Electron. Commun. Probab. **26** (2021), article no. 66, 1–15.
https://doi.org/10.1214/21-ECP436
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

A time-invariant random graph with splitting events*

Agelos Georgakopoulos† John Haslegrave‡

Abstract

We introduce a process where a connected rooted multigraph evolves by splitting events on its vertices, occurring randomly in continuous time. When a vertex splits, its incoming edges are randomly assigned between its offspring and a Poisson random number of edges are added between them. The process is parametrised by a positive real λ which governs the limiting average degree. We show that for each value of λ there is a unique random connected rooted multigraph $M(\lambda)$ invariant under this evolution. As a consequence, starting from any finite graph G the process will almost surely converge in distribution to $M(\lambda)$, which does not depend on G. We show that this limit has finite expected size. The same process naturally extends to one in which connectedness is not necessarily preserved, and we give a sharp threshold for connectedness of this version.

This is an asynchronous version, which is more realistic from the real-world network point of view, of a process we studied in [8, 9].

Keywords: random graphs; reproducing graphs; convergence; birth process.
MSC2020 subject classifications: Primary 05C82, Secondary 05C80; 60C05; 90B15.
Submitted to ECP on December 4, 2019, final version accepted on October 29, 2021.
Supersedes arXiv:1911.09630v2.

1 Introduction

We consider a random network model with reproduction which evolves in continuous time. Each vertex independently, at rate 1, splits into two. When a vertex splits, each of its existing edges is randomly rerouted to one of the two vertices produced, and these two vertices are connected by a random number of edges with distribution $\mathrm{Po}(\lambda/2)$, where $\lambda > 0$ is a fixed parameter. If the resulting graph is disconnected, only the component of the root is retained (the precise definition is given in the next section). We show that there is a unique random multigraph $M(\lambda)$ which is time-invariant under this evolution and has finite average degree (Theorem 1.4), and analyse some of its properties. As a consequence, if we run our process starting from any finite graph G, it will almost surely converge in distribution to $M(\lambda)$.

This model arose naturally in our recent work [9]: there, we considered the variant of the above evolution where all vertices split simultaneously in regular time intervals.

*Both authors were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 639046). J.H. was also partially supported by the UK Research and Innovation Future Leaders Fellowship MR/S016325/1.
†University of Warwick.
‡University of Warwick. E-mail: j.haslegrave@cantab.net

We observed that there is a unique finite-degree random multigraph $G(\lambda)$ which is time-invariant under this evolution too. We will refer to $G(\lambda)$ as the *synchronous version* of $M(\lambda)$. Moreover, we showed that $G(\lambda)$ is identically distributed with the cluster of the origin in an instance of long-range percolation on the infinitely-generated group $\bigoplus_{i\in\mathbb{N}} \mathbb{Z}_2$. Perhaps surprisingly, given its alternative definition as a cluster of a percolation model on a group, and given that most percolation models on finitely generated groups undergo a phase transition [6], $G(\lambda)$ is almost surely finite for any value of the intensity λ, and its expected size is finite. In this paper we show the analogous result for $M(\lambda)$ (Theorem 1.5).

Our splits can be thought of as reproduction of vertices, in the sense that a vertex produces a child and then passes on some of its connections to its child. In this sense, our first definition of $G(\lambda)$ is reminiscent of the models for random reproducing graphs studied by Jordan [12], building on earlier deterministic models for social networks [15, 5], with the key distinction being that in Jordan's model all connections of the parent are retained, whether or not they are inherited by the child.

However, simultaneous, discrete-time reproduction by the whole population is not a realistic model for real-life networks. It is therefore natural to consider a variant in which reproduction events are independent and may occur at any time, which is part of the motivation of the current paper. Mechanisms for growing networks based on repeated vertex duplications have previously been proposed as plausible for the development of the web graph [13] and for evolution of biochemical networks [3, 17]. Mathematical analysis of such a model was carried out non-rigorously by Pastor-Satorras, Smith and Sole [14], suggesting a limiting degree distribution which is power-law with an exponential cutoff, although subsequent rigorous work by Bebek, Berenbrink, Cooper, Friedetzky, Nadeau, and Sahinalp [2] showed that this is not the case. Another related model, motivated by duplication of genetic material, has been studied by Thörnblad [16] and by Backhausz and Móri [1]; however, the graph structure of this model is particularly simple, being a collection of disjoint cliques. A similar model for a fixed population size, which has richer behaviour owing to the random loss of individual edges, was introduced by Bienvenu, Débarre, and Lambert [4].

Although the continuous-time model $M(\lambda)$ studied here is more natural in certain respects, its analysis is significantly more challenging than that of the synchronous version $G(\lambda)$ for the following reason. A basic tool in the analysis of both models is the underlying *genealogical tree* T, containing all vertices in our evolution, and joining each vertex to its children by an edge. Starting with T, we can alternatively define our random graphs by joining pairs of leaves of T with random independent edges with appropriately chosen probabilities. In the synchronous case, this T is very simple: it is a binary tree of depth n when we run the process for n steps starting from a single vertex, and it is the so-called canopy tree when we start with $G(\lambda)$. When we start with $M(\lambda)$ however, T is a random tree with a non-trivial distribution: it can be thought of as the local limit of the ball $B(t)$ of radius t in first passage percolation on the full binary tree after re-rooting $B(t)$ at a leaf (see Section 2 for more details). Thus our main results Theorem 1.4 and Theorem 1.5 below were much harder to prove than their analogues in [9].

1.1 Model and results

It will be convenient for some proofs and statements of results to define both the main process defined above and a "full" version of the process in which other components are not discarded. In fact it is simpler to define the latter first. A *multigraph* is a graph in which two vertices may be joined by several parallel edges. The multigraphs of this paper do not have loops, i.e. edges that start and end at the same vertex.

Definition 1.1. *For a rooted connected multigraph, (G, o), the* full process $(\overline{G}_t, o_t)_{t\geq 0}$ *with parameter $\lambda > 0$ is defined as follows. Set $(\overline{G}_0, o_0) = (G, o)$. Give each vertex v a splitting time τ_v, where splitting times are i.i.d.* $\mathrm{Exp}(1)$ *variables. When $t = \tau_v$, replace v with two new vertices v_1, v_2, and give each a splitting time of $t + \mathrm{Exp}(1)$. Add* $\mathrm{Po}(\lambda/2)$ *edges between v_1 and v_2. Moreover, replace each edge of the form uv with one of the edges uv_1, uv_2 chosen uniformly at random. If v was the root, update the root to be v_1 or v_2, each with probability $1/2$. All these random choices are made independently from each other. Set $(\overline{G}_t, o_t)$ to be the resultant graph.*

We will frequently consider a single-vertex starting graph; we write $\overline{G}_t^\circ$ in this case.

Remark 1.2. The number of vertices of $\overline{G}_t^\circ$ over time, which is independent of all edge-related events, is a Yule process with rate $r = 1$, that is, a pure birth process where the birth rate is r times the population. Its value at time t has a geometric distribution with mean e^{rt}; see [7, Section XVII.3].

Definition 1.3. *The* cluster process (G_t, o_t) *with parameter λ is the rooted connected multigraph formed by the component of the root in $\overline{G}_t$.*

It is natural to think of the cluster process as a reproduction process where individuals die when they leave the component of the root. In this sense it resembles a general branching process, or Crump–Mode–Jagers process, (see e.g. [11, Chapter 6]); however, these processes assume independence of the lifespans of different individuals, whereas in our model death events are highly interdependent.

We prove three main results about these processes, listed below.

Theorem 1.4. *For each $\lambda > 0$ there is a unique random rooted connected multigraph with finite expected root degree, $(M(\lambda), o)$, which is invariant under the cluster process in the sense that $(M(\lambda)_t, o_t)$ has the same distribution for any $t \geq 0$.*

It is not immediately obvious that $M(\lambda)$ is almost surely finite. However, we prove a much stronger result.

Theorem 1.5. $\mathbb{E}(|M(\lambda)|) < \infty$ *for every $\lambda > 0$.*

When considering the full process, a natural question is when it becomes disconnected, or equivalently when the full and cluster processes first differ.

Theorem 1.6. *The time $t = \lambda$ is a sharp threshold for both connectedness of $\overline{G}_t^\circ$ and the existence of isolated vertices, that is, for any $\varepsilon > 0$, with high probability as $\lambda \to \infty$ the graph $\overline{G}_{(1-\varepsilon)\lambda}^\circ$ is connected but $\overline{G}_{(1+\varepsilon)\lambda}^\circ$ is disconnected with isolated vertices.*

1.2 Questions

In [9] we conjectured that $\mathbb{E}(|G(\lambda)|) \sim \lambda^{c\lambda}$ in agreement with computer simulation data. Simulations on $\mathbb{E}(|M(\lambda)|)$ showed a similar behaviour to $\mathbb{E}(|G(\lambda)|)$, and the same conjecture can be made. We know that $\mathbb{E}(|G(\lambda)|)$ is an analytic function of λ because of results in percolation theory [10]. For $\mathbb{E}(|M(\lambda)|)$ we do not even have a proof of continuity. Apart from obtaining more detailed results about the behaviour of $M(\lambda)$, it would also be interesting to modify our splitting rule in order to obtain other random graph models with temporal invariance.

2 Convergence to a limit

In this section we prove Theorem 1.4; throughout the section we assume the parameter $\lambda > 0$ is fixed. Let (G, o) be a random rooted graph such that $\mathbb{E}(d(o))$ is finite. Let (G°, o) be the single-vertex loopless graph with the same root o. Run the cluster process (G_t, o_t) given in Definition 1.3, and let H_t be the subgraph of G_t induced by descendants of o. Note that $o_t \in H_t$ and (H_t, o_t) evolves according to the law of the cluster process

(G_t°, o_t), so has the same distribution.

Lemma 2.1. *With probability* 1*, for sufficiently large* t *we have* $(G_t, o_t) = (H_t, o_t)$.

Proof. We refer to edges of $\overline{G}_t$ which were added after time 0 as *new* edges, and those which correspond (after replacements when vertices split) to edges of G as *old* edges. Let $e \in E(G)$ be an edge from the root, and let the corresponding edge at time t meet o'_t, where o'_t is a descendant of the root. We say that e has been *killed* by time t if, for some $s \le t$, we have $o'_s \neq o_s$ and no new edges meet o'_s. If e has been killed by time t, then at time s all paths from o_s to o'_s must use at least one old edge, and this property is preserved by splitting events, so the same is true for t. Thus, if a path from the root in $\overline{G}_t$ uses any old edge, the first old edge in that path must not have been killed by time t, meaning that the old edges which have not been killed by time t form a cut separating H_t from the rest of G_t. It therefore suffices to show that with probability 1 eventually every old edge has been killed.

For a specified edge e, consider the first time that the root splits and $o'_t \neq o_t$; call this t_1. Let $t_2, t_3, \dots$ be the subsequent times that o'_t splits, and let X_k be the number of new edges meeting o'_{t_k}. Then $X_{k+1} \sim \mathrm{Bin}(X_k, 1/2) + \mathrm{Po}(\lambda/2)$. This gives an irreducible Markov chain on $\mathbb{N}$ with a stationary distribution $\mathrm{Po}(\lambda)$. As a result, the chain is positive recurrent and in particular hits 0 in finite time, killing e, with probability 1. Since there were finitely many old edges, all of them are killed in finite time with probability 1. □

Before proceeding to the proof of Theorem 1.4, we first recall the *Poisson edge model* of [9]. This is a long-range percolation model on the leaves of the canopy tree. We may label the complete binary trees of height $0, 1, \dots$ in such a way that each tree is a subtree of the next, with each leaf also being a leaf of the next tree. The (binary) *canopy tree* is then the union of this sequence of trees, and has an infinite sequence of leaves. The Poisson edge model is a random multigraph whose vertices are the leaves of the canopy tree, and whose edges are given by independently placing $\mathrm{Po}(2^{1-d(x,y)}\lambda)$ edges between each pair of leaves x, y, where $d(x, y)$ is the graph distance on the canopy tree. In [9] it is shown that the unique random rooted connected multigraph having finite expected root degree which is invariant under the synchronous version of the cluster process is given by the cluster of the root in the Poisson edge model. For the cluster process of Definition 1.3, the picture will be more complicated. Note that we may define the T-Poisson edge model for any binary tree T in the same way: it is the random multigraph on the leaves of T, with $\mathrm{Po}(2^{1-d_T(x,y)}\lambda)$ edges independently between each pair of leaves x, y. We shall need a simple observation about the T-Poisson edge model.

Let T be any binary tree, and fix an edge uv. We say that an edge of the T-Poisson edge model *crosses* uv if its endpoints are in different components of $T - uv$.

Lemma 2.2. *The probability that the* T*-Poisson edge model has no edges which cross* uv *is at least* $\mathrm{e}^{-\lambda}$.

Proof. Write L_u, L_v for the leaves of the components containing u and v respectively. The number of such edges is $\mathrm{Po}(z\lambda)$ where

$$\begin{aligned} z &= \sum_{x \in L_u} \sum_{y \in L_v} 2^{1-d_T(x,y)} \\ &= \Big(\sum_{x \in L_u} 2^{-d_T(x,u)} \Big) \Big(\sum_{y \in L_v} 2^{-d_T(v,y)} \Big). \end{aligned}$$

We must therefore check that $z \le 1$. Consider a random walk on the component of $T - uv$ containing u started at u and constrained to increase the distance from u at every step, stopping if it reaches a leaf. Then for $x \in L_u$ the probability this walk stops at x

is $2^{-d_T(x,u)}$, since there are two possible moves at each step. Thus $\sum_{x\in L_u} 2^{-d_T(x,u)} \leq 1$, and the same argument applies to L_v, giving the result. □

Remark 2.3. In fact provided that T has countably many ends we have equality in Lemma 2.2, since both walks terminate almost surely.

Proof of Theorem 1.4. We will construct a random multigraph $(M(\lambda), o)$ with the property that $(M(\lambda)_t, o_t)$ has the same distribution for any $t \geq 0$. To show uniqueness, we will show that (G_t°, o_t) converges in distribution to $(M(\lambda), o)$, and apply Lemma 2.1.

Our construction of $(M(\lambda), o)$ will use the T-Poisson edge model, working with a random tree T. (This tree can be thought of as the local limit of the Yule tree at time t, or equivalently the ball of radius t in first passage percolation on the full binary tree with $\mathrm{Exp}(1)$ edge costs, after re-rooting at the leaf reached by a simple forward random walk from the root.)

To begin with, we construct some finite random trees $T(t)$ that will form the building blocks in the construction of T. Given a parameter $t > 0$, we define a random rooted binary tree $T(t)$ as follows. Start from a single-vertex rooted tree, with an exponential clock of rate 1 on the root. Whenever a clock on a vertex v rings, add two children of v, each with their own independent exponential clocks of rate 1 (do not replace the clock on v; each vertex rings at most once). Continue until time t. Note that $T(t)$ is almost surely finite. Next we construct an infinite random tree T. Start from an infinite path $P = v_0 v_1 \cdots$, and label its edges with an infinite sequence $s_1, s_2, \ldots$ of i.i.d. $\mathrm{Exp}(1)$ random variables. For each $i > 0$, sample a copy T_i of $T(\sum_{j\leq i} s_j)$, denote its root by w_i, and join T_i to P with the edge $v_i w_i$. Here each T_i is sampled independently.

Having constructed T, consider the T-Poisson edge model. We let $M(\lambda)$ be the component of v_0 in this random multigraph, and let v_0 be the root of $M(\lambda)$. For $n \in \mathbb{N}$, let L_n be the leaves of the component of $T - v_n v_{n+1}$ containing v_n.

Claim 2.4. *With probability* 1, $V(M(\lambda)) \subseteq L_n$ *for* n *sufficiently large.*

Proof of Claim. Starting from $k = 0$, iteratively reveal the number of edges of the T-Poisson edge model between pairs of vertices until an edge crossing $v_k v_{k+1}$ is found. If this happens, update k to be the smallest value such that no edge yet revealed crosses $v_k v_{k+1}$ and continue revealing. By Lemma 2.2, for each different value of k considered there is a probability of at least $\mathrm{e}^{-\lambda}$ that no suitable edge is ever found, no matter what was previously revealed. Thus almost surely one of the edges $v_k v_{k+1}$ is not crossed, meaning that $V(M(\lambda)) \subseteq L_k$. ♢

Thus $M(\lambda)$ almost surely contains vertices from finitely many of the subtrees T_i. In particular, since each T_i is almost surely finite, so is $M(\lambda)$.

Claim 2.5. $(M(\lambda)_t, o_t)$ *has the same distribution as* $(M(\lambda), o) = (M(\lambda)_0, o_0)$.

Proof of Claim. Recall that the construction of $M(\lambda)$ was based on the randomly edge-labelled path P. Let us denote by $G(P, \lambda)$ the random graph constructed from any path P with edges bearing positive real labels by following the above procedure. To compare $M(\lambda)$ with $M(\lambda)_t$, we will express the latter as $G(P_t, \lambda)$ for an appropriate randomly labelled path P_t: consider a Poisson point process $R = (-t_1, -t_2, \ldots, -t_k), k \geq 0$ on the interval $[-t, 0]$ (where we assume that $t_i \geq t_{i+1}$) governed by Lebesgue measure and with duration 1. We obtain P_t from P as follows. We change the label s_1 of the first edge of P into $s_1 + t_k$ if $k \geq 1$, or into $s_1 + t$ if $k = 0$. Moreover, we append k edges at the start of P, and label them as follows. The first edge is labelled $t - t_1$, and for $i = 2, \ldots, k$, the ith edge is labelled $t_{i-1} - t_i$. It is straightforward to check that $G(P_t, \lambda)$ is identically distributed with $(M(\lambda)_t, o_t)$ by identifying the times at which the root is split with the

reversal $t_k, \dots, t_2, t_1$ of R, using the fact that $t_{i-1} - t_i$ has distribution $\mathrm{Exp}(1)$, and so do t_k and $t - t_1$.

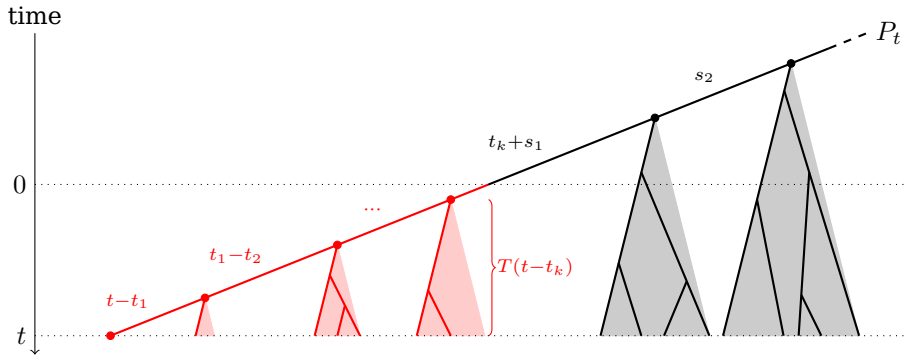

Figure 1: Construction of $M(\lambda)_t$ via P_t (proof of Claim 2.5).

To finish the proof that $(M(\lambda)_t, o_t) = G(P_t, \lambda)$ has the same distribution as $(M(\lambda), o) = G(P, \lambda)$, it suffices to prove that P_t has the same distribution as P. To prove this, note that we can sample the labels $s_1, s_2, \dots$ of P as a Poisson point process on the real axis $[0, \infty)$ governed by Lebesgue measure and with duration 1. Similarly, we can sample the labels of P_t as the gaps of a Poisson point process on $[-t, \infty)$. But these two Poisson point processes are identically distributed once we shift by t, as required. ♢

Next, we show that G_t° converges in distribution to $M(\lambda)$. To begin with, we can obtain G_t° by a construction similar to that of $M(\lambda)$, by keeping track of the genealogical tree T_t of the vertices of G_t°: the vertex set of T comprises all vertices that appeared throughout the process $G_s^\circ, 0 \le s \le t$, and if a vertex v was replaced with v_1, v_2 at some time $s \le t$, we join v with an edge to each of v_1, v_2. Note that the vertex set of G_t° is contained in the set of leaves of T_t. To sample the edges of G_t°, we put $\mathrm{Po}(2^{1-d_{T_t}(x,y)}\lambda)$ parallel edges independently between any two leaves x, y of T_t, and identify G_t° with the component of o in the resulting multigraph.

The times $t_1, \dots, t_k$ when the root of G_t° splits are, by definition, given by a Poisson point process on $[0, t]$ governed by Lebesgue measure on that interval. Consequently, the "reversed" sequence of times $t - t_k, \dots, t - t_1$ has the same distribution as $t_1, \dots, t_k$. Using this fact, we may equivalently construct G_t° using $t - t_k, \dots, t - t_1$ as the splitting times of the root, while leaving the rest of the construction unchanged. This realisation of G_t° coincides, by definition, with the following construction. Start with a random path P_t with k edges $e_1, \dots, e_k$, where as above k is the number of splittings of o in the time interval $[0, t]$, labelling e_i with the time gap $s_i = t_{k+1-i} - t_{k-i}$ if $i = 2, \dots, k$ or $s_i = t - t_k$ if $i = 1$. Attach to the endvertex v_i of e_i an independent copy of $T(\sum_{j \le i} s_j)$ as above, and finally define a random graph on the leaves of the resulting tree by taking the component of the root in its Poisson edge model.

Appropriately coupled, $M(\lambda)$ and G_t° therefore give the same result so long as $M(\lambda)$ does not reach the end of the finite path P_t in the above construction. Write E_n for the event that $M(\lambda)$ does not extend past v_n. Given $\varepsilon > 0$, choose n such that $\mathbb{P}(E_n) < \varepsilon/2$ (which is possible by Claim 2.4) and t such that $\mathbb{P}(\mathrm{Po}(t) < n) < \varepsilon/2$.

For any set of isomorphism classes of rooted connected graphs $\mathcal{S}$, we have

$$\begin{aligned}\mathbb{P}(G_t^\circ \in \mathcal{S}) &\le \mathbb{P}(M(\lambda) \in \mathcal{S} \wedge E_n \wedge (s_1 + \dots + s_n < t)) + \mathbb{P}(E_n^\complement) + \mathbb{P}(s_1 + \dots + s_n \ge t) \\ &< \mathbb{P}(M(\lambda) \in \mathcal{S}) + \varepsilon,\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(G_t^\circ \in \mathcal{S}) &\geq \mathbb{P}(M(\lambda) \in \mathcal{S} \wedge E_n \wedge (s_1 + \cdots + s_n < t)) \\ &\geq \mathbb{P}(M(\lambda) \in \mathcal{S}) - \mathbb{P}(E_n^{\complement}) - \mathbb{P}(s_1 + \cdots + s_n \geq t) \\ &> \mathbb{P}(M(\lambda) \in \mathcal{S}) - \varepsilon.\end{aligned}$$

Thus G_t° converges in distribution to $M(\lambda)$ as $t \to \infty$. The uniqueness of $M(\lambda)$ now follows from Lemma 2.1, since if G is a random graph with G_t identically distributed for every t, that lemma implies that the distribution of G is the limit of the distribution of G_t°. □

The random multigraph $M(\lambda)$ described above differs from the corresponding multigraph $G(\lambda)$ for the synchronous case studied in [9], that is, the component of the root in the original Poisson edge model on the canopy tree. To see this, it is sufficient to consider the probability, conditional on $d(o) = 2$, of a double edge from the root. For $M(\lambda)$ this is $\sum_{x \neq o} 4^{1-d(o,x)}$, where the sum is taken over all other leaves of the random tree T. Note that the probability that w_1 is a leaf is $\mathbb{P}(\tau(w_1) < s_1)$, where $\tau(w_1)$ is the length of w_1's clock. Since $\tau(w_1)$ and s_1 are i.i.d., we have $\mathbb{P}(w_1 \text{ a leaf}) = 1/2$; clearly w_i is less likely to be a leaf than w_1 if $i > 1$, so each w_i is a leaf with probability at most $1/2$. For each $i \geq 1$, the probability of a double edge to a descendent of w_i is 4^{-i} if w_i is a leaf, and at most 4^{-i-1} otherwise (being maximised when both its offspring are leaves). So the probability of a double edge is at most $\sum_{i \geq 1}(4^{-i} + 4^{-i-1})/2 = 1/4$. For the canopy tree version $G(\lambda)$, the probability of a double edge is $\sum_{h \geq 1} 2^{h-1} 4^{1-2h} = 2/7$, and so $M(\lambda)$ has a strictly smaller double-edge probability.

3 Finite expected size

In this section, we consider the expected size $\mathbb{E}(|M(\lambda)|)$. While the expected size of $G(\lambda)$ is finite for every $\lambda > 0$ [9], it is not immediately clear whether the same is true of $M(\lambda)$. Since $M(\lambda)$ arises from the T-Poisson edge model on a random tree T, and we know that the expected cluster size is finite for the Poisson edge model on the canopy tree, and that the cluster size of the Poisson edge model on any binary tree is almost surely finite (Claim 2.4), one might hope to prove a universal bound (depending on λ) on the expected cluster size for any binary tree, whence the desired result would follow by averaging. However, no such bound exists; indeed, there are binary trees on which the expected cluster size of the Poisson edge model is infinite for sufficiently large λ. One example may be obtained by replacing each edge of the canopy tree by a two-edge path with a pendant leaf attached to the new vertex. If v was a leaf of the canopy tree at distance $2k$ from o, then the new tree contains a sequence of $2k + 2$ leaves, starting at o and ending at v, such that each consecutive pair is at distance 4. Each of these pairs is adjacent in the Poisson edge model on this tree with probability $1 - \mathrm{e}^{-\lambda/8}$, and so every such v is in the component of o with probability at least $(1 - \mathrm{e}^{-\lambda/8})^{2k+1}$. Provided $\lambda \geq 8 \log(2 + \sqrt{2})$, it follows that the expected size of this component is infinite.

Of course, the initial sections of such a tree are not typical Yule trees, and so this example does not rule out the possibility of exploiting the large-scale structure of the tree T constructed in the previous section. However, we will find it easier to use a more local approach: rather than showing that an initial section of T is typically well-behaved everywhere, we work directly with Definition 1.3 and explore the component of the root in $\overline{G}_t$. This means that we only need T, which corresponds to the splitting events, to behave well in such parts as we encounter during this exploration. In the remainder of the section, we prove Theorem 1.5.

3.1 Outline of proof

Fix $\lambda > 0$. Note that since G_t° converges in distribution to $M(\lambda)$ and both G_t° and $M(\lambda)$ are almost surely finite we have $\mathbb{E}(|G_t^\circ|) \to \mathbb{E}(|M(\lambda)|)$ as $t \to \infty$. Our basic strategy is to prove a bound $f(t)$ on $\mathbb{E}(|G_t^\circ|)$ which changes only slowly with t, and has a finite limit. Recall that G_t° is the component of the root in $\overline{G}_t^\circ$. First we bound the size of G_t° at time $t + \varepsilon$.

Lemma 3.1. *Fix times $t \geq 0$ and $\varepsilon > 0$, and let $X_\varepsilon = |\overline{G}_\varepsilon^\circ|$ be the total number of vertices in the full process at time ε. Then we have*

$$\mathbb{E}(|G_{t+\varepsilon}^\circ|) < (1-\varepsilon)\mathbb{E}(|G_t^\circ|) + \varepsilon\mathbb{E}(|G_{t+\varepsilon}^\circ| \mid X_\varepsilon = 2) + 4\varepsilon^2 \mathrm{e}^{t+\varepsilon}.$$

Proof. Conditioning on the value of X_ε, we have

$$\begin{aligned}\mathbb{E}(|G_{t+\varepsilon}^\circ|) = \mathbb{P}(X_\varepsilon = 1)\mathbb{E}(|G_{t+\varepsilon}^\circ| \mid X_\varepsilon = 1) + \mathbb{P}(X_\varepsilon = 2)\mathbb{E}(|G_{t+\varepsilon}^\circ| \mid X_\varepsilon = 2) \\ + \mathbb{P}(X_\varepsilon > 2)\mathbb{E}(|G_{t+\varepsilon}^\circ| \mid X_\varepsilon > 2).\end{aligned}$$

Note that, conditioned on $X_\varepsilon = 1$, $G_{t+\varepsilon}^\circ$ is just the result of letting the single vertex at time ε evolve for an additional time t, and $\mathbb{P}(X_\varepsilon = 1) = \mathrm{e}^{-\varepsilon} < 1 - \varepsilon + \varepsilon^2$, so

$$\begin{aligned}\mathbb{P}(X_\varepsilon = 1)\mathbb{E}(|G_{t+\varepsilon}^\circ| \mid X_\varepsilon = 1) &< (1 - \varepsilon + \varepsilon^2)\mathbb{E}(|G_t^\circ|) \\ &< (1-\varepsilon)\mathbb{E}(|G_t^\circ|) + \varepsilon^2 \mathrm{e}^{t+\varepsilon}.\end{aligned}$$

Also, $\mathbb{P}(X_\varepsilon = 2) < \varepsilon$, which gives the required second term.

To deal with the third term, recall from Remark 1.2 that $X_\varepsilon \sim \mathrm{Geo}(\mathrm{e}^{-\varepsilon})$ and so $\mathbb{P}(X_\varepsilon > 2) = (1 - \mathrm{e}^{-\varepsilon})^2 < \varepsilon^2$. Now suppose that $X_\varepsilon > 2$. This means that there is some random time $\eta_2 < \varepsilon$ at which the second splitting event occurs. Nothing that happens after η_2 can affect the event $X_\varepsilon > 2$, and so we may condition on η_2. At time η_2 there are three vertices, which may or may not be connected by edges. Certainly $|G_{t+\varepsilon}^\circ|$ is dominated by $|\overline{G}_{t+\varepsilon}^\circ|$, which, conditioned on η_2, has expectation $3\mathrm{e}^{t+\varepsilon-\eta_2} < 3\mathrm{e}^{t+\varepsilon}$. Thus the final term is less than $3\varepsilon^2\mathrm{e}^{t+\varepsilon}$, as required. □

Conditioned on $X_\varepsilon = 2$, $\overline{G}_{t+\varepsilon}^\circ$ is distributed as two independent copies of the full process run for time t with some edges between them, rooted at the root of the first copy. We will show that the probability of some of these edges touching the component of the root in the first copy is exponentially small. If this does happen, we argue that the expected number of edges between the two copies is not much larger than its unconditional expectation (i.e. $\lambda/2$), and that consequently we connect together (on average) not too many components. The main issue with this is that conditioning on this unlikely event might change the expected size of a component significantly, so we must control this. If we can do this, we will have shown that

$$\mathbb{E}(|G_{t+\varepsilon}^\circ| \mid X_\varepsilon = 2) \leq (1 + h(t))\mathbb{E}(|G_t^\circ|), \tag{3.1}$$

where $h(t)$ is some function that decays exponentially in t. It will follow, from (3.1) and Lemma 3.1, that for any fixed $t \geq 0$ we have,

$$\limsup_{\varepsilon \to 0+} \frac{\mathbb{E}(|G_{t+\varepsilon}^\circ|) - \mathbb{E}(|G_t^\circ|)}{\varepsilon} \leq h(t)\mathbb{E}(|G_t^\circ|),$$

and so if $f : [0, \infty) \to [0, \infty)$ is a function satisfying $f(0) = 1$ and $f'(t) = h(t)f(t)$, then $f(t) \geq \mathbb{E}(|G_t^\circ|)$ for each t. Now for this f we have $\frac{\mathrm{d}}{\mathrm{d}t} \log f(t) = h(t)$ and so

$$\lim_{t \to \infty} f(t) = \exp \int_0^\infty h(s)\mathrm{d}s < \infty.$$

Write $G_t^{\circ\circ}$ for the result of running the cluster process for time t starting from two vertices with $N \sim \mathrm{Po}(\lambda/2)$ number of edges between. We consider the descendants of the two original vertices in the corresponding full process $\overline{G}_t^{\circ\circ}$ as two independent copies of $\overline{G}_t^{\circ}$, with the "left" copy being descendants of the original root, and say that the N edges between the two copies are *old*, and others are *new*. Note that the component of the root in the subgraph induced by the left copy is distributed as G_t°. We follow what happens to the left-endpoints of all old edges, and to the root. Recall that an old edge is *killed* by a splitting event if after that event its left-endpoint is not the root, and meets no new edges. Consider the following four events, for a fixed time t and $0 < \alpha < 1$; note that some of these events may depend on what happens after time t.

A: the left-endpoint of some old edge splits less than αt times by time t.

B: after the left-endpoint of some old edge splits $\alpha t/3$ times, it is either the root or the left-endpoint of more than one old edge.

C: B does not occur, but some new edge meets the left-endpoint of some old edge for the entire period between the $(\alpha t/3)$th and $(2\alpha t/3)$th splits of the latter.

D: B and C do not occur, but some old edge is not killed between its $(2\alpha t/3)$th and αtth splits.

Writing $\mathbb{I}_A$, etc., for the indicator functions of these events, we have

$$\begin{aligned}\mathbb{E}(|G_t^{\circ\circ}|) &\le \mathbb{E}(|G_t^{\circ\circ}|(\mathbb{I}_A + \mathbb{I}_B + \mathbb{I}_C + \mathbb{I}_D + \mathbb{I}_{(A\cup B\cup C\cup D)^{\complement}})) \\ &\le \sum_{E\in\{A,B,C,D\}} \mathbb{E}(|G_t^{\circ\circ}| \mid E)\mathbb{P}(E) + \mathbb{E}(|G_t^{\circ\circ}| \mid (A\cup B\cup C\cup D)^{\complement}). \qquad (3.2)\end{aligned}$$

3.2 Dealing with event A

Note that the left-endpoint of a given edge splits $\mathrm{Po}(t)$ times in time t, and

$$\mathbb{P}(\mathrm{Po}(t) \le \lfloor \alpha t \rfloor) \le (\lfloor \alpha t \rfloor + 1)\frac{\mathrm{e}^{-t}t^{\lfloor \alpha t \rfloor}}{\lfloor \alpha t \rfloor!} = O(\sqrt{t})(\mathrm{e}^{\alpha-1}\alpha^{-\alpha})^t.$$

Since $\lim_{\alpha\to 0+}\alpha^\alpha = 1$, we may choose $\alpha > 0$ such that

$$\mathbb{P}(A) \le \frac{\lambda}{2}\mathbb{P}(\mathrm{Po}(t) \le \lfloor \alpha t \rfloor) = O\big(\mathrm{e}^{(\mathrm{e}^{-\lambda}/2-1)t}\big). \qquad (3.3)$$

We next define a variant of the full process: the *singleton-free process* S_t starts from a single vertex with $\mathrm{Po}(\lambda/2)$ tokens. It proceeds as the full process with tokens distributed randomly between the offspring when a vertex splits, but with the exception that any vertex which is isolated and has no tokens is immediately discarded.

First we will show that $\mathbb{E}(|S_t|)$ is bounded by the expected size of a Yule process of rate $r = r(\lambda) < 1$. The intuition here is that each splitting event has at least a constant probability of producing an isolated vertex, and we can just ignore these events, resulting in a thinning of the rate by a constant factor. However, we need to be slightly careful to check that the lower bound on the probability of creating an isolated vertex still holds even conditioned on the splitting vertex not having been isolated at any point in its history. We will need the following lemma, which will be used again for the other events.

Lemma 3.2. *If $X \sim \mathrm{Po}(m)$ and $Y \sim \mathrm{Bin}(X, p)$ for some $p \in (0, 1]$, then $X \mid (Y \ge k)$ is stochastically dominated by $k + \mathrm{Po}(m)$.*

Proof. We handle the case $p = 1$; the case $p < 1$ follows by noting that $Y \sim \mathrm{Po}(pm)$ and $X - Y$ are independent. We may sample $X \mid (X \ge k)$ by repeatedly sampling X, keeping

the first value which is at least k. Since we can take the rth sample of X as the number of points occurring in the interval $[r-1, r]$ in a Poisson process of intensity m, this is the same as letting the Poisson process run until the first time we have seen k points since the last integer, then continuing until the next integer. This is clearly dominated by letting the process run to the first time we have seen k points since the last integer, then continuing for time 1, which gives the required distribution. □

If we only have the weaker condition that $X \leq_{\mathrm{st}} \mathrm{Po}(m)$ we cannot get any bound on $\mathbb{E}(X \mid Y \geq 1)$, but the following bounds are sufficient for our purpose.

Corollary 3.3. *Suppose $X \leq_{\mathrm{st}} \mathrm{Po}(m)$ is nonnegative, and $Y \sim \mathrm{Bin}(X, p)$ for some $p \in (0, 1]$. Then $\mathbb{P}(Y \geq 1) \leq pm$ and $\mathbb{P}(Y \geq 1)\mathbb{E}(X \mid (Y \geq 1)) \leq pm(m+1)$.*

Proof. We can couple X with a variable $X' \sim \mathrm{Po}(m)$, and couple Y with $Y' \sim \mathrm{Bin}(X', p)$ in the natural way, so that $(Y \geq 1) \subseteq (Y' \geq 1)$. Then, since $0 \leq X \leq X'$ we have

$$\begin{aligned}\mathbb{P}(Y \geq 1)\mathbb{E}(X \mid Y \geq 1) &\leq \mathbb{P}(Y \geq 1)\mathbb{E}(X' \mid Y \geq 1)\\ &\leq \mathbb{P}(Y' \geq 1)\mathbb{E}(X' \mid Y' \geq 1).\end{aligned}$$

Lemma 3.2 gives $\mathbb{E}(X' \mid Y' \geq 1) \leq m+1$, and $\mathbb{P}(Y \geq 1) \leq \mathbb{P}(Y' \geq 1) \leq \mathbb{E}(Y') = mp$. □

We may sample S_t by using a Yule process Y_t of rate 1 for the splitting events, then determining the movement of new edges and removing any vertices which were isolated at any point in their history. Similarly, we can simulate a Yule process $Y_t^{(r)}$ of rate $r < 1$ from the same copy of Y_t by, independently for each splitting event, removing all descendants of one offspring with probability $1-r$. Conditional on Y_t, each vertex which has undergone k splitting events has probability $\left(\frac{r+1}{2}\right)^k$ of surviving in $Y_t^{(r)}$. In S_t, conditional on a vertex having survived j splits without being isolated, we argue by induction on j that the number of edge-ends meeting it is dominated by $\mathrm{Po}(1+\lambda)$. This is true for $j = 0$. Assuming the statement holds for j, a vertex which has split $j+1$ times may inherit the first edge-end from its parent, and receives at most $\mathrm{Po}(\lambda)$ other edge-ends; conditioning on not being isolated does not change the number of additional edge-ends if it did inherit the first edge, and increases it to at most $1 + \mathrm{Po}(\lambda)$ if not. So the result holds for all j. Consequently, given that a vertex has survived j splits without being isolated, its offspring after the next split have at most $\mathrm{Ber}(1/2) + \mathrm{Po}(\lambda)$ edge-ends, so are each isolated with probability at least $\frac{\mathrm{e}^{-\lambda}}{2}$; it follows that each vertex in Y_t which underwent k splitting events has probability $\left(\frac{2-\mathrm{e}^{-\lambda}}{2}\right)^k$ of surviving in S_t. For $r = 1 - \mathrm{e}^{-\lambda}$, each vertex has a higher probability of surviving in $Y_t^{(r)}$ than S_t, and so we have

$$\mathbb{E}(|S_t|) \leq \mathbb{E}(|Y_t^{(r)}|) = \mathrm{e}^{(1-\mathrm{e}^{-\lambda})t}.$$

Lemma 3.2 implies that $N \mid A \leq_{\mathrm{st}} 1 + \mathrm{Po}(\lambda/2)$. To see this, note that we may first condition on the tree of splitting events. For each possible tree T, each old edge independently has some probability p_T of following a path in the tree which splits fewer than αt times; we may ignore trees for which $p_T = 0$. Thus $N \mid T, A \leq_{\mathrm{st}} 1 + \mathrm{Po}(\lambda/2)$, and the result follows by averaging over T.

Now we consider the singleton-free process conditional on A. Suppose a vertex meeting an old edge or root splits, and one of the new vertices created, v, does not meet a new edge or the root. Conditioning on A does not affect the future evolution of v, and it evolves as the non-root half of a singleton-free process (or is discarded if it has no new edges). Thus the expected number of descendants of v is at most $\mathbb{E}(|S_t|)/2 = \mathrm{e}^{(1-\mathrm{e}^{-\lambda})t}$. For each old edge, the expected number of times it splits is t before conditioning on A, and cannot increase after conditioning; the same applies to the root. Thus the expected

number of times such a vertex is created is at most $(2+\lambda)t$. Since every vertex at time t is either a descendant of such a vertex, meets an old edge, or is the root, we have

$$\mathbb{E}(|G_t^{\circ\circ}| \mid A) \le \mathbb{E}(|S_t| \mid A) \le 2 + \lambda + (2+\lambda)t\mathrm{e}^{(1-\mathrm{e}^{-\lambda})t},$$

and so (recalling (3.3)) we have

$$\mathbb{P}(A)\mathbb{E}(|G_t^{\circ\circ}| \mid A) = O(t\mathrm{e}^{-t\mathrm{e}^{-\lambda}/2}). \tag{3.4}$$

3.3 Dealing with event B

Since there are $X \sim \mathrm{Po}(\lambda/2)$ left-endpoints of old edges and one root, and each pair has probability $2^{-\alpha t/3}$ of coinciding after $\alpha t/3$ splits, a union bound gives

$$\mathbb{P}(B) \le \mathbb{E}\left(\binom{X+1}{2}\right)2^{-\alpha t/3} = \left(\frac{\lambda^2}{8} + \frac{\lambda}{2}\right)2^{-\alpha t/3}. \tag{3.5}$$

Consider the full tree of possible locations for left-endpoints after $\alpha t/3$ splits. Order these locations $v_1, \dots, v_{2^{\alpha t/3}}$; without loss of generality we may assume the root is at v_1 after $\alpha t/3$ splits. Writing X_i for the number of old edges at location v_i after $\alpha t/3$ splits, the X_i are i.i.d. $\mathrm{Po}(2^{-\alpha t/3}\lambda/2)$ random variables. We will control the expected number of old edges conditioned on B. B occurs if and only if either $X_1 \ge 1$ or $X_i \ge 2$ for some $i > 1$. For each $i \ge 2$, let B_i be the event that $X_i \ge 2$, and $X_j \le 1$ for each $j > i$. Let B_1 be the event that $X_1 \ge 1$ but $X_j \le 1$ for each $j > 1$. Now the events $(B_i)_{i=1}^{2^{\alpha t/3}}$ form a partition of B, and $N = \sum_{j=1}^{2^{\alpha t/3}} X_j$. Lemma 3.2 gives $\mathbb{E}(X_i \mid B_i) \le \mathbb{E}(X_i) + 2$, and $\mathbb{E}(X_j \mid B_i) \le \mathbb{E}(X_j)$ if $j \ne i$, so $\mathbb{E}(N \mid B_i) \le \mathbb{E}(N) + 2$ for each i. Thus

$$\mathbb{E}(N \mid B) = \sum_{i=1}^{2^{\alpha t/3}} \mathbb{P}(B_i \mid B)\mathbb{E}(N \mid B_i) \le \lambda/2 + 2.$$

The old edges therefore combine, on average and conditional on B, at most $\lambda/2 + 3$ components from the two copies. Since B does not depend on splitting times or new edges, each component has expected size $\mathbb{E}(|G_t^{\circ}|)$. Thus, recalling (3.5), we have

$$\mathbb{P}(B)\mathbb{E}(|G_t^{\circ\circ}| \mid B) = O(2^{-\alpha t/3})\mathbb{E}(|G_t^{\circ}|). \tag{3.6}$$

3.4 Dealing with event C

Randomly designate one end of each new edge to be the "head", and the other the "tail", so that the number of edges xy with head x and the number with head y are independent. We set C_{h} (C_{t}) to be the event that the head (the tail) of some new edge coincides with the left end of some old edge for the period in question. Since $C = C_{\mathrm{h}} \cup C_{\mathrm{t}}$ and by symmetry of $C_{\mathrm{h}}, C_{\mathrm{t}}$, we have $\mathbb{P}(C)\mathbb{E}(|G_t^{\circ\circ}| \mid C) \le 2\mathbb{P}(C_{\mathrm{h}})\mathbb{E}(|G_t^{\circ\circ}| \mid C_{\mathrm{h}})$.

We first condition on $B^{\complement}$; since $\mathbb{P}(B \mid N = n)$ is increasing in n, $(N \mid B^{\complement}) \le_{\mathrm{st}} \mathrm{Po}(\lambda/2)$. Given $B^{\complement}$, each of the $N \mid B^{\complement}$ old edges coincides with the head of a new edge for the period between its $(\alpha t/3)$th and $(2\alpha t/3)$th splits independently and with equal probability p. Since the number of heads coinciding with a given old edge at the start of the period is distributed $\mathrm{Po}(\kappa)$ for some fixed $\kappa \le \lambda/2$, a union bound gives $p \le 2^{-\alpha t/3}\lambda/2$. Thus

$$\mathbb{P}(C) \le 2\mathbb{P}(C_{\mathrm{h}} \mid B^{\complement}) \le 2^{-\alpha t/3}\lambda^2/2. \tag{3.7}$$

Also, Corollary 3.3 gives $\mathbb{P}(C_{\mathrm{h}} \mid B^{\complement})\mathbb{E}(N \mid C_{\mathrm{h}}) \le 2^{-\alpha t/3}(\lambda/2)^2(1+\lambda/2)$.

Next we bound $\mathbb{P}(C_{\mathrm{h}})\mathbb{E}(|G_t^{\circ\circ}| \mid C_{\mathrm{h}})$. Lemma 3.2 implies that the number of heads which coincide with a given old edge, after conditioning on C_{h}, is dominated by $1+\mathrm{Po}(p\kappa)$,

and the number of other new edges is independent of C_{h}. Thus we may couple the new edges conditioned on C_{h} as a subgraph of the unconditioned new edges together with at most $N \mid C_{\mathrm{h}}$ additional new edges. The expected size of the component of a given vertex in the subgraph of unconditional new edges is $\mathbb{E}(|G_t^\circ|)$, and since the old edges and additional new edges merge at most $2\mathbb{E}(N \mid C_{\mathrm{h}}) + 1$ components on average,

$$\begin{aligned}\mathbb{P}(C)\mathbb{E}(|G_t^{\circ\circ}| \mid C) &\leq 2\mathbb{P}(C_{\mathrm{h}})(2\mathbb{E}(N \mid C_{\mathrm{h}}) + 1)\mathbb{E}(|G_t^\circ|) \\ &= \mathbb{E}(|G_t^\circ|)O(2^{-\alpha t/3}). \end{aligned} \tag{3.8}$$

3.5 Dealing with event D

Now suppose that B and C do not occur. Again, we have $N \mid B^{\complement} \cap C^{\complement} \leq_{\mathrm{st}} \mathrm{Po}(\lambda/2)$. Since C does not occur, all new edges that meet left ends of old edges after $2\alpha t/3$ splits were created after the $(\alpha t/3)$th split, and since B does not occur, each of these meets only one old edge. Thus every old edge is not killed between its $(2\alpha t/3)$th and αtth splits independently with some probability p. Corollary 3.3 therefore gives $\mathbb{P}(D)\mathbb{E}(N \mid D) \leq p(\lambda/2)(\lambda/2 + 1)$.

We next bound p. Note that being killed is monotone on adding new edges. Suppose that after a given split an old edge e meets $X_0 \leq_{\mathrm{st}} 1 + \mathrm{Po}(\lambda)$ new edges. Adding extra new edges, if necessary, we may assume e meets $1 + \mathrm{Po}(\lambda)$ new edges. After the next split, conditioned on e meeting at least one new edge, we claim that it meets $X_1 \leq_{\mathrm{st}} 1 + \mathrm{Po}(\lambda)$ new edges. If the first of the $1 + \mathrm{Po}(\lambda)$ new edges still meets e, there are $\mathrm{Po}(\lambda/2) + \mathrm{Po}(\lambda/2) \sim \mathrm{Po}(\lambda/2)$ other new edges meeting e, whereas if not we have $\mathrm{Po}(\lambda)$ edges meeting e, conditioned to be positive, and by Lemma 3.2 this is dominated by $1 + \mathrm{Po}(\lambda)$. Thus conditioning on not having been killed at the previous step leaves at most $1 + \mathrm{Po}(\lambda)$ new edges meeting e, giving a probability of at least $\frac{1}{2}\mathrm{e}^{-\lambda/2}$ of being killed at the next step; write $c_\lambda = 1 - \frac{1}{2}\mathrm{e}^{-\lambda/2}$. It follows that $p \leq c_\lambda^{\alpha t/3}$ and so

$$\mathbb{P}(D) \leq \mathbb{E}(N \mid B^{\complement} \cap C^{\complement})p \leq c_\lambda^{\alpha t/3}\lambda/2. \tag{3.9}$$

For each old edge e, we associate each new edge e' which meets e at any point between its $(2\alpha t/3)$th and αtth splits with the interval for which it meets e, i.e. the set of indices in $\{\alpha t/3, \dots, \alpha t\}$ of splits after which e and e' meet. Denote the number of new edges meeting e for an interval I by $X_{e,I}$; note that $X_{e,I} \sim \mathrm{Po}(\kappa_I)$ for some κ_I depending only on I, and all these are independent. We now condition on the number of old edges and which pairs e, I have $X_{e,I} \geq 1$; this is sufficient information to determine whether D occurs. Lemma 3.2 gives $X_{e,I} \mid D \leq_{\mathrm{st}} 1 + \mathrm{Po}(\kappa_I)$ for each e, I. We can thus couple the new edges conditioned on D as a subgraph of the unconditioned new edges together with at most $(N \mid D)(2\alpha t/3)^2$ additional new edges. As in Section 3.4, it follows that

$$\mathbb{P}(D)\mathbb{E}(|G_t^{\circ\circ}| \mid D) = \mathbb{E}(|G_t^\circ|)O(t^2 c_\lambda^{\alpha t/3}). \tag{3.10}$$

3.6 Final bounds

If none of A, B, C, D occur then all old edges have been killed. Since this means any path from the root uses only new edges (see the proof of Lemma 2.1), the component of the root is entirely within the left half, and thus we have

$$\begin{aligned}\mathbb{E}(|G_t^{\circ\circ}| \mid (A \cup B \cup C \cup D)^{\complement}) &= \mathbb{E}(|G_t^\circ| \mid (A \cup B \cup C \cup D)^{\complement}) \\ &\leq \mathbb{E}(|G_t^\circ|)/\mathbb{P}((A \cup B \cup C \cup D)^{\complement}).\end{aligned}$$

Combining (3.3), (3.4), (3.5), (3.6), (3.7), (3.8), (3.9) and (3.10) using (3.2), we have

$$\mathbb{E}(|G_t^{\circ\circ}|) = (1 + o(\zeta^t))\mathbb{E}(|G_t^\circ|),$$

for some $\zeta < 1$, as required for (3.1), which thus completes the proof of finiteness.

4 A sharp threshold for connectedness

In this section we prove Theorem 1.6, giving a sharp threshold for connectedness of $\overline{G}_t^\circ$. We show that, as for the binomial random graph, it coincides with the threshold for isolated vertices to appear. Our methods in this section will follow those of [9] closely. For both directions we will need the following simple concentration bound.

Lemma 4.1. *Let $f(t) : (0,\infty) \to (0,\infty)$ be any function with $f(t) \to \infty$ as $t \to \infty$. Then with high probability we have $\mathrm{e}^{t-f(t)} < |\overline{G}_t^\circ| < \mathrm{e}^{t+f(t)}$.*

Proof. Set $n_1 = \lceil \mathrm{e}^{t-f(t)} \rceil$ and $n_2 = \lfloor \mathrm{e}^{t+f(t)} \rfloor$. Since $|\overline{G}_t^\circ| \sim \mathrm{Geo}(\mathrm{e}^{-t})$, we have

$$\mathbb{P}(n_1 < |\overline{G}_t^\circ| < n_2) = (1-\mathrm{e}^{-t})^{n_1} - (1-\mathrm{e}^{-t})^{n_2-1} \underset{t\to\infty}{\longrightarrow} 1. \qquad \square$$

We first show that isolated vertices appear soon after time λ.

Proposition 4.2. *Let $f(t) : (0,\infty) \to (0,\infty)$ be any function with $f(\lambda) \to \infty$ as $\lambda \to \infty$. If $t > \lambda + f(\lambda)$ then as $\lambda \to \infty$ with high probability $\overline{G}_t^\circ$ has an isolated vertex.*

Proof. For technical reasons we prove the same statement for the modified process obtained by adding $\mathrm{Po}(\lambda/2)$ extra edges at the first splitting event. This ensures that each vertex has the same probability $\mathrm{e}^{-\lambda}$ of being isolated. Conditioned on the tree $T(t)$ defined in the proof of Theorem 1.4, [9, Lemma 7.1] applies and gives $\mathbb{P}(X \mid T(t)) \le 2/(2+|\overline{G}_t^\circ|\mathrm{e}^{-\lambda})$, where X is the event that no vertex is isolated. Note that $\lambda + f(\lambda)/2 < t - g(t)$, where $g(t)$ is another function satisfying $g(t) \to \infty$. By Lemma 4.1, with high probability $|\overline{G}_t^\circ| > \mathrm{e}^{t-g(t)}$; conditional on this we have $\mathbb{P}(X) \le 2/(2+\mathrm{e}^{f(\lambda)/2}) = o(1)$. $\square$

To complete the proof of Theorem 1.6, we must show that with high probability $\overline{G}_t^\circ(\lambda)$ is connected shortly before $t = \lambda$. For this we need another result from [9], but first we define some terms used. Fix a finite binary tree T representing descendants of a marked apex vertex, and $k \in \mathbb{N}$. We say that two vertices are *siblings* if they have the same parent, and two pairs of siblings are *k-cousins* if they have a common ancestor which is at most distance k on T from all of them. Let G be a graph whose vertices are leaves of T. We say two siblings x, y are *strongly linked* by G if G contains an edge between a descendant of x and a descendant of y, and *weakly linked* by G if there is some vertex z of T which is a sibling of one of the k lowest ancestors of x, y, such that G contains edges between a descendant of x and one of z, and between a descendant of y and one of z. [9, Lemma 7.2] says that the following set of conditions is sufficient for G to be connected:

(i) every pair of siblings in T is either strongly linked or weakly linked by G;

(ii) of every two pairs which are k-cousins, at least one is strongly linked by G;

(iii) any pair of siblings within the top k layers of T are strongly linked by G.

Proposition 4.3. *For any $\alpha > 1$, if $t \le \lambda - \alpha \log \lambda$ then as $\lambda \to \infty$ with high probability $\overline{G}_t^\circ$ is connected.*

Proof. Regarding $\overline{G}_t^\circ$ as a random graph on the leaves $L(t)$ of $T(t)$, we will show the conditions above hold with high probability, for some suitable k. Choose $\alpha' > 0$ such that $\alpha - \alpha' > 1$; then, by Lemma 4.1, with high probability $|L(t)| < \mathrm{e}^{t+\alpha' \log t} < \mathrm{e}^{t+\alpha' \log \lambda}$, so

$$|T(t)| < 2\mathrm{e}^{t+\alpha' \log \lambda}. \tag{4.1}$$

Suppose (4.1) holds, and set $k = \log_2 \lambda$. The probability that a particular pair of siblings fails to be strongly linked is $\mathrm{e}^{-\lambda/2}$, and since each pair of siblings has at most $k2^k = \lambda \log_2 \lambda$ pairs of k-cousins, the total number of ways to choose two pairs of siblings

which are k-cousins is at most $\mathrm{e}^{t+\alpha' \log \lambda} \lambda \log_2 \lambda = \mathrm{e}^{t+(1+\alpha') \log \lambda} \log_2 \lambda = o(\mathrm{e}^{\lambda})$. For each such choice, the probability that neither pair is strongly linked by $\overline{G}_t^{\circ}(\lambda)$ is $\mathrm{e}^{-\lambda}$ and so with high probability (ii) holds. There are at most λ pairs of siblings in the top k layers of $T(t)$, and so (iii) also holds with high probability. Finally, for a fixed pair of siblings below this point the probability that they are not strongly or weakly linked by $\overline{G}_t^{\circ}$ is

$$\mathrm{e}^{-\lambda/2}\big(1-(1-\mathrm{e}^{-\lambda/4})^2\big)\cdots\big(1-(1-\mathrm{e}^{-\lambda/2^{k-1}})^2\big) < 2^{k-1}\mathrm{e}^{-\lambda(1-2^{-k})}.$$

Thus the probability that some pair fails to be strongly or weakly linked is at most

$$\begin{aligned} 2^k \mathrm{e}^{-\lambda(1-2^{-k})} \mathrm{e}^{t+\alpha' \log \lambda} &= \lambda \mathrm{e}^{(t+\alpha' \log \lambda)-(t+\alpha \log \lambda)(1-1/\lambda)} \\ &= O(\lambda^{1+\alpha'-\alpha}) = o(1). \end{aligned}$$

□

References

[1] Backhausz, Á. and Móri, T. F. (2016) Further properties of a random graph with duplications and deletions. *Stoch. Models* **32**, 99–120. MR3457123

[2] Bebek, G., Berenbrink, P., Cooper, C., Friedetzky, T., Nadeau, J. and Sahinalp, S. C. (2006) The Degree Distribution of the Generalized Duplication Model. *Theoret. Comput. Sci.* **369**, 234–249. MR2277572

[3] Bhan, A., Galas, D. J. and Dewey, T. G. (2002) A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486–1493.

[4] Bienvenu, F., Débarre, F. and Lambert, A. (2019) The split-and-drift random graph, a null model for speciation. *Stochastic Process. Appl.* **129**, 2010–2048. MR3958422

[5] Bonato, A., Hadi, N., Horn, P., Prałat, P. and Wang, C. (2011) Models of on-line social networks. *Internet Math.* **6**, 285–313. MR2798106

[6] Duminil-Copin, H., Goswami, S., Raoufi, A., Severo, F. and Yadin, A. Existence of phase transition for percolation using the Gaussian Free Field. *Duke Math. J.* **169** (2020), no. 18, 3539–3563. MR4181032

[7] Feller, W. (1957) *An Introduction to Probability Theory and its Applications* vol. 1, 2nd ed. John Wiley & Sons, New York. MR0088081

[8] Georgakopoulos, A. (2016) Group-Walk Random Graphs. In *Groups, Graphs, and Random Walks* (T. Ceccherini-Silberstein, M. Salvatori and E. Sava-Huss, eds), Vol. 436 of the *LMS Lecture Note Series*, Cambridge University Press, pp. 190–204. MR3644009

[9] Georgakopoulos, A. and Haslegrave, J. (2020) Percolation on an infinitely-generated group. *Combin. Probab. Comput.* **29**, 587–615. MR4132522

[10] Georgakopoulos, A. and Panagiotis, C. Analyticity results in Bernoulli Percolation. *Mem. Amer. Math. Soc.*, to appear.

[11] Jagers, P. (1975) *Branching Processes with Biological Applications*. John Wiley & Sons, New York. MR0488341

[12] Jordan, J. (2011) Randomised reproducing graphs. *Electron. J. Probab.* **16**, 1549–1562. MR2827470

[13] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. and Upfal, E. (2000) Stochastic models for the web graph. *41st Annual Symposium on Foundations of Computer Science*, 57–65. MR1931804

[14] Pastor-Satorras, R., Smith, E. and Sole, R. V. (2003) Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199–210. MR2070214

[15] Southwell, R. and Cannings, C. (2010) Some models of reproducing graphs: I Pure reproduction. *Appl. Math.* **1**, 137–145.

[16] Thörnblad, E. (2015) Asymptotic degree distribution of a duplication-deletion random graph model. *Internet Math.* **11**, 289–305. MR3344243

[17] Vázquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003) Modelling of protein interaction networks. *Complexus* **1**, 38–44.

Acknowledgments. We are grateful to the anonymous referee for their very helpful comments.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System `http://www.vtex.lt/en/ejms.html`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`