

$$(\sigma'_{N,k,1}, \dots, \sigma'_{N,k,\nu_{N,k}}) := (\sigma_{N,k,\nu_{N,k}}, \dots, \sigma_{N,k,1}). \tag{4.3}$$

Also, suppose that we condition on $\lambda_{N,k} = j$ for some $1 \leqslant j \leqslant N-k$, so that if C_i for $1 \leqslant i \leqslant j-1$ is the indicator of whether i is a partial sum of the sequence $(\sigma'_{N,k,m}, 1 \leqslant m \leqslant \nu_{N,k})$, then the variables $(C_i, 1 \leqslant i \leqslant j-1)$, conditioned on $\lambda_{N,k} = j$, are independent Bernoulli($\frac{1}{k+j+1-i}$) random variables. We know from the proof of Theorem 1.1 that $\mathbb{P}(\sigma'_{N,k,1} = j) = \frac{k+1}{k+j}$ (since this corresponds to there being only one colour in the loop-erased walk); however, because of Theorem 1.1, we now also see that, for $1 \leqslant n \leqslant j-1$,

$$\begin{aligned}&\mathbb{P}(\sigma'_{N,k,1} = n)\\ &\quad = \left(1 - \frac{1}{k+j}\right)\left(1 - \frac{1}{k+j-1}\right)\cdots\left(1 - \frac{1}{k+j+2-n}\right)\frac{1}{k+j+1-n} = \frac{1}{k+j},\end{aligned}$$

so that $\sigma'_{N,k,1}$ is equally likely to be any of $1, 2, \dots, j-1$.

It is easy to generalize this, to see that conditional on $(\sigma'_{N,k,1}, \dots, \sigma'_{N,k,m}) = (k_1, \dots, k_m)$ with $c := k_1 + \cdots + k_m < j$,

$$\mathbb{P}(\sigma'_{N,k,m+1} = j - c) = \frac{k+1}{k+j-c}, \tag{4.4}$$

$$\mathbb{P}(\sigma'_{N,k,m+1} = k_{m+1}) = \frac{1}{k+j-c}, \quad 1 \leqslant k_{m+1} \leqslant j-c-1. \tag{4.5}$$

Now since $\lambda_{N,k} \to \infty$ in probability as $N \to \infty$, we see that as $N \to \infty$,

$$\lambda_{N,k}^{-1}(\sigma'_{N,k,1}, \dots, \sigma'_{N,k,\nu_{N,k}}) \xrightarrow{d} (U_1, (1-U_1)U_2, (1-U_1)(1-U_2)U_3, \dots) \tag{4.6}$$

where the $(U_i, i = 1, 2, \dots)$ are independent uniform $(0,1)$ random variables. The right-hand side of (4.6) is known as the *continuous uniform stick-breaking process* defined by the $(U_i, i = 1, 2, \dots)$.

4.3. Markovian and non-Markovian properties of the coloured walk

Although the sequence of lengths L_n of the loop-erased walk is a Markov chain, we will now argue that the sequence of compositions of L_n defined by the lengths of coloured segments of the loop-erased walk is not a Markov chain.

Recall that a *composition* of a positive integer ℓ is a sequence of positive integers with sum ℓ. Here the terms of the sequence represent the lengths of stretches of edges of the same colour in a loop-erased walk with total length ℓ. Such a composition of ℓ is conveniently encoded by the string of ℓ binary *bits* $(y_1, \dots, y_\ell)$, where $y_1 = 1$ and y_i is the indicator of a colour change between the $(i-1)$th and ith edges. So the number of compositions of ℓ is $2^{\ell-1}$. If $y = (y_1, \dots, y_\ell)$ and $z = (z_1, \dots, z_m)$ are compositions of ℓ and m respectively, call y a *truncation* of z if $\ell \leqslant m$, $y_i = z_i$ for $1 \leqslant i \leqslant \ell - 1$, and $y_\ell \leqslant z_\ell$. We also introduce a *trivial composition*, corresponding to a sequence with no terms, which is regarded as a truncation of every composition of a positive integer.

For $n = 0, 1, \ldots$, let C_n denote the composition induced by colouring the segments of the loop-erasure of $(X_0, \ldots, X_n)$, where it is understood that if the loop-erasure has one vertex and no edges, then C_n is the trivial composition. Observe that X_{n+1} belongs to the set of values in the loop-erasure of $(X_0, \ldots, X_n)$ if and only if C_{n+1} is a truncation of C_n. The sequence $(C_n, n = 0, 1, \ldots)$ has the following dynamics.

- If C_n is the trivial composition, then C_{n+1} is the trivial composition with probability $1/N$, and $C_{n+1} = (1)$ with probability $1 - 1/N$.
- If C_n is some non-trivial composition c of ℓ, then C_n is either some truncation of C_{n-1} or an extension of C_{n-1} by one term, according to whether or not X_n creates a loop:
  - if X_n creates a loop, then C_{n+1} extends C_n by adding bit 1 with probability $(N - \ell - 1)/N$, while C_{n+1} is equally likely to be each of the $\ell + 1$ possible truncations of C_n;
  - if X_n does not create a loop, then C_{n+1} extends C_n by adding bit 0 with probability $(N - \ell - 1)/N$, while C_{n+1} is equally likely to be each of the $\ell + 1$ possible truncations of C_n.

The sequence $(C_n, n = 0, 1, \ldots)$ is not a Markov chain because the transition dynamics from C_n to C_{n+1} depend on whether X_n created a loop, which is determined by the relationship between C_{n-1} and C_n. However, this analysis does show that the sequence of pairs of compositions, $((C_n, C_{n+1}), n = 0, 1, \ldots)$, is a Markov chain.

5. Open questions

This article has examined loop-erased random walk which stops when we reach some marked subset of $[N]$ of size k, at which point the walk has a random length $\lambda_{N,k}$; in this case, we showed that, conditional on the value of $\lambda_{N,k}$, the composition of colour segments was distributed as the length of spacings between 1s of a sequence of Bernoulli random variables. What if we stop the walk when the length of the loop-erasure reaches m, for some $1 \leqslant m \leqslant N - k$? Will the resulting composition of colours have a similar distribution, as the lengths of spacings? How will the composition be distributed if we just stop at some fixed finite time m, and look at the coloured loop-erasure of $(X_0, \ldots, X_m)$? The situation appears to be like the Ray–Knight description of the distribution of Brownian local times, where results are much simpler for suitable stopping times than for fixed times.

In Section 4, we derived a number of results which hold at the end of the stopped walk. But in this situation too, what happens at an intermediate stage? In the case of $k = \lfloor \mu\sqrt{N} \rfloor$, it seems that the length of the loop-erased random walk, considered as a stochastic process, should converge to the standard Rayleigh process.

Finally, it seems that there is some sort of 'critical' behaviour when k is of order of magnitude $\sqrt{N}$, as detailed in Section 4. What happens to the coloured walk when $k = \omega(\sqrt{N})$ or $k = o(\sqrt{N})$? In the former case, it is not too hard to see that the probability of only one colour occurring in the walk goes to 1: if R_N is the first repeat time, and $\zeta_{N,k}$ is the first time that simple random walk on $[N]$ started at N hits $[k]$, then it is clear that $R_N/\sqrt{N} \xrightarrow{d} R$ as $N \to \infty$, where R is the standard Rayleigh distribution, and that if $k = \omega(\sqrt{N})$, then $\zeta_{N,k}/\sqrt{N} \xrightarrow{d} 0$, so that $(R_N - \zeta_{N,k})/\sqrt{N} \xrightarrow{d} R$, and thus $\mathbb{P}(R_N - \zeta_{N,k} > 0) \to 1$, which implies that with probability approaching 1, $[k]$ will be hit before the walk has a loop, and thus only one colour will be used. In the latter case of $k = o(\sqrt{N})$, the walk does not get stopped by an exponential time soon enough, and it appears that the number of colours of the stopped walk should increase as $\log N$.

References

[1] Aldous, D. J. (1991) The continuum random tree I. *Ann. Probab.* **19** 1–28.

[2] Camarri, M. and Pitman, J. (2000) Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electron. J. Probab.* **5** 1–18.

[3] Durrett, R. (1996) *Probability: Theory and Examples*, Duxbury.

[4] Evans, S., Pitman, J. and Winter, A. (2006) Rayleigh processes, real trees, and root growth with re-grafting. *Probab. Theory Rel. Fields* **134** 81–126.

[5] Feller, W. (1966) *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley.

[6] Lawler, G. F. (1991) *Intersections of Random Walks*, Birkhäuser.

[7] Lyons, R. and Peres, Y. (2003) *Probability on Trees and Networks*, Cambridge University Press. In preparation. Current version available at `http://php.indiana.edu/~rdlyons/`.

[8] Meir, A. and Moon, J. W. (1970) The distance between points in random trees. *J. Combin. Theory* **8** 99–103.

[9] Pittel, B. Note on exact and asymptotic distributions of the parameters of the loop-erased random walk on the complete graph. In *Mathematics and Computer Science II* (Versailles 2002), *Trends in Mathematics*, Birkhäuser.

[10] Wilson, D. B. (1996) Generating random spanning trees more quickly than the cover time. In *Proc. 28th Annual ACM Symposium on the Theory of Copmputing*, ACM, pp. 296–303.

[11] Schweinsberg, J. (2008) Loop-erased random walk on finite graphs and the Rayleigh process. *J. Theoret. Probab.* **21** 378–396.