KINGMAN'S COALESCENT ON A RANDOM GRAPH

LOUIGI ADDARIO-BERRY, CAELAN ATAMANCHUK, AND MAXWELL KAYE

Abstract

We introduce a generalization of Kingman's coalescent on $[n]$ that we call the *Kingman coalescent* on a graph $G = ([n], E)$. Specifically, we generalize a forest valued representation of the coalescent introduced in [ABE18]. The difference between the Kingman coalescent on G and the normal Kingman coalescent on $[n]$ is that two trees T_1, T_2 with roots ρ_1, ρ_2 can merge if and only if $\{\rho_1, \rho_2\} \in E$. When this process finishes (when there are no trees left that can merge anymore), we are left with a random spanning forest that we call a *Kingman forest* of G. In this article, we study the Kingman coalescent on Erdős-Rényi random graphs, $G_{n,p}$. We derive a relationship between the Kingman coalescent on $G_{n,p}$ and uniform random recursive trees, which provides many answers concerning structural questions about the corresponding Kingman forests. We explore the heights of Kingman forests as well as the sizes of their trees as illustrative examples of how to use the connection. Our main results concern the number of trees, $C_{n,p}$, in a Kingman forest of $G_{n,p}$. For fixed $p \in (0, 1)$, we prove that $C_{n,p}$ converges in distribution to an almost surely finite random variable as $n \to \infty$. For $p = p(n)$ such that $p \to 0$ and $np \to \infty$ as $n \to \infty$, we prove that $C_{n,p}$ converges in probability to $\frac{2(1-p)}{p}$.

1 INTRODUCTION

1.1 DEFINITIONS AND RESULTS

Let G be a finite graph with $|V(G)| = n$. A *rooted spanning forest* of G is a set $\{T_j, j \in [k]\}$ of vertex-disjoint, rooted subtrees of G with $\cup_{i=1}^{k} V(T_i) = V(G)$. For a rooted tree T we write $\rho(T)$ to denote the root of T. We always view the edges of a rooted tree as directed towards the root.

The *Kingman coalescent on G* is defined as follows. Let f_0 be the empty rooted spanning forest of G, with n elements, each of which is a rooted tree of size 1. For $i \geq 0$, if $\{\rho(T) : T \in f_i\}$ is an independent set in G, then set $f_{i+1} = f_i$. Otherwise, there exists at least one edge connecting distinct roots of trees in f_i; choose one such edge $\{\rho(T), \rho(T')\}$ uniformly at random, orient it uniformly at random as $(\rho(T), \rho(T'))$, and add it to f_i to form f_{i+1}. The unique tree in $f_{i+1} \setminus f_i$ has vertex set $v(T) \cup v(T')$ and root $\rho(T')$. Note that $f_m = f_{n-1}$ for all $m \geq n - 1$. We write $F(G) = f_{n-1}$ for the final forest built by the process, which we call the *Kingman forest* of G, and we write $(f_i, i \geq 0) \overset{d}{=} \text{KINGMAN}(G)$ for the process as a whole.

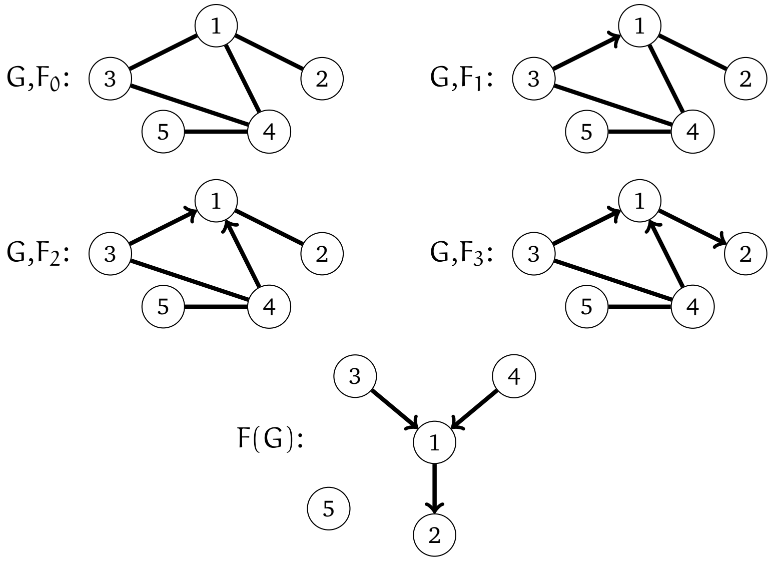

Figure 1: A realization of the Kingman coalescent on the graph G. The oriented edges are those in the forest F_k. After the third edge is added to F_k, there are no edges between the roots 2 and 5, so $F_k = F_3$ for all $k \geq 3$.

Note that if $G = K_n$ is the complete graph with n vertices, then to form f_{i+1} from f_i, a uniformly random pair of trees of f_i is chosen and merged. In this case, writing Π_i for the partition of $V(G)$ formed by the vertex sets of the trees of f_i, then $(\Pi_i, 0 \leq i \leq n-1)$ is distributed as the (discrete time) Kingman's coalescent on a set of size n. As such, the above process expands the traditional definition of Kingman's coalescent to a collection of processes in which some coalescent events may be forbidden.

In this paper, we study the structure and number of trees of $F(G)$ when the underlying graph G is an Erdős-Rényi random graph $G_{n,p}$ for $p \in (0,1)$ fixed. Throughout the paper, we let $C_{n,p}$ denote the number of trees in $F(G_{n,p})$. Our main result is the following theorem.

Theorem 1.1. *There exists a family of random variables* $(C_p : p \in (0,1))$ *such that*

(i) $C_{n,p} \xrightarrow{d} C_p$ *and* $\mathbf{E}[C_{n,p}] \to \mathbf{E}[C_p]$ *as* $n \to \infty$ *for any fixed* $p \in (0,1)$*; and*

(ii) $\frac{p}{2(1-p)} C_p \xrightarrow{\mathbb{P}} 1$ *and* $\frac{p}{2(1-p)} \mathbf{E}[C_p] \to 1$ *as* $p \to 0$.

The main tools that we use in this proof also provide information in the case when $p \to 0$ as $n \to \infty$.

Theorem 1.2. *Choose* $p = p(n)$ *such that* $p \to 0$ *and* $np \to \infty$ *as* $n \to \infty$*. Then* $\frac{(1-p)}{2p} C_{n,p} \xrightarrow{\mathbb{P}} 1$ *and* $\frac{(1-p)}{2p} \mathbf{E}[C_{n,p}] \to 1$ *as* $n \to \infty$.

Our analysis of $\textsc{Kingman}(G_{n,p})$ uses a coupling with the Kingman coalescent on the complete graph, $\textsc{Kingman}(K_n)$, which is possible by the symmetry and independence of the existence of edges in $G_{n,p}$. This coupling, along with Theorem 1.1, allows us to deduce several structural properties of the resulting Kingman forests for the case when p is fixed. We state the following theorem about the asymptotic behavior of tree sizes and heights as an illustrative example, but one could use the same connection to derive information about many other statistics.

In a rooted forest F, we define the height of a vertex v, $\text{height}(v)$, to be its graph distance from the root of its tree. We set $\text{height}(F) = \max_{v \in V(F)} \text{height}(v)$, and call this quantity the height of the forest. For a fixed $k \geq 0$ and $\alpha_1, ..., \alpha_k \in \mathbb{N}$, we say that a random vector $X = (X_1, ..., X_k)$ has a Dirichlet distribution with parameters $\alpha_1, ..., \alpha_k$, and write $X \overset{d}{=} \text{Dirichlet}(\alpha_1, ..., \alpha_k)$ if it has a density function

$$f_X(t_1, ..., t_k) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} x^{\alpha_j - 1}$$

with respect to the Lebesgue measure on the unit simplex $(t_1, ..., t_k)$, where Γ denotes the standard gamma function.

Theorem 1.3. *For fixed $p \in (0, 1)$, the following results hold.*

(i) *Let $X_n = (|T_1|, ..., |T_{C_{n,p}}|)$ be the sizes of the trees in $F(G_{n,p})$, listed in random order. Then, $(\frac{1}{n} X_n, C_{n,p}) \xrightarrow{d} (X, C_p)$ as $n \to \infty$, where C_p is the random variable from Theorem 1.1 and, conditionally given C_p, X has Dirichlet distribution with parameters $\alpha_1 = 1, \ldots, \alpha_{C_p} = 1$, i.e., X has the uniform density $f_X(t_1, ..., t_{C_p}) = \Gamma(C_p) \prod_{j=1}^{C_p} t_j^0 = \Gamma(C_p)$ on the unit simplex.*

(ii) *There exists $K > 0$ such that $|\mathbf{E}[\text{height}(F(G_{n,p}))] - e \log(n) + \frac{3}{2} \log \log(n)| \leq K$ for all n. Moreover, it holds that $\frac{\text{height}(F(G_{n,p}))}{e \log(n)} \xrightarrow{\mathbb{P}} 1$ as $n \to \infty$.*

1.2 Outline of the sections

In Section 2 we motivate the work done in this paper and cover some background information on coalescing graph processes. In Section 3 we introduce and study the edge reveal process, a coupling between the Kingman coalescent and the underlying random graph that is key to our study of the Kingman coalescent. Section 4 is dedicated to proving Theorem 1.1 and 1.2. In Section 5 we describe the aforementioned coupling with $\textsc{Kingman}(K_n)$, and prove Theorem 1.3. In Section 6 we provide proof of some bounds which are used in the proof of Theorems 1.1 and 1.2. Section 7 concludes the paper with some open questions and ideas for future research.

1.3 Notation

Before moving forward we pause to collect some notation. For $x, y \in \mathbb{R}$, we define $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$. The set $\mathbb{N}$ denotes the natural numbers with 0 excluded, and $\mathbb{Z}_{\geq 0} = \mathbb{N} \cup \{0\}$. For a set S and $k \in \mathbb{Z}_{\geq 0}$, we let $\binom{S}{k}$ denote the collection of all subsets of S of size exactly k. For $k \in \mathbb{N}$, we write $[k] := \{1, ..., k\}$. For an undirected graph $G = (V, E)$ and a vertex $v \in V$, we define $\deg_G(v) = |\{e \in E : u \in e\}|$. We write $E(G)$ to refer to the edge set of a graph G and $V(G)$ to refer to its vertex set. If we say "G is a graph on V", we mean that G is a graph with $V(G) = V$. For a graph $G = (V, E)$ and $e \in \binom{V}{2}$, we write $G + e$ for the graph $(V, E \cup \{e\})$; if $e \in E$ then $G + e = G$. A rooted tree $t = (V(t), E(t))$ is *increasing* if $V(t) \subset \mathbb{N}$ and vertex labels increase along any root-to-leaf path (equivalently, if every non-root vertex's label is strictly larger than that of its parent. Finally, $\rho(F)$ is the set of all the roots of trees in the rooted forest F.

We use $\overset{d}{=}$ to denote distributional equality, $\xrightarrow{d}$ to denote convergence in distribution, and $\xrightarrow{\mathbb{P}}$ to denote convergence in probability. For two random variables X and Y, we say that X stochastically dominates Y, writing $Y \preceq X$, if $\mathbf{P}(X \geq x) \geq \mathbf{P}(Y \geq x)$ for all $x \in \mathbb{R}$. For a finite set S, we say that $X \overset{d}{=} \text{Unif}(S)$ if for all $s \in S$, $\mathbf{P}(X = s) = |S|^{-1}$. We say that X is a geometric random variable with parameter p, and write $X \overset{d}{=} \text{Geo}(p)$, if $\mathbf{P}(X = k) = (1-p)^k p$ for $k \in \mathbb{Z}_{\geq 0}$. We say that X has a negative binomial distribution with parameters $r \in \mathbb{N}$ and $p \in (0, 1]$, and write $X \overset{d}{=} \text{N-Bin}(r, p)$, if $\mathbf{P}(X = k) = \binom{k+r-1}{k}(1-p)^k p^r$ for $k \in \mathbb{Z}_{\geq 0}$. We say that X has a hypergeometric distribution with parameters $n \in \mathbb{Z}_{\geq 0}, m \in [n], k \in [n]$, and write $X \overset{d}{=} \text{HG}(k, m, n)$, if

$$\mathbf{P}(X = j) = \frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}.$$

Throughout the article, we let $\mathcal{F}_{n,k}$ be the set of rooted forests on $[n]$ with k edges that are each given a unique label in $[k]$, such that edge labels decrease along all root-to-leaf paths. We let $\mathcal{G}_{n,k}$ denote the set of graphs on $[n]$ with k edges. For $S \subseteq [n]$, we define $\mathcal{G}_{S,k}$ to be the set of all graphs on S with k edges. We let $\mathcal{G}^{(r)}_{n,k} = \cup_{S \subseteq [n]:|S|=r} \mathcal{G}_{S,k}$ be all graphs with k edges and a vertex set of size r drawn from the set $[n]$.

2 Background and motivations

An n-*coalescent* is a stochastic process $(P_k)_{k=0}^{\infty}$ consisting of partitions of $[n] = \{1, ..., n\}$, where $P_0 = \{\{1\}, ..., \{n\}\}$ and P_{k+1} is derived from P_k by merging two distinct portions A and B with probability proportional to some function $\kappa(|A|, |B|)$. Three particularly well studied examples are $\kappa(x, y) = 1$, $\kappa(x, y) = x + y$, and $\kappa(x, y) = xy$.

These choices are referred to as Kingman's coalescent, the additive coalescent, and the multiplicative coalescent respectively [Kin82b, Pit99a, Ald97].

The study of n-coalescent processes has motivations coming from across the sciences [Ber09]. Some of the early mathematical work on coalescent models was due to Kingman [Kin82a, Kin82b], with motivation coming from the area of population genetics. Since then, coalescent processes have become part of the standard toolkit of population genetics for studying ancestral recombination graphs [CSD25, NVD25]. For a second source of inspiration one can look towards statistical physics, where coalescent processes have naturally emerged within the study of spin glasses [BS98, Pit99b, GM05].

The three coalescents mentioned above are often viewed as a sequence of forests $(F_k)_{k=0}^{\infty}$ on the vertex set $[n]$ [AB15] with F_0 being $([n], \emptyset)$. For the Kingman coalescent, the sequence exactly corresponds to $\textsc{Kingman}(K_n)$. For the additive coalescent, we sample a uniform pair (x, y) such that $y \in [n]$ and x is the root of a tree in F_k that does not contain y. Then, F_{k+1} is formed by adding the edge (x, y) to F_k, which results in x no longer being a root. The multiplicative coalescent is typically seen as a sequence of unrooted forests where F_{k+1} is derived from F_k by adding a uniform edge to F_k from among edges whose addition would not create a cycle.

Coalescent graph processes have frequently appeared in the random graph theory literature. Various versions of the Kingman coalescent [Kin82b] have been used to study recursively growing random trees via direct distributional equivalences [DR76, ABE18, BBRKK25]. The additive coalescent has appeared naturally in the study of uniform trees, as the forest valued version of the process produces a tree that is distributed uniformly over all labelled trees [Pit99a, AB15]. The multiplicative coalescent has appeared in both the study of component sizes in critical Erdös-Rényi random graphs as well as the study of minimum spanning trees [Ald97, ABBR09, ABBGM17].

These three coalescents have all been studied in depth when there are no "external" constraints, in the sense that all mergers permitted by the coalescent rule in question are permitted. However, only a small amount of work has been put towards understanding the forests that emerge when we add the restriction that all edges must come from a set of allowed edges E, i.e., when we run the coalescents on an underlying graph G. For all three coalescents, the size and structure of the forests may be greatly affected by structure of the underlying graph, and this is a primary motivation for our investigation of the structure of $\textsc{Kingman}(G_{n,p})$. There is some work on the structure of the multiplicative coalescent in non-complete geometries, due its connection with minimum spanning trees [ABBGM17, ABS21, GPS18]. Thus far, we are unaware of any research into the structure of additive coalescents on non-complete graphs.

3 The edge reveal process

When $G \overset{d}{=} G_{n,p}$, there is a useful Markov chain which couples the construction of $\text{Kingman}(G)$ to a construction of G itself. We call this coupling the *edge reveal process*; we shall use it to analyse the number of trees in $\text{Kingman}(G)$.

3.1 Definition and distributional identities

Let $B = \left\{B_e : e \in \binom{[n]}{2}\right\}$ be a collection of independent $\text{Ber}(p)$ random variables, so if $E = \left\{e \in \binom{[n]}{2} : B_e = 1\right\}$, then the graph $([n], E)$ is distributed like $G_{n,p}$. Independent of B, let $(e_k)_{k=1}^{\infty}$ be a sequence of independent $\text{Unif}\binom{[n]}{2}$ random variables. We call B the *bits* of the edge reveal process and the sequence $(e_k)_{k=0}^{\infty}$ the *queried pairs* of the edge reveal process. We set $R_0 = [n]$, $F_0 = ([n], \emptyset)$, and $G_0 = ([n], \emptyset)$. Then, for $k \geq 0$ we inductively define $(R_{k+1}, F_{k+1}, G_{k+1})$, and $L_{k+1} : E(F_{k+1}) \to \mathbb{N}$, as:

(i) If $B_{e_{k+1}} = 0$: set $(R_{k+1}, F_{k+1}, G_{k+1}) = (R_k, F_k, G_k)$.

(ii) If $B_{e_{k+1}} = 1$ and $e_{k+1} \not\subseteq R_k$: Set $(R_{k+1}, F_{k+1}, G_{k+1}) = (R_k, F_k, G_k + e_{k+1})$. (Note that e_{k+1} may already be in G_k; the graph does not change in this case).

(iii) If $B_{e_{k+1}} = 1$ and $e_{k+1} \subseteq R_k$: let $O_{k+1} = (u_{k+1}, v_{k+1})$ be a uniformly random orientation of e_{k+1}. Define $(R_{k+1}, F_{k+1}, G_{k+1}) = (R_k \backslash u_{k+1}, F_k + O_{k+1}, G_k + e_{k+1})$ and let L_{k+1} be such that $L_{k+1}(e_j) = |E(F_j)| + 1$ for all $1 \leq j \leq k+1$. (Note that with this labelling convention, the labelled forests (F_k, L_k) are elements of $\mathcal{F}_{n,|E(F_k)|}$).

We write $(R_k, F_k, G_k)_{k=0}^{\infty} \overset{d}{=} \text{ERP}(n, p)$. This sequence is infinite, though once all pairs in $\binom{[n]}{2}$ have been queried, the sequence never changes. Let $\tau_0 = 0$, and let $\tau_k = \inf\{j > \tau_{k-1} : F_j \neq F_{j-1}\}$ for all $k \geq 0$ be the times when updates of type (iii) happen (note that these times can be infinite). The snapshots of the sequence (R_k, F_k, G_k) at the times $\tau_0, \ldots, \tau_{n-1}$ are the main points of interest. We call $\tau_0, \ldots, \tau_{n-1}$ the *coalescing times* of $\text{ERP}(n, p)$.

Lemma 3.1. *Let* $(R_k, F_k, G_k)_{k=0}^{\infty} \overset{d}{=} \text{ERP}(n, p)$ *and set* $F^*_k = F_{\tau_k \wedge \tau_{k^*}}$ *for all* $0 \leq k \leq \infty$. *Then,* $(F^*_k)_{k=0}^{\infty} \overset{d}{=} \text{Kingman}(G_{n,p})$.

Proof. The forest $F^*_{\tau_{j+1} \wedge \tau_{k^*}}$ is obtained from $F^*_{\tau_j \wedge \tau_{k^*}}$ by adding a single new edge that is sampled uniformly from the set

$$\left\{e \in \binom{[n]}{2} : e \subseteq \rho(F_{\tau_{j-1}}),\ B_e = 1\right\},$$

with uniformly random orientation. This is exactly the rule for how the edges are added for the Kingman coalescent. Since each bit B_e is $\text{Ber}(p)$ distributed, we have $\left([n], \{e \in \binom{[n]}{2} : B_e = 1\}\right) \overset{d}{=} G_{n,p}$. Thus, $(F^*_k)_{k=0}^{\infty} \overset{d}{=} \text{Kingman}(G_{n,p})$. □

For the next lemma we introduce the *complement process* of $\mathrm{ERP}(n, p)$, which is the sequence $(G^*_k)_{k=0}^{\infty}$ given by $G^*_k = (\rho(F_k), \{e_1, ..., e_k\} \cap \binom{\rho(F_k)}{2} \setminus E(G_k))$. The vertex set of G^*_k is the set of roots of F_k; its edges are exactly the pairs of roots that have been queried by time k and are *not* in G_k. That is, all the edges in G^*_k that have been "verified" to not be in the underlying graph by time k. We often refer to edges that have $B_e = 0$ (and thus, all edges in G^*_k for any $k \geq 0$) as the *non-edges* of the edge reveal process, and we see updates of type (i) as revealing a non-edge.

In Section 4, we use the sequence $(G^*_k)_{k=0}^{\infty}$ to study the number of trees in a Kingman forest of $G_{n,p}$. A very useful property for this purpose is that F_k and G^*_k are each uniformly random conditional on their number of edges. In particular, this implies that their structure is uniform at coalescing times.

Lemma 3.2. *Let* $(R_k, F_k, G_k)_{k=0}^{\infty} \overset{d}{=} \mathrm{ERP}(n, p)$. *Let*

$$\mathcal{S}(m, \ell) = \left\{(f, g) : f \in \mathcal{F}_{n,m},\ g \in \mathcal{G}_{\rho(f),\ell}\right\}.$$

Then for any $0 \leq m + \ell \leq k$ *and any* $(f, g) \in \mathcal{S}(m, \ell)$*, we have*

$$\mathbf{P}\Big(F_k = f,\ G^*_k = g \;\Big|\; |E(F_k)| = m, |E(G^*_k)| = \ell\Big) = \frac{1}{|\mathcal{S}(m, \ell)|}.$$

Thus,

(i) *for any* $f \in \mathcal{F}_{n,m}$ *with* $m \leq k$*,* $\mathbf{P}(F_k = f \mid |E(F_k)| = m) = 1/|\mathcal{F}_{n,m}|$*; and*

(ii) *for any* $g \in \mathcal{G}^{(r)}_{n,m}$ *with* $m \leq k$*,* $\mathbf{P}(G^*_k = g \mid |E(G^*_k)| = m,\ |R_k| = r) = 1/|\mathcal{G}^{(r)}_{n,m}|$.

Proof. We prove the first identity via induction on k. Instead of showing it directly, we argue that $\mathbf{P}(F_k = f,\ G^*_k = g) = \mathbf{P}(F_k = \hat{f},\ G^*_k = \hat{g})$ for any $(f, g), (\hat{f}, \hat{g}) \in \mathcal{S}(m, \ell)$ with $m, \ell \geq 0$ arbitrary. The base case is immediate as F_0 and G^*_0 are deterministic. Suppose that the identity holds for some $k \geq 0$, and let $f, g \in \mathcal{S}(m, \ell)$ for some arbitrary $m, \ell \geq 0$. Let (u, v) be the edge in f of largest label. Consider the event $\{F_{k+1} = f, G^*_{k+1} = g\}$. Exactly one of the three following (disjoint) events must occur if $\{F_{k+1} = f, G^*_{k+1} = g\}$ is to occur:

(i) $(F_k, G^*_k) = (f, g)$: In this case we have that $(F_{k+1}, G^*_{k+1}) = (f, g)$ if and only if $e_{k+1} \in E(f) \cup E(g) \cup \binom{[n]\setminus\rho(f)}{2}$.

(ii) $(F_k, G^*_k) = (f \setminus (u, v), g')$, where g' is a graph on the vertex set $\rho(f) \cup \{u\}$ and edge set $E(g) \cup S$ for $S \subseteq \{\{u, w\} : w \in \rho(f) \setminus \{u, v\}\}$: In this case we have that $(F_{k+1}, G^*_{k+1}) = (f, g)$ if and only if $e_{k+1} = \{u, v\}$, $B_{\{u,v\}} = 1$, and the orientation chosen for the edge is $O_{k+1} = (u, v)$.

(iii) $(F_k, G^*_k) = (f, g \setminus e)$ for some $e \in E(g)$: In this case we have that $(F_{k+1}, G^*_{k+1}) = (f, g)$ if and only if $e_{k+1} = e$ and $B_e = 0$.

By the definition of the edge reveal process, one can see that there are no other possibilities for (F_k, G_k^*) which allow for $(F_{k+1}, G_{k+1}^*) = (f, g)$ to occur. Denote the set of graphs g' in (ii) by $S(g)$, and note that $|V(g')| = |V(g) \cup \{u\}| = m + 1$ for all $g' \in S(g)$. Let

$$\begin{aligned}
A(f,g) &= \{(F_k, G_k^*) = (f,g)\} \bigcap \left\{ e_{k+1} \in E(f) \cup E(g) \cup \binom{[n] \setminus \rho(f)}{2} \right\}, \\
B(f,g) &= \bigcup_{g' \in S(g)} \left\{ (F_k, G_k^*) = (f \setminus (u,v), g') \right\} \cap \left\{ e_{k+1} = \{u,v\},\ B_{\{u,v\}} = 1,\ O_{k+1} = (u,v) \right\}, \\
C(f,g) &= \bigcup_{e \in E(g)} \left\{ (F_k, G_k^*) = (f, g \setminus e) \right\} \cap \{e_{k+1} = e,\ B_e = 0\}
\end{aligned}$$

be the events from (i), (ii), and (iii). Then,

$$\mathbf{P}(F_k = f,\ G_k^* = g) = \mathbf{P}(A(f,g)) + \mathbf{P}(B(f,g)) + \mathbf{P}(C(f,g)).$$

Via explicit counting, we obtain the following identities:

$$\begin{aligned}
\mathbf{P}(A(f,g)) &= \mathbf{P}(F_k = f,\ G_k^* = g)\mathbf{P}\left(e_{k+1} \in E(f) \cup E(g) \cup \binom{[n] \setminus \rho(f)}{2} \right) \\
&= \mathbf{P}(F_k = f,\ G_k^* = g) \frac{|E(f)| + |E(g)| + \binom{n - |\rho(f)|}{2}}{\binom{n}{2}}, \\
\mathbf{P}(B(f,g)) &= \sum_{g' \in S(g)} \mathbf{P}(F_k = f,\ G_k^* = g')\mathbf{P}(e_{k+1} = \{u,v\},\ B_{\{u,v\}} = 1,\ O_{k+1} = (u,v)) \\
&= \sum_{g' \in S(g)} \mathbf{P}(F_k = f,\ G_k^* = g') \cdot \frac{p}{2\binom{n}{2}}, \\
\mathbf{P}(C(f,g)) &= \sum_{e \in E(g)} \mathbf{P}(F_k = f,\ G_k^* = g \setminus e)\mathbf{P}(e_{k+1} = e,\ B_e = 0) \\
&= |E(g)| \cdot \mathbf{P}(F_k = f,\ G_k^* = g \setminus e) \cdot \frac{1-p}{\binom{n}{2}}.
\end{aligned}$$

From here, induction yields that $\mathbf{P}(A(f,g)) = \mathbf{P}(A(\hat{f}, \hat{g}))$ and $\mathbf{P}(C(f,g)) = \mathbf{P}(C(\hat{f}, \hat{g}))$ for any other pair $(\hat{f}, \hat{g}) \in S(m, \ell)$. For $B(f,g)$, to see that $\mathbf{P}(B(f,g)) = \mathbf{P}(B(\hat{f}, \hat{g}))$, we remark that there are exactly $\binom{n-m}{j}$ graphs $g' \in S(g)$ such that $|E(g')| = |E(g)| + j$ for all $0 \le j \le k - 1$. Since this does not depend on the structure of g, only on its numbers of vertices and edges, we can apply the inductive hypothesis again to get that $\mathbf{P}(B(f,g)) = \mathbf{P}(B(\hat{f}, \hat{g}))$.

The first identity in the lemma follows from the fact that $\mathbf{P}(F_k = f,\ G_k^* = g) = \mathbf{P}(F_k = f',\ G_k^* = g')$ for all $(f,g),\ (f',g') \in S(m,\ell)$. The second and third identities follow straightforwardly from the first and the fact that $\mathbf{P}((F_k, G_k^*) \in S(m,\ell)) > 0$ if $m + \ell \le k$. □

3.2 Edge counts and monotonicity in the complement process

Tracking the number of edges in the complement process of ERP(n, p), which we do in Section 4, is a key part of the analysis of the number of trees in $F(G_{n,p})$. Write $N_k = |E(G_k^*)|$. Since N_k is the number of pairs in $\binom{R_k}{2}$ that have been verified to be non-edges by step k, if $N_k = \binom{R_k}{2}$ then the edge reveal process has terminated in the sense that $F_k = F_j$ for all $j \geq k$. Let $K^* = \inf\{k \geq 0 : N_k \geq \binom{R_k}{2}\}$. We define $(M_k)_{k=0}^{n-1} := (N_{\tau_k \wedge K^*})_{k=0}^{n-1}$, which we call the *edge count walk* of the edge reveal process. By Lemma 3.1 and the discussion above, if $J^* := \inf\{j \geq 0 : M_j \geq \binom{n-j}{2}\}$, then $n - J^* + 1 \overset{d}{=} C_{n,p}$. The plus one appears because in the step where we terminate, a vertex does not get removed.

We now turn our attention to describing a single step, $M_{k+1} - M_k$, of the edge-count walk. First suppose that $\tau_k < \infty$. For all $0 \leq k \leq n-2$, let $S_k = \{\tau_k + 1, \ldots, K^* \wedge (\tau_{k+1})\}$ and let

$$X_k = \left|\left\{j \in S_k : e_j \in \binom{R_{\tau_k}}{2},\ B_{e_j} = 0,\ e_j \notin \{e_1, \ldots, e_{j-1}\}\right\}\right|,$$

The random variable X_k is precisely the number of pairs that are verified to be non-edges between times τ_k and $K^* \wedge \tau_{k+1}$, so if $|E(G_{\tau_k}^*)| < \binom{n-k}{2}$ then $K^* \wedge \tau_{k+1} > \tau_k$ and $X_k = |E(G_{K^* \wedge \tau_{k+1}^* - 1})| - |E(G_{\tau_k}^*)|$. On the other hand, if $|E(G_{\tau_k}^*)| = \binom{n-k}{2}$ then either $\tau_k = \infty$ or else $K^* = \tau_k$, and in either case $X_k = 0$. Moreover, provided that $|E(G_{\tau_k}^*)| < \binom{n-k}{2}$, by the definition of the process, the number of pairs that are verified to be non-edges between times τ_k and $K^* \wedge \tau_{k+1}$ is distributed as a Geo(p) random variable truncated at $\binom{n-k}{2} - |E(G_{\tau_k}^*)|$.

Let $Y_k = \deg_{G^*_{(\tau_{k+1}-1)}}(u_{k+1}) \mathbf{1}_{\{\tau_{k+1} < \infty\}}$, where u_{k+1} is the tail of the oriented edge O_{k+1}. When $\tau_{k+1} = \infty$, Y_k is defined to be zero, and so the fact that the orientation O_{k+1} doesn't exist is not a problem for the definition. When $\tau_k = \infty$ we set $X_k = Y_k = 0$. By the definition of the edge reveal process we have that $M_{k+1} = M_k + X_k - Y_k$ for all $0 \leq k \leq n-2$.

Lemma 3.3. *For all $0 \leq k \leq n-2$ we have the following:*

(i) *Conditionally given M_k, X_k is a* Geo(p) *random variable truncated at $\binom{n-k}{2} - M_k$.*

(ii) *Conditionally given M_k and X_k, Y_k is distributed like*

$$\mathrm{HG}\left(n-k-2, M_k + X_k, \binom{n-k}{2} - 1\right) \mathbf{1}_{\{M_k + X_k \leq \binom{n-k}{2} - 1\}}. \tag{1}$$

Proof. The first identity was verified prior to the proof, so we only need to prove the second one. For this, note that that $\tau_{k+1} = \infty$ if and only if $X_k \geq \binom{n-k}{2} - M_k$ and that no vertex is removed from the complement process after time τ_k if $\tau_{k+1} = \infty$.

Since $Y_k = \deg_{G^*_{(\tau_{k+1}-1)}}(u_{k+1})\mathbf{1}_{\{\tau_{k+1}<\infty\}}$, this fact explains the indicator in (1). When $X_k < \binom{n-k}{2} - M_k$, it holds that $\tau_{k+1} < \infty$. Letting $e_{\tau_{k+1}} = \{u, v\}$, by Lemma 3.2, conditionally given M_k and X_k, $G^*_{(\tau_{k+1}-1)}$ is distributed uniformly over the set

$$\left\{ g \in \mathcal{G}^{(n-k)}_{n,M_k+X_k} : \{u, v\} \notin g \right\}.$$

From here, recalling the well known fact that the degree of a typical vertex in a graph drawn uniformly from $\mathcal{G}^{(r)}_{n,k}$ has a $\mathrm{HG}(k, n-1, \binom{n}{2})$ distribution justifies the first factor in (1), as we are conditioning on the edge $\{u_{k+1}, v_{k+1}\}$ to not be in $G^*_{(\tau_{k+1}-1)}$. □

Using the relationship between M_k, X_k, and Y_k that is derived in Lemma 3.3 we can show that the edge count walk is monotone with respect to n.

Lemma 3.4. *Let* $m \le n$, *and let* $(M^{(n)}_k)^{n-1}_{k=0}$ *and* $(M^{(m)}_k)^{m-1}_{k=0}$ *be the edge count walks for two edge reveal processes,* $\mathrm{ERP}(n, p)$ *and* $\mathrm{ERP}(m, p)$ *respectively. Then,* $M^{(m)}_k \preceq M^{(n)}_{k+(n-m)}$ *for all* $0 \le k \le m-1$.

In the proof of the above lemma, we use the following property about hypergeometric random variables.

Lemma 3.5. $X \overset{d}{=} \mathrm{HG}(k, m, n)$ *and* $Y \overset{d}{=} \mathrm{HG}(k, m', n)$, *where* $0 \le m \le m' \le n$. *Then,* $m - X \preceq m' - Y$.

Proof. Consider a population of n balls, with m coloured red, $m'-m$ coloured blue, and the rest coloured black. Draw k balls from the population without replacement and let X be the number of red balls drawn and Y the number of red or blue balls drawn. Then, $Y - X \le m' - m$, and so $m - X \le m' - Y$. □

Proof of Lemma 3.4. Let the coalescing times of the respective processes be $(\tau^{(n)}_k)^{n-1}_{k=0}$ and $(\tau^{(m)}_k)^{m-1}_{k=0}$. The argument proceeds via induction, with the base case being immediate from the fact that $M^{(m)}_0 = 0$ deterministically. Suppose that $M^{(m)}_k \preceq M^{(n)}_{k+(n-m)}$ for some $0 \le k \le m-2$. Let $(X^{(n)}_k, Y^{(n)}_k)^{n-1}_{k=0}$ and $(X^{(m)}_k, Y^{(m)}_k)^{m-1}_{k=0}$ be defined as before Lemma 3.3 for $\mathrm{ERP}(n, p)$ and $\mathrm{ERP}(m, p)$ respectively.

First suppose that $\tau^{(n)}_{k+(n-m)} = \infty$. In this case, we have $M^{(n)}_{k+1+(n-m)} = M^{(n)}_{k+(n-m)} \ge \binom{m-k}{2}$. If $\tau^{(m)}_k = \infty$, then we can, by induction, take a coupling such that $M^{(m)}_{k+1} = M^{(m)}_k \le M^{(n)}_{k+(n-m)} = M^{(n)}_{k+1+(n-m)}$. If $\tau^{(m)}_k < \infty$, then we necessarily have $M^{(m)}_{k+1} \le \binom{m-k}{2} \le M^{(n)}_{k+1+(n-m)}$.

Now assume $\tau^{(n)}_{k+(n-m)} < \infty$ (and hence, $\tau^{(m)}_k < \infty$ as well). Recall from Lemma 3.3 that, under this conditioning, $X^{(m)}_k$ is distributed like $G \wedge (\binom{m-k}{2} - M^{(m)}_k)$ and

$X^{(n)}_{k+(n-m)}$ like $G \wedge (\binom{m-k}{2} - M^{(n)}_{k+(n-m)})$ for $G \overset{d}{=} \mathrm{Geo}(p)$ sampled independently of $M^{(m)}_k$ and $M^{(n)}_{k+(n-m)}$. Take some coupling where $M^{(m)}_k \leq M^{(n)}_{k+(n-m)}$ and let $G \overset{d}{=} \mathrm{Geo}(p)$ be independent. Then we have

$$\begin{aligned}
&M^{(n)}_{k+(n-m)} + X^{(n)}_{k+(n-m)} \\
=&M^{(m)}_k + (M^{(n)}_{k+(n-m)} - M^{(m)}_k) + G \wedge \left(\binom{m-k}{2} - M^{(m)}_k - (M^{(n)}_{k+(n-m)} - M^{(m)}_k) \right) \\
\geq&M^{(m)}_k + (M^{(n)}_{k+(n-m)} - M^{(m)}_k) + G \wedge \left(\binom{m-k}{2} - M^{(m)}_k \right) - (M^{(n)}_{k+(n-m)} - M^{(m)}_k) \\
=&M^{(m)}_k + X^{(m)}_k,
\end{aligned}$$

proving that $M^{(m)}_k + X^{(m)}_k \preceq M^{(n)}_{k+(n-m)} + X^{(n)}_{k+(n-m)}$.

If $\tau^{(m)}_{k+1} = \infty$, then $\tau^{(n)}_{k+1+(n-m)} = \infty$ as well and so $Y^{(m)}_k = Y^{(n)}_{k+(n-m)} = 0$, which would complete the proof. If $\tau^{(n)}_{k+1+(n-m)} = \infty$ and $\tau^{(m)}_{k+1} < \infty$, then we can similarly finish the proof immediately as $Y^{(m)}_k \geq Y^{(n)}_{k+(n-m)} = 0$ in this case. Hence, we can suppose that $\tau^{(m)}_{k+1}, \tau^{(n)}_{k+1+(n-m)} < \infty$. In this case, conditionally given $M^{(m)}_k$, $M^{(n)}_{k+(n-m)}$, $X^{(m)}_k$, and $X^{(n)}_{k+(n-m)}$, we have that $Y^{(n)}_{k+(n-m)} \overset{d}{=} \mathrm{HG}(m-k-2, M^{(n)}_{k+(n-m)} + X^{(n)}_{k+(n-m)}, \binom{m-k}{2})$ and $Y^{(n)}_k \overset{d}{=} \mathrm{HG}(m-k-2, M^{(m)}_k + X^{(m)}_k, \binom{m-k}{2})$. Since $M^{(n)}_{k+1+(n-m)} = (M^{(n)}_{k+(n-m)} + X^{(n)}_{k+(n-m)}) - Y^{(n)}_{k+(n-m)}$ and $M^{(m)}_{k+1} = (M^{(m)}_k + X^{(m)}_k) - Y^{(m)}_k$, the result then follows from Lemma 3.5. □

Recall the fact, stated in discussion at the start of this subsection, that $C_{n,p} \overset{d}{=} n - J^* + 1$, where $J^* := \inf\{j \geq 0 : M_j \geq \binom{n-j}{2}\}$. Combining these definitions with the above lemma directly implies an important monotonicity result for the Kingman coalescent on $G_{n,p}$.

Corollary 3.6. *Let $n \geq m \geq 0$. Then, $C_{m,p} \preceq C_{n,p}$.*

4 The number of trees in $F(G_{n,p})$

In this section, we prove Theorems 1.1 and 1.2. By definition, obtaining results on the number of trees is essentially equivalent to obtaining results on J^*, which requires some bounds on $(M_k)_{k=0}^{n-1}$. We postpone the proofs until Section 6, though we record the results now. For the rest of the paper, we introduce the notation $K^-_{p,\epsilon} = \lfloor \frac{2(1-\epsilon)(1-p)}{p} \rfloor$ and $K^+_{p,\epsilon} = \lceil \frac{2(1+\epsilon)(1-p)}{p} \rceil$.

Lemma 4.1. *For any $\eta, \epsilon, \delta \in (0,1)$, there exist $C, L, c > 0$ such that the following hold:*

(i) *Fix* $p \in (0, \eta)$ *and integers* n *and* ℓ *such that* $n \geq \ell \geq L \vee K^+_{p,\epsilon}$. *Then,*

$$\mathbf{P}\left(\bigcup_{k=1}^{n-\ell}\left\{M_k \geq (1+\epsilon)\frac{(1-p)(n-k)}{p}\right\}\right) \leq Ce^{-c\ell}.$$

(ii) *For any* $n \geq 0$ *and any* $p \in (0, \eta)$ *such that* $K^-_{p,\epsilon} \geq L$, *we have,*

$$\mathbf{P}\left(M_{n-K^-_{p,\epsilon}} \leq \binom{K^-_{p,\epsilon}}{2}\right) \leq \delta + \frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right).$$

A brief computation shows that $\frac{(1+\epsilon)(1-p)(n-k)}{p}$ crosses above $\binom{n-k}{2}$ around the value k for which $n - k = K^+_{p,\epsilon}$. Thus, Lemma 4.1 tells us that J^* is likely to be between $K^-_{p,\epsilon}$ and $K^+_{p,\epsilon}$ for n large. Using Lemma 4.1 we can make this intuition into a quantitative result.

Lemma 4.2. *Let* $p = p(n) < 1$ *be such that* $np \to \infty$ *as* $n \to \infty$ *and* $\limsup_{n\to\infty} p(n) < 1$. *For any* $\delta, \epsilon \in (0,1)$, *there exists* $L \geq 0$ *such that, if* $\liminf_{n\to\infty} K^-_{p,\epsilon} \geq L$, *then*

$$\limsup_{n\to\infty}\left|\frac{p\mathbf{E}[C_{n,p}]}{2(1-p)} - 1\right| \leq \delta,$$

and

$$\limsup_{n\to\infty}\mathbf{P}\left(\left|C_{n,p} - \frac{2(1-p)}{p}\right| \geq \frac{2\epsilon(1-p)}{p}\right) \leq \delta.$$

Proof. Since $C_{n,p} \overset{d}{=} n - J^* + 1$, we have

$$\mathbf{E}[C_{n,p} - 1] = \sum_{k=0}^{n-1}\mathbf{P}(C_{n,p} \geq k+1) = \sum_{k=0}^{n-1}\mathbf{P}(J^* \leq n-k). \tag{2}$$

Using the definition of J^*, we obtain

$$\mathbf{E}[C_{n,p} - 1] \geq \sum_{k=1}^{K^-_{p,\epsilon}}\mathbf{P}\left(\bigcup_{j=0}^{n-k}\left\{M_j \geq \binom{n-j}{2}\right\}\right) \geq K^-_{p,\epsilon}\mathbf{P}\left(M_{n-K^-_{p,\epsilon}} \geq \binom{K^-_{p,\epsilon}}{2}\right). \tag{3}$$

By Lemma 4.1 (ii), there exists $L_1 \geq 0$ such that for any n and p satisfying $K^-_{p,\epsilon} \geq L_1$ we have

$$\mathbf{P}\left(M_{n-K^-_{p,\epsilon}} \geq \binom{K^-_{p,\epsilon}}{2}\right) \geq 1 - \delta - \frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right). \tag{4}$$

Note that, by our assumptions on n and p, it holds that $\lim_{n\to\infty} K^-_{p,\epsilon}n^{-1} = 0$. Using this fact along with (3) gives

$$\liminf_{n\to\infty} \frac{\mathbf{E}[C_{n,p}-1]}{K^-_{p,\epsilon}} \geq \liminf_{n\to\infty} \left(1-\delta-\frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right)\right) = (1-\delta),$$

whenever $\liminf_{n\to\infty} K^-_{p,\epsilon} \geq L_1$. By increasing L_1 if needed, using the definition of $K^-_{p,\epsilon}$, we obtain the bound,

$$\liminf_{n\to\infty} \frac{2(1-p)\mathbf{E}[C_{n,p}]}{p} \geq (1-\delta)^2(1-\epsilon). \tag{5}$$

On the other hand, from (2) and the definition of $C_{n,p}$, we also have

$$\mathbf{E}[C_{n,p}-1] \leq K^+_{p,\epsilon} + \sum_{k=K^+_{p,\epsilon}+1}^{n-1} \mathbf{P}(C_{n,p} \geq k+1) = K^+_{p,\epsilon} + \sum_{k=K^+_{p,\epsilon}+1}^{n-1} \mathbf{P}\left(M_{n-k} \geq \binom{k}{2}\right).$$

Since, for $k \geq K^+_{p,\epsilon}+1$ we have

$$\binom{k}{2} \geq \frac{1}{2}kK^+_{p,\epsilon} = \frac{(1+\epsilon)(1-p)k}{p}, \tag{6}$$

we can apply Lemma 4.1 (i) to obtain $c, C, L_2 \geq 0$ such that

$$\mathbf{E}[C_{n,p}-1] \leq K^+_{p,\epsilon} + C\sum_{k=K^+_{p,\epsilon}+1}^{n} e^{-ck}$$

for $K^+_{p,\epsilon} \geq L_2$. Evaluating the sum on the right side we obtain an $L_3 > 0$ such that, when $K^+_{p,\epsilon} \geq L_3$,

$$C\sum_{k=K^+_{p,\epsilon}+1}^{n} e^{-ck} \leq \delta K^+_{p,\epsilon}.$$

Hence, for $K^+_{p,\epsilon} \geq L_2 \vee L_3$ we have

$$\limsup_{n\to\infty} \frac{\mathbf{E}[C_{n,p}-1]}{K^+_{p,\epsilon}} \leq (1+\delta).$$

As before, we may increase L_2 or L_3 in order to obtain the bound,

$$\limsup_{n\to\infty} \frac{2(1-p)\mathbf{E}[C_{n,p}]}{p} \leq (1+\delta)^2(1+\epsilon). \tag{7}$$

Given our freedom over the choice of δ and ϵ, the first result of Lemma 4.2 follows straightforwardly from combining (5) with (7).

The second result follows from a similar approach. Recalling (6), we have

$$\limsup_{n\to\infty} \mathbf{P}\left(C_{n,p} > K^+_{p,\epsilon} + 1\right) \le \limsup_{n\to\infty} \mathbf{P}\left(\bigcup_{k=K^+_{p,\epsilon}+1}^{n} \left\{ M_{n-k} \ge \frac{(1+\epsilon)(1-p)k}{p} \right\} \right).$$

Then, by Lemma 4.1 (i), we obtain $c', C', L_4 > 0$ such that, for $n \ge \ell \ge L_4$,

$$\limsup_{n\to\infty} \mathbf{P}\left(\bigcup_{k=K^+_{p,\epsilon}+1}^{n} \left\{ M_{n-k} \ge \frac{(1+\epsilon)(1-p)k}{p} \right\} \right) \le C' e^{-c' K^+_{p,\epsilon}}.$$

Since $K^+_{p,\epsilon} \ge K^-_{p,\epsilon}$, there exists $L_5 > 0$ such that when $K^-_{p,\epsilon} \ge L_5$ we have $C' e^{-c' K^+_{p,\epsilon}} \le \delta$, which establishes the claimed upper bound. For the lower bound we use (4) along with the aforementioned fact that $\lim_{n\to\infty} K^-_{p,\epsilon} n^{-1} = 0$ to get the existence of a constant $L_1 \ge 0$ such that, when $\liminf_{n\to\infty} K^-_{p,\epsilon} \ge L_1$,

$$\begin{aligned} \limsup_{n\to\infty} \mathbf{P}(C_{n,p} \le K^-_{p,\epsilon} + 1) &\le \limsup_{n\to\infty} \mathbf{P}\left(M_{n-K^-_{p,\epsilon}} \le \binom{K^-_{p,\epsilon}}{2} \right) \\ &\le \limsup_{n\to\infty} \left(\delta + \frac{1}{2\epsilon} \left(\frac{K^+_{p,\epsilon} - 1}{n-1} \right) \right) = \delta. \end{aligned}$$

Taking $K^-_{p,\epsilon} \ge L_1 \vee L_4 \vee L_5$ we obtain the desired result. □

Equipped with Lemma 4.2, Theorems 1.1 and 1.2 follow without too much extra effort. We restate Theorem 1.1 for reference:

Theorem. *There exists a family of random variables* $(C_p : p \in (0,1))$ *such that*

(i) $C_{n,p} \xrightarrow{d} C_p$ *and* $\mathbf{E}[C_{n,p}] \to \mathbf{E}[C_p]$ *as* $n \to \infty$ *for any fixed* $p \in (0,1)$*; and*

(ii) $\frac{p}{2(1-p)} C_p \xrightarrow{\mathbb{P}} 1$ *and* $\frac{p}{2(1-p)} \mathbf{E}[C_p] \to 1$ *as* $p \to 0$.

Proof of Theorem 1.1. We only prove (i), as (ii) can be easily proven by simply combining (i) with the previous lemma. By applying the second result in Lemma 4.2 we see that, for fixed p, the sequence $(C_{n,p})_{n=0}^{\infty}$ is a tight family of random variables. By Prokhorov's Theorem, there is some subsequence $(C_{n_k,p})_{k=0}^{\infty}$ and some random variable C_p on $\mathbb{N}$ such that $C_{n_k,p} \xrightarrow{d} C_p$ as $k \to \infty$ [Bil13]. The monotonicity from Corollary 3.6 combined with the subsequential convergence implies that $C_{n,p} \xrightarrow{d} C_p$ as $n \to \infty$.

Since the sequence $(C_{n,p})_{n=1}^{\infty}$ is such that $C_{m,p} \preceq C_{n,p}$ for all $0 \le m \le n$ by Corollary 3.6, it follows from the monotone convergence theorem that $\mathbb{E}[C_{n,p}] \to \mathbb{E}[C_p]$ as $n \to \infty$. To see this, for each $n \ge 0$ and $k \ge 1$, let $q_{n,k} = \mathbf{P}(C_{n,p} \ge k)$ and set $q_k = \mathbf{P}(C_p \geq k)$. Let $U \overset{d}{=} \mathrm{Unif}[0,1]$ and define random variables $(X_n)_{n=0}^{\infty}$ and X as follows:

$$X_n = \sum_{k=1}^{\infty} k\mathbf{1}_{\{q_{n,k} \leq U < q_{n,k+1}\}},\ X = \sum_{k=1}^{\infty} k\mathbf{1}_{\{q_k \leq U < q_{k+1}\}}.$$

Note that, by definition, $X_n \overset{d}{=} C_{n,p}$ for all $n \geq 0$ and that $X \overset{d}{=} C_p$. By Corollary 3.6 it holds that $X_m(\omega) \leq X_n(\omega)$ for all $0 \leq m \leq n$. By the fact that $\mathbf{P}(C_{n,p} \geq k) \to \mathbf{P}(C_p \geq k)$, we have that $X_n(\omega) \leq X(\omega)$ for all $n \geq 0$. Moreover, $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$ since

$$\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right\} \subseteq \{q_1, q_2, \ldots\}.$$

Now apply the monotone convergence theorem to conclude. □

Since $K_{p,\epsilon}^{-} \to \infty$ as $n \to \infty$ whenever $p \to 0$ as $n \to \infty$, Theorem 1.2 follows directly from Lemma 4.2.

5 Structural properties of $F(G_{n,p})$

Many statistics of the trees in a Kingman forest of a $G_{n,p}$ can be determined by using the useful connection between $\textsc{Kingman}(K_n)$ and uniform random recursive trees, which we briefly introduce now. The *uniform random recursive tree process* is an infinite sequence of random rooted trees $(T_n)_{n=1}^{\infty}$, where T_1 consists of a root labelled 1, and T_{n+1} is derived from T_n by attaching a vertex labelled $n+1$ to a uniform vertex from T_n. The tree T_n is called a uniform random recursive tree of size n. Much is known about the structural properties of T_n as $n \to \infty$, including statistics like the height, max degree, and profile [Dev87, Pit94, DL95, DF99, GS02, Jan05, FHN06, Zha15]. It turns out [Dev87, ABE18, Esl22] that a Kingman forest of K_n, upon relabelling the vertices and edges in a way that we describe later, is distributed like T_n. Since labellings do not affect the structure of the trees, this connection can be leveraged to deduce information about label–independent properties of either model by studying the other.

The *uniform random recursive forest process with* k *trees* is a sequence of forests $(F_{n,k})_{n=k}^{\infty}$ defined recursively. First, $F_{k,k}$ is a graph with k roots labelled $1, \ldots, k$ and no edges. $F_{n+1,k}$ is derived from $F_{n,k}$ by adding an directed edge from a new vertex with the label $n+1$ to a uniformly chosen vertex in $F_{n,k}$. We write $(F_{n,k})_{n=k}^{\infty} \overset{d}{=} \mathrm{URRF}_k$ and $F_{n,k} \overset{d}{=} \mathrm{URRF}_k(n)$.

Let $f \in \mathcal{F}_{n,n-k}$. Suppose that its roots are $x_1 \leq \ldots \leq x_k$. For all $i \in [n]\backslash\{x_1, \ldots, x_k\}$, let $\ell_f(i)$ be the label of the unique edge that has i as its tail. We define a new random labelling of the vertices $L_f : [n] \to [n]$ as follows:

$$L_f(i) = \begin{cases} j, \text{ if } i = x_j \\ n - \ell_f(i) + 1, \text{ if } i \notin \{x_1, \ldots, x_k\} \end{cases}.$$

Let $\Phi(f)$ be forest that is obtained from f by removing the edge labellings, and relabelling the vertices by L_f.

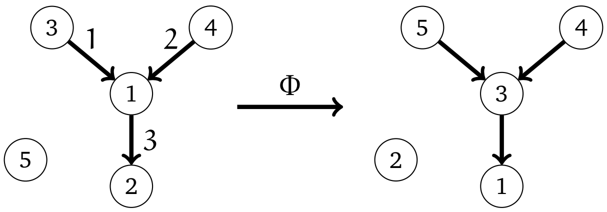

Figure 2: A forest in $\mathcal{F}_{5,3}$ and its image under Φ.

Lemma 5.1. *Let* $F \overset{d}{=} \text{Unif}(\mathcal{F}_{n,n-k})$. *Then,* $\Phi(F) \overset{d}{=} \text{URRF}_k(n)$.

Proof. Let $\mathcal{R}_{n,k}$ be the set of all forests on n vertices, with k trees, whose trees are increasing. Simple inductive arguments on k show that $|\mathcal{R}_{n,k}| = \frac{(n-1)!}{(k-1)!}$, that $F_{n,k} \overset{d}{=} \text{Unif}(\mathcal{R}_{n,k})$, and that $|\mathcal{F}_{n,n-k}| = \frac{n!(n-1)!}{k!(k-1)!}$. From these observations we conclude that, to show the desired result, it suffices to show that $\Phi : \mathcal{F}_{n,n-k} \to \mathcal{R}_{n,k}$ is an $\frac{n!}{k!}$ to 1 surjection.

Let $f \in \mathcal{R}_{n,k}$. It is easy to see that $|\Phi^{-1}(f)| > 0$ by considering the forest obtained by labelling each edge, (u, v), of f by $n - u + 1$ and relabelling the non-root vertices arbitrarily in the set $[n] \setminus [k]$. Then, observe that applying two distinct permutations σ, τ to the vertex labels of any particular $f \in \Phi^{-1}(f_2)$ that satisfies $\sigma(x_1) \leq \ldots \leq \sigma(x_k)$ and $\tau(x_1) \leq \ldots \leq \tau(x_k)$ yields two distinct forests f_σ and f_τ that both are in the set $\Phi^{-1}(f_2)$. From this observation, we get that $|\Phi^{-1}(f_2)| \geq \frac{n!}{k!}$ (the number of permutations that satisfy the described constraint).

Next, suppose that $f_2, f_2' \in \Phi^{-1}(f)$. Let $x_1, \ldots, x_k$ and $x_1', \ldots, x_k'$ be the roots of f_2 and f_2' respectively. We define a function $\sigma : [n] \to [n]$ as follows. First, we set $\sigma(x_j) = x_j'$ for all $1 \leq j \leq k$. Then, for all $i \in [n] \setminus \{x_1, \ldots, x_k\}$, we set $\sigma(i)$ to be the unique vertex j in f_2' such that $\ell_{f_2}(i) = \ell_{f_2'}(j)$. σ is clearly a bijection and is clearly edge-label-preserving. If we can show that it is a graph isomorphism between f_2 and f_2', then it follows that $|\Phi^{-1}(f)| \leq \frac{n!}{k!}$, and so $|\Phi^{-1}(f)| = \frac{n!}{k!}$.

By the symmetry of the two forests, to show that σ is an isomorphism, it suffices to show that $(\sigma(u), \sigma(v)) \in E(f_2')$ for all $(u, v) \in E(f_2)$. By the definitions of σ and Φ, we have that $L_{f_2}(u) = L_{f_2'}(\sigma(u))$ and $L_{f_2}(v) = L_{f_2'}(\sigma(v))$. Since $\Phi(f_2) = f$, it holds that $(L_{f_2}(u), L_{f_2}(v)) \in E(f)$, and so $(L_{f_2'}(\sigma(u)), L_{f_2'}(\sigma(v))) \in E(f)$ as well. Since $L_{f_2'}$ is just a relabelling of the vertices, we conclude that $(\sigma(u), \sigma(v)) \in E(f_2')$.

□

Now fix $p \in (0, 1)$. From Lemma 3.2 (i), for $n \in \mathbb{N}$ we have that, conditional upon $C_{n,p}$, $F(G_{n,p})$ is a uniform element of $\mathcal{F}_{n,n-C_{n,p}}$. By Lemma 5.1, a uniform element of $\mathcal{F}_{n,n-C_{n,p}}$ has the same graph structure as a uniform random recursive forest with $C_{n,p}$ trees. We can use this fact to derive information about the structure of $F(G_{n,p})$. First we cover the sizes of the trees in the forest.

Theorem 5.2. *Fix* $p \in (0,1)$ *and let* $X_n = (|S_1|, ..., |S_{C_{n,p}}|)$ *be the sizes of the trees in* $F(G_{n,p})$. *Then,* $(\frac{1}{n}X_n, C_{n,p}) \xrightarrow{d} (X, C_p)$ *as* $n \to \infty$, *where* C_p *is the random variable from Theorem 1.1 and, conditional upon* C_p, X *has a Dirichlet distribution with parameters* $\alpha_1 = ... = \alpha_{C_p} = 1$.

Proof. For all $n \geq k \geq 0$, let $Y_n^{(k)} = (Y_{n,1}^{(k)}, ..., Y_{n,k}^{(k)})$ be distributed like the number of balls of each colour $1, ..., k$ in a standard Pólya urn that is initialized with one ball of each colour after $(n-k)$ balls have been added to the system. Let $Z_k = (Z_{k,1}, ..., Z_{k,k})$ be a Dirichlet random variable with parameters $\alpha_1 = \cdots = \alpha_k = 1$. By applying Lemmas 3.2 and 5.1 it holds for any $0 \leq x_1, ..., x_k \leq n$ that

$$\mathbf{P}\left(|S_1| \geq x_1, ..., |S_k| \geq x_k \mid C_{n,p} = k\right) = \mathbf{P}(Y_{n,1}^{(k)} \geq x_1, ..., Y_{n,k}^{(k)} \geq x_k). \qquad (8)$$

To see this simply note that, at any step of the uniform random recursive forest process, the conditional probability that the process adds a vertex to a given tree is exactly proportional to the size of the tree, so the vector of tree sizes in a sample from $\text{URRF}_k(n)$ is distributed as $Y_n^{(k)}$. It is a well-known result from the theory of Pólya urns (see e.g., [Pem07] Theorem 2.1 or [Mah08] Theorem 3.2) that, for any $0 \leq x_1, ..., x_k \leq 1$,

$$\mathbf{P}\left(\frac{1}{n}Y_{n,1}^{(k)} \geq x_1, ..., \frac{1}{n}Y_{n,k}^{(k)} \geq x_k\right) \to \mathbf{P}(Z_{k,1} \geq x_1, ..., Z_{k,k} \geq x_k)$$

as $n \to \infty$. By combining this convergence with Theorem 1.1 and (8) we get,

$$\mathbf{P}\left(\left\{\frac{1}{n}|S_1| \geq x_1, ..., \frac{1}{n}|S_k| \geq x_k\right\} \bigcap \{C_{n,p} = k\}\right) \to \mathbf{P}(Z_{k,1} \geq x_1, ..., Z_{k,k} \geq x_k, C_p = k)$$

as $n \to \infty$. □

We finish this section by identifying the asymptotic height of $F(G_{n,p})$.

Theorem 5.3. *The following two points hold:*

(i) *There exists a constant* $K > 0$ *such that*

$$|\mathbf{E}[\text{height}(F(G_{n,p}))] - e\log(n) + \frac{3}{2}\log\log(n)| \leq K$$

for all $n \geq 1$.

(ii) *We have,* $\frac{\text{height}(F(G_{n,p}))}{e\log(n)} \xrightarrow{\mathbb{P}} 1$ *as* $n \to \infty$.

Proof. Let $(T_m)_{m=1}^{\infty}$ be distributed as the uniform random recursive tree process. Then, it is immediate from the respective definitions that $([m], E(T_m) \setminus \binom{[m]}{2})_{m=k}^{\infty} \overset{d}{=}$ URRF_k. This fact, in combination with Lemma 3.2 (i), implies that we can generate $F(G_{n,p})$ by sampling T_n and $C_{n,p}$ independently, then deleting the first $C_{n,p}$ edges from T_n. Since the deletion of all $k-1$ edges in $\binom{[k]}{2}$ from a uniform random recursive tree can at most reduce the height by $k-1$, this coupling gives us, for all $x > 0$,

$$\mathbf{P}(\text{height}(T_n) - C_{n,p} \geq x) \leq \mathbf{P}\left(\text{height}(F(G_{n,p})) \geq x\right) \leq \mathbf{P}\left(\text{height}(T_n) \geq x\right) . \qquad (9)$$

By Corollary 1.3 of [ABF13], it holds that

$$\Lambda := \sup_{n \geq 1} \mathbf{E} \left| \text{height}(T_n) - e\log(n) + \frac{3}{2}\log\log(n) \right| < \infty.$$

Also, from Theorem 1.1, we have that $\mathbf{E}[C_{n,p}] \to \mathbf{E}[C_p] < \infty$ as $n \to \infty$. Combining these two facts with (9) we get,

$$\sup_{n\geq 1} \left| \mathbf{E}\Big[\text{height}\left(F(G_{n,p})\right)\Big] - e\log(n) + \frac{3}{2}\log\log(n) \right| \leq \sup_{n \geq 1} \mathbf{E}[C_{n,p}] + \Lambda < \infty,$$

proving the first result. For the second result, first note that, since $\Lambda < \infty$, Markov's inequality implies that $\frac{\text{height}(T_n)}{e\log(n)} \xrightarrow{\mathbb{P}} 1$ as $n \to \infty$. Then, since $\mathbf{E}[C_p] < \infty$, we have that $\frac{C_{n,p}}{e\log(n)} \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. Combining these two convergences with (9) completes the proof of the second result. □

Computing other statistics in $F(G_{n,p})$

The usefulness of the coupling from the proof of Theorem 5.3, where we generate $F(G_{n,p})$ from independently sampled $C_{n,p}$ and T_n, is not limited in its usage to only discussion of the height. For example, this fact almost implies that, for any $i \in [n]$, $\deg_{F(G_{n,p})}(i)$ can be coupled with $\deg_{T_n}(i)$ so that $|\deg_{F(G_{n,p})}(i) - \deg_{T_n}(i)| \leq 1$. If we take $i = i(n) \to \infty$ as $n \to \infty$, we even have that $\mathbf{P}(\deg_{F(G_{n,p})}(i) \neq \deg_{T_n}(i)) \to 0$ as $n \to \infty$. From this we could derive a variety of results concerning the degrees in Kingman forests of $G_{n,p}$ with almost no extra effort. More generally, one can compute almost any statistic of interest that is understood for uniform random recursive trees by leveraging the fact that $([m], E(T_m) \setminus \binom{[m]}{2})_{m=k}^{\infty} \overset{d}{=} \mathrm{URRF}_k$.

6 Quantitative results on M_k

In this section we prove the results on M_k in Lemma 4.1 that we used in the proof of our main results. We require the use of many fairly standard tail bounds for familiar collections of random variables in our analysis of M_k.

Lemma 6.1. *Let* $0 < \delta < 1$*. Let* $(X_k)_{k=1}^n$ *be an independent collection of random variables with* $X_k \overset{d}{=} \mathrm{HG}(d_k, m_k, n_k)$ *for some* $d_k, m_k \leq n_k$ *and set* $X = \sum_{k=1}^n X_k$ *and* $\mu = \mathbf{E}[X] = \sum_{k=1}^n \frac{d_k m_k}{n_k}$*. Then,*

$$\mathbf{P}\left(|X - \mu| \geq \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{3}\right).$$

Let $X \overset{d}{=} \text{N-Bin}(r, p)$*. Then,*

$$\mathbf{P}\left(X \geq \frac{(1+\delta)r(1-p)}{p}\right) \leq \exp\left(-\frac{((1-p)\delta)^2 r}{6}\right),$$

and

$$\mathbf{P}\left(X \leq (1-\delta)\frac{r(1-p)}{p}\right) \leq \exp\left(-\frac{((1-p)\delta)^2 r}{3(1-\delta(1-p))}\right).$$

Proof. The first bound follows from an extension of Hoeffding's inequality to the setting of sampling without replacement [Hoe94, Theorems 2 and 4].

The second and third inequalities follow from the close relationship between binomial and negative binomial random variables. For a N-Bin(r, p) random variable to be at least k, we need to observe at most r successes from $r + k$ independent Ber(p) trials. Hence,

$$\mathbf{P}\left(X \geq (1+\delta)\frac{r(1-p)}{p}\right) = \mathbf{P}\left(\mathrm{Bin}\left(r + (1+\delta)\mu, p\right) \leq r\right),$$

where $\mu = \frac{r(1-p)}{p}$. Using a Chernoff bound gives

$$\mathbf{P}\left(X \geq (1+\delta)\frac{r(1-p)}{p}\right) \leq \exp\left(-\frac{1}{3}\left(1 - \frac{r}{\mu^+}\right)^2 \mu^+\right),$$

where

$$\mu^+ := ((1+\delta)\mu + r)p = (1+\delta)r(1-p) + rp = (1+\delta(1-p))r.$$

From here one can simplify the expression in a straightforward way to derive the final result:

$$\begin{aligned}\exp\left(-\frac{1}{3}\left(1 - \frac{r}{\mu^+}\right)^2 \mu^+\right) &= \exp\left(-\frac{1}{3}\left((1+(1-p)\delta)r - 2r + \frac{r}{(1+(1-p)\delta)}\right)\right) \\ &= \exp\left(-\frac{1}{3}\left(\frac{((1-p)\delta)^2 r}{1+(1-p)\delta}\right)\right)\end{aligned}$$

$$\leq \exp\left(-\frac{((1-p)\delta)^2 r}{6}\right).$$

The corresponding lower bound is derived in an almost identical fashion, so we omit the proof. □

Using the bounds from Lemma 6.1 we can prove (i) in Lemma 4.1, which we restate in the following lemma.

Lemma 6.2. *For any* $\eta, \epsilon \in (0,1)$*, there exist* $C, L, c > 0$ *such that the following holds. Fix* $p \in (0,\eta)$ *and integers* n *and* ℓ *such that* $n \geq \ell \geq L \vee K^+_{p,\epsilon}$*. Then,*

$$\mathbf{P}\left(\bigcup_{k=0}^{n-\ell}\left\{M_k \geq (1+\epsilon)\frac{(1-p)(n-k)}{p}\right\}\right) \leq Ce^{-c\ell}.$$

Proof. Let

$$E = \bigcup_{k=0}^{n-\ell}\left\{M_k \geq (1+\epsilon)\frac{(1-p)(n-k)}{p}\right\}.$$

For all $0 \leq k \leq n$, set

$$I_k = \left[\left(1+\frac{\epsilon}{2}\right)\frac{(1-p)(n-k)}{p}, (1+\epsilon)\frac{(1-p)(n-k)}{p}\right].$$

For each $1 \leq k \leq n-\ell$ and $1 \leq j \leq n-\ell-k$, let $A_{k,j}$ denote the event that the following three conditions (i)-(iii) hold:

(i) $M_i \in I_i$ for $k < i < k+j$,

(ii) $M_{k+j} \geq (1+\epsilon)\frac{(1-p)(n-k-j)}{p} = \sup I_{k+j}$, and

(iii) $M_k \leq \left(1+\frac{\epsilon}{2}\right)\frac{(1-p)(n-k)}{p} = \inf I_k$.

Since $M_0 = 0$, if E occurs then there must be some $1 \leq k \leq n-\ell$ and $1 \leq j \leq n-\ell-k$ such that $A_{k,j}$ occurs. Set

$$\Delta_{k,j} := \sup I_{k+j} - \inf I_k = \frac{\epsilon(1-p)(n-k)}{2p} - \frac{(1+\epsilon)(1-p)j}{p},$$

and $T_k = \frac{\epsilon(n-k)}{4(1+\epsilon)}$. We split the bounding of $\mathbf{P}(E)$ into two cases with the union bound,

$$\mathbf{P}(E) \leq \underbrace{\sum_{k=1}^{n-\ell}\sum_{j=1}^{T_k}\mathbf{P}(A_{k,j})}_{:=(\mathrm{I})} + \underbrace{\sum_{k=1}^{n-\ell}\sum_{j=T_k+1}^{n-\ell-k}\mathbf{P}(A_{k,j})}_{:=(\mathrm{II})}. \tag{10}$$

We shall bound (I) and (II) separately, beginning with (I). One can show by a brief computation that $\Delta_{k,j} \geq \frac{(1+\epsilon)(1-p)T_k}{p}$ for $(k,j) \in \{(i_1, i_2) : 1 \leq i_1 \leq n-\ell,\ 1 \leq i_2 \leq T_k\}$. Indeed, for $1 \leq k \leq n-\ell$ and $1 \leq j \leq T_k$, the map $j \mapsto \Delta_{k,j}$ is decreasing, so $\Delta_{k,j} \geq \Delta_{k,T_k}$. Using the definitions of $\Delta_{k,j}$ and T_k, we then compute

$$\begin{aligned}\Delta_{k,j} &\geq \frac{\epsilon(1-p)(n-k)}{2p} - \frac{(1+\epsilon)(1-p)T_k}{p} \\ &= \frac{1-p}{p}\left(\frac{\epsilon(n-k)}{2} - (1+\epsilon)T_k\right) \\ &= \frac{1-p}{p}\left(\frac{\epsilon(n-k)}{2} - (1+\epsilon)\cdot\frac{\epsilon(n-k)}{4(1+\epsilon)}\right) \\ &= \frac{1-p}{p}\cdot\frac{\epsilon(n-k)}{4} = \frac{(1+\epsilon)(1-p)T_k}{p}.\end{aligned}$$

By using the characterization of $(M_k)_{k=0}^{n-2}$ given in Lemma 3.3, we have that $M_{k+j} - M_j$ is stochastically dominated by a sum of j independent Geo(p)–distributed random variables, which is N-Bin(j, p)–distributed. Using this along with the negative binomial bound from Lemma 6.1 we get that, for $1 \leq j \leq T_k$,

$$\mathbf{P}(A_{k,j}) \leq \mathbf{P}\left(\text{N-Bin}(T_k, p) \geq \frac{(1+\epsilon)(1-p)T_k}{p}\right) \leq 2\exp\left(-\frac{(1-p)^2\epsilon^2 T_k}{6}\right).$$

Since $p \leq \eta$, we may compress all of the constants into some $c_1 = c_1(\epsilon, \eta) > 0$ to get

$$\begin{aligned}(\mathrm{I}) &\leq \frac{\epsilon}{2(1+\epsilon)}\sum_{k=1}^{n-\ell}(n-k)\exp\left(-c_1(n-k)\right) \\ &\leq \frac{\epsilon}{2(1+\epsilon)}\sum_{k=\ell}^{\infty} k\exp\left(-c_1 k\right).\end{aligned}$$

Doing a routine comparison of the above sum with an integral we can obtain a second constant $c_2 = c_2(\epsilon, \eta) > 0$ such that

$$(\mathrm{I}) \leq c_2 \ell e^{-c_1\ell}. \tag{11}$$

To bound (II), we need to consider the edges that are removed during the complement process as well. Essential to proceeding computations is the following claim that bounds $\mathbf{P}(A_{k,j})$ by the probability of an event concerning sums of i.i.d. random variables.

Claim. *For* $(k,j) \in \{(i_1,i_2) : 1 \le i_1 \le n-\ell,\ T_k + 1 \le i_2 \le n-\ell-k\}$*, we have that*

$$\mathbf{P}(A_{k,j}) \le \mathbf{P}\left(X^*_{k+j-1} + \sum_{i=0}^{j-2} (X^*_{k+i} - Y^*_{k+i}) \ge \Delta_{k,j} \right), \tag{12}$$

where all the random variables $(X^*_{k+i}, Y^*_{k+i})_{0 \le i \le j-1}$ *are independent, with* $X^*_{k+i} \overset{d}{=} \mathrm{Geo}(p)$ *and*

$$Y^*_{k+i} \overset{d}{=} \mathrm{HG}\left(n-k-i-2, \left\lfloor \frac{(1+\epsilon/2)(1-p)(n-k-i)}{p} \right\rfloor, \binom{n-k-i}{2} \right).$$

Proof. Let $\mathcal{F}$ be the sigma algebra generated by the whole edge reveal process. First, we note by the definition of $A_{k,j}$ that

$$\mathbf{P}(A_{k,j}) \le \mathbf{P}\left(\left(\bigcap_{i=1}^{j-1} \{M_i \in I_i\} \right) \bigcap \left\{ X_{k+j-1} + \sum_{i=0}^{j-2} (X_{k+i} - Y_{k+i}) \ge \Delta_{k,j} \right\} \right).$$

From here, we complete the proof with a direct coupling. We define, for all $0 \le i \le j-1$ conditionally given X_{k+i} and M_{k+i},

$$X^*_{k+i} = X_{k+i} + Z_{k+i} \mathbf{1}_{\{M_{k+i} + X_{k+i} = \binom{n-k-i}{2}\}},$$

where $(Z_{k+i} : 0 \le i \le j-1)$ is a collection of independent $\mathrm{Geo}(p)$ random variables that is also independent of $\mathcal{F}$. By the memoryless property, we have that $X^*_{k+i} \overset{d}{=} \mathrm{Geo}(p)$. Recall that, when $\tau_{k+i} < \infty$, we have that $Y_{k+i} = \deg_{G^*_{\tau_{k+i}-1}}(u_{k+i})$. For all $0 \le i \le j-2$, if $\tau_{k+i} < \infty$ and $G^*_{G^*_{\tau_{k+i}-1}}$ has more than $\lfloor \frac{(1+\epsilon/2)(1-p)(n-k-i)}{p} \rfloor$ edges, let $Y^*_{k+i} = \deg_{H_{k+i}}(u_{k+i})$, where H_{k+i} is a uniformly chosen subgraph of $G^*_{G^*_{\tau_{k+i}-1}}$ with $\lfloor \frac{(1+\epsilon/2)(1-p)(n-k-i)}{p} \rfloor$ edges. Otherwise, we just set Y^*_{k+i} to be a hypergeometric random variable with our desired distribution, independent of $\mathcal{F}$. By construction, X^*_{k+i} and Y^*_{k+i} are independent, and have the correct distribution. Finally, since $X_{k+i} \le X^*_{k+i}$ for all $0 \le i \le j-1$, and since $Y_{k+i} \ge Y^*_{k+i}$ on the event that $M_{k+i} + X_{k+i} \le \binom{n-k-i}{2}$ (which is a subset of the event that $M_{k+i+1} \in I_{k+i+1}$), we have that

$$\left(\bigcap_{i=1}^{j-1} \{M_i \in I_i\} \right) \bigcap \left\{ X_{k+j-1} + \sum_{i=0}^{j-2} (X_{k+i} - Y_{k+i}) \ge \Delta_{k,j} \right\}$$

is a subset of

$$\left\{ X^*_{k+j-1} + \sum_{i=0}^{j-2} (X^*_{k+i} - Y^*_{k+i}) \ge \Delta_{k,j} \right\},$$

which is enough to complete the proof. □

With the claim proven, we can begin to work on bounding (II), by bounding the expression in the right side of (12). Set, for all $1 \leq k \leq n-\ell$ and $T_k+1 \leq j \leq n-\ell-k$,

$$\mu_{k,j} = \mathbf{E}\left[X^*_{k+j-1} + \sum_{i=0}^{j-2} X^*_{k+i}\right] = \frac{j(1-p)}{p},$$
$$\nu_{k,j} = \mathbf{E}\left[\sum_{i=0}^{j-2} Y^*_{k+i}\right] = \sum_{i=0}^{j-2}\left(1+\frac{\epsilon}{2}\right)\frac{2(1-p)(n-k-i-2)}{p(n-k-i-1)}.$$

From $(n-k-i-2)/(n-k-i-1) = 1 - \frac{1}{n-k-i-1}$ we obtain

$$\begin{aligned}\mu_{k,j} - \nu_{k,j} &= \frac{j(1-p)}{p} - \left(1+\frac{\epsilon}{2}\right)\frac{2(1-p)}{p}\sum_{i=0}^{j-2}\left(1 - \frac{1}{n-k-i-1}\right)\\ &\leq \frac{1-p}{p}\Big(-\big((1+\epsilon)j-(2+\epsilon)\big) + 4\sum_{r=n-k-j+1}^{n-k}\frac{1}{r}\Big)\\ &\leq \frac{1-p}{p}\Big(-\big((1+\epsilon)j-(2+\epsilon)\big) + 4\log(n-k)\Big),\end{aligned}$$

where in the last step we used the crude bound $\sum_{r=n-k-j+1}^{n-k} r^{-1} \leq \log(n-k)$.

Fix $\delta = \epsilon/20 \in (0,1)$. Since $\epsilon < 1$ we have $\mu_{k,j} + \nu_{k,j} \leq 5j(1-p)/p$, so

$$(1+\delta)\mu_{k,j} - (1-\delta)\nu_{k,j} \leq \frac{1-p}{p}\Big(-\big((1+\epsilon-5\delta)j - (2+\epsilon)\big) + \log(n-k)\Big).$$

Using that $j \leq n-k-\ell$ and $5\delta = \epsilon/4$, as well as the definition of $\Delta_{k,j}$, we obtain

$$\begin{aligned}(1+\delta)\mu_{k,j} - (1-\delta)\nu_{k,j} - \Delta_{k,j} &\leq \frac{1-p}{p}\left(5\delta j + 2 + \epsilon - \frac{\epsilon}{2}(n-k) + \log(n-k)\right)\\ &\leq \frac{1-p}{p}\left(3 - 5\delta\,\ell - \frac{\epsilon}{4}(n-k) + \log(n-k)\right).\end{aligned}$$

Since the linear term dominates the logarithm, there exists $L = L(\epsilon,\eta) > 0$ such that for all $s \geq L$, $-\frac{\epsilon}{4}s + \log s + 3 \leq -\frac{\epsilon}{8}s$. Hence, for all $n-k \geq L$ and all $T_k \leq j \leq n-k-\ell$,

$$(1+\delta)\mu_{k,j} - (1-\delta)\nu_{k,j} \leq \Delta_{k,j} - \frac{\epsilon}{8}\frac{1-p}{p}(n-k) - \frac{5\delta(1-p)}{p}\,\ell.$$

(Since $n-k \geq \ell$, for the above bound to hold it suffices that $\ell \geq L$ and $T_k \leq j \leq$

$n - k - \ell$.) Consequently,

$$\Big\{ \sum_{i=0}^{j-1} X_{k+i}^* \le (1+\delta)\mu_{k,j} \Big\} \bigcap \Big\{ \sum_{i=0}^{j-2} Y_{k+i}^* \ge (1-\delta)\nu_{k,j} \Big\}$$
$$\subseteq \Big\{ X_{k+j-1}^* + \sum_{i=0}^{j-2} (X_{k+i}^* - Y_{k+i}^*) < \Delta_{k,j} \Big\}.$$

Thus, by (12) and a union bound, if $\ell \ge L$ and $T_k \le j \le n - k - \ell$ then

$$\mathbf{P}(A_{k,j}) \le \mathbf{P}\left(\sum_{i=0}^{j-1} X_{k+i}^* \ge (1+\delta)\mu_{k,j} \right) + \mathbf{P}\left(\sum_{i=0}^{j-2} Y_{k+i}^* \le (1-\delta)\nu_{k,j} \right).$$

Lemma 6.1 yields a constant $c_3 = c_3(\epsilon, \eta) > 0$ such that

$$\mathbf{P}\left(\sum_{i=0}^{j-1} X_{k+i}^* \ge (1+\delta)\mu_{k,j} \right) \le \exp\left(-\frac{((1-p)\delta)^2}{6} j \right) \le \exp(-c_3\,(n-k))\,, \tag{13}$$

since $j \ge T_k = \frac{\epsilon}{4(1+\epsilon)}(n-k)$ in case (II). Also, applying the first inequality in Lemma 6.1 to the family $(Y_{k+i}^*)_{i=0}^{j-2}$, and using the crude bound

$$\nu_{k,j} = \sum_{i=0}^{j-2} \left(1 + \frac{\epsilon}{2}\right) \frac{2(1-p)(n-k-i-2)}{p(n-k-i-1)} \ge \left(1 + \frac{\epsilon}{2}\right) \frac{1-p}{p} (j-1),$$

since $n - k \ge \ell$, we obtain $c_4 = c_4(\epsilon, \eta) > 0$ such that, for all $T_k \le j \le n - k - \ell$,

$$\mathbf{P}\left(\sum_{i=0}^{j-2} Y_{k+i}^* \le (1-\delta)\nu_{k,j} \right) \le 2\exp\left(-\frac{\delta^2}{3}\,\nu_{k,j} \right) \le 2\exp(-c_4\,(n-k))\,. \tag{14}$$

Putting the pieces together. Combining (13) and (14) gives

$$\mathbf{P}(A_{k,j}) \le \exp(-c_3(n-k)) + 2\exp(-c_4(n-k))\,, \tag{15}$$

for all $\ell \ge L \vee K_{p,\epsilon}^+$ and $T_k \le j \le n - k - \ell$. Hence, summing over j and k in the contribution (II) from (10), we obtain

$$(\mathrm{II}) \le \sum_{k=1}^{n-\ell} \sum_{j=T_k}^{n-\ell-k} \mathbf{P}(A_{k,j}) \le \sum_{k=\ell}^{\infty} k\,e^{-c_3 k} + 2\sum_{k=\ell}^{\infty} k\,e^{-c_4 k}.$$

Using this together with (11) in (10), the result follows. □

Since $M_0 = 0$, proving a lower bound matching the one provided in Lemma 6.2 requires a different approach. Specifically, we are faced with the new problem of verifying that M_k ever gets within $\frac{\epsilon(1-p)(n-k)}{p}$ of the desired value of $\frac{(1-p)(n-k)}{p}$. The next two lemmas combine to show this.

Lemma 6.3. *For all* $\epsilon, \eta \in (0,1)$ *there exists* $L > 0$ *such that, for all* $p \in (0,\eta)$ *and integers* ℓ, n *with* $n \geq \ell \geq L \vee K^+_{p,\epsilon}$,

$$\mathbf{E}[M_{n-\ell}] \;\geq\; \frac{(1-\epsilon)(1-p)}{p}\,\ell\Big(1 - \frac{\ell-1}{n-1}\Big).$$

Proof. Let $\mathcal{F}_k = \sigma\big((X_j, Y_j)_{j<k}\big)$ and recall that $M_{k+1} = M_k + X_k - Y_k$. As in the exposure process,

$$\mathbf{E}[Y_k \,|\, \mathcal{F}_k, X_k] \;\leq\; \frac{2}{n-k}\,(M_k + X_k),$$

and X_k is a $\mathrm{Geo}(p)$ truncated at $B_k := \binom{n-k}{2} - M_k$, so

$$\mathbf{E}[X_k \mid \mathcal{F}_k] = \frac{1-p}{p}\,\big(1 - p^{B_k}\big).$$

Therefore, for $0 \leq k \leq n-2$,

$$\begin{aligned}\mathbf{E}[M_{k+1} \mid \mathcal{F}_k] &= M_k + \mathbf{E}[X_k - \mathbf{E}[Y_k \,|\, \mathcal{F}_k, X_k] | \,\mathcal{F}_k] \\ &\geq \Big(1 - \frac{2}{n-k}\Big)M_k + \frac{1-p}{p}\Big(1 - \frac{2}{n-k} - p^{B_k}\Big). \qquad (16)\end{aligned}$$

Choose L_1 so that $\frac{2}{n-k} \leq \epsilon/2$ whenever $n-k \geq L_1$. Next, with

$$t_k := \Big(1 + \frac{\epsilon}{2}\Big)\frac{(1-p)(n-k)}{p},$$

we have

$$\mathbf{E}[p^{B_k}] \;=\; \mathbf{E}[p^{B_k}\mathbf{1}_{\{M_k \geq t_k\}}] + \mathbf{E}[p^{B_k}\mathbf{1}_{\{M_k < t_k\}}] \;\leq\; \mathbf{P}(M_k \geq t_k) + p^{\binom{n-k}{2} - t_k}.$$

By Lemma 6.2, there exist $C, c, L_2 > 0$ such that $\mathbf{P}(M_k \geq t_k) \leq Ce^{-c(n-k)}$ for $n-k \geq L_2 \vee K^+_{p,\epsilon}$. Moreover,

$$\binom{n-k}{2} - t_k = \binom{n-k}{2} - \Big(1 + \frac{\epsilon}{2}\Big)\frac{(1-p)(n-k)}{p} \;\geq\; \frac{\epsilon(1-p)}{2p}\,(n-k)$$

whenever $n-k \geq \frac{2(1+\epsilon)(1-p)}{p} + 1$, hence $p^{B_k - t_k} \leq e^{-c'(n-k)}$ (since $p \leq \eta < 1$). Thus there exists L_3 such that

$$\mathbf{E}[p^{B_k}] \le \epsilon/2 \qquad \text{whenever } n-k \ge L_3 \vee \frac{2(1+\epsilon)(1-p)}{p}.$$

Taking expectations in (16) and using the two bounds above, we obtain for all such k,

$$\mathbf{E}[M_{k+1}] \;\ge\; \Big(1-\frac{2}{n-k}\Big)\mathbf{E}[M_k] + \frac{(1-\epsilon)(1-p)}{p}. \tag{17}$$

Set $c := \frac{(1-\epsilon)(1-p)}{p}$ and $b_k := 1-\frac{2}{n-k}$. A particular solution of the recurrence $a_{k+1} = b_k a_k + c$ is $a_k^* := c(n-k)$. Writing $d_k := \mathbf{E}[M_k] - a_k^*$, (17) gives $d_{k+1} \ge b_k d_k$. With $M_0 = 0$ we have $d_0 = -cn$, and therefore

$$\mathbf{E}[M_k] \;\ge\; c(n-k) - cn\prod_{i=0}^{k-1}\Big(1-\frac{2}{n-i}\Big).$$

For $k = n-\ell$ this product telescopes to yield

$$\prod_{i=0}^{n-\ell-1}\Big(1-\frac{2}{n-i}\Big) = \prod_{j=\ell+1}^{n}\frac{j-2}{j} = \frac{\ell(\ell-1)}{n(n-1)}.$$

Hence, for all $n \ge \ell \ge L \vee K^+_{p,\epsilon}$ with $L := L_1 \vee L_3$,

$$\mathbf{E}[M_{n-\ell}] \;\ge\; c\Big(\ell - \frac{\ell(\ell-1)}{n-1}\Big) = \frac{(1-\epsilon)(1-p)}{p}\,\ell\Big(1-\frac{\ell-1}{n-1}\Big),$$

which is the stated bound. □

Combining the previous lemma with the exponential tail bound from Lemma 6.2 we can prove that M_k will cross above $\frac{(1-\epsilon)(1-p)(n-k)}{p}$ before the edge reveal process terminates.

Lemma 6.4. *Let $\eta, \epsilon, \delta \in (0,1)$. There exists $L \ge 0$ such that, for any $n \ge 0$ and any $p \in (0,\eta)$ such that $K^+_{p,\epsilon} \ge L$, we have,*

$$\mathbf{P}\left(M_{n-K^+_{p,\epsilon}} \le \frac{(1-\epsilon)(1-p)K^+_{p,\epsilon}}{p}\right) \le \delta + \frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right).$$

In the proof, we use the following basic probability fact.

Lemma 6.5. *Let X be a random variable with $\mathbf{E}[X] \in \mathbb{R}$, and let $a, b \in \mathbb{R}$. Then,*

$$\mathbf{P}(X \le a) \le \frac{1}{b-a}\Big(b + \mathbf{E}[X : X \ge b] - \mathbf{E}[X]\Big)$$

Proof. By the definition of the expectation we have,

$$\mathbf{E}[X] \le a\mathbf{P}(X \le a) + b\mathbf{P}(a < X < b) + \mathbf{E}[X : X \ge b].$$

Upper bounding $\mathbf{P}(a < X < b)$ with $1 - \mathbf{P}(X \le a)$ and then re-arranging terms gives the desired result. □

Proof of Lemma 6.4. Let

$$a = \frac{(1-\epsilon)(1-p)K^+_{p,\epsilon}}{p}, \quad b = \frac{(1+\epsilon\delta)(1-p)K^+_{p,\epsilon}}{p}.$$

Our goal is to apply Lemma 6.5 to obtain our desired result, and towards that we need to bound $\mathbf{E}[M_{n-K^+_{p,\epsilon}}]$ and $\mathbf{E}[M_{n-K^+_{p,\epsilon}} : M_{n-K^+_{p,\epsilon}} \ge b]$. By Lemma 6.3 we have an $L_1 > 0$ such that, when p is such that $K^+_{p,\epsilon} \ge L_1$,

$$\mathbb{E}[M_{n-K^+_{p,\epsilon}}] \ge \frac{\left(1 - \frac{\epsilon\delta}{2}\right)(1-p)}{p} K^+_{p,\epsilon}\left(1 - \frac{K^+_{p,\epsilon} - 1}{n-1}\right) \tag{18}$$

By Lemma 6.2 there exists c, C, L_1 such that

$$\mathbf{P}\left(M_{n-\ell} \ge \left(1 + \frac{\epsilon\delta}{2}\right)\frac{(1-p)\ell}{p}\right) \le Ce^{-c\ell}. \tag{19}$$

whenever $\ell \ge L_2 \vee K^+_{p,\epsilon\delta/2}$. Using the fact that $M_{n-K^+_{p,\epsilon}}$ can only take values in $\{\binom{k}{2} : k \ge K^+_{p,\epsilon}\}$ on the event that $M_{n-K^+_{p,\epsilon}} \ge \binom{K^+_{p,\epsilon}}{2}$, we have,

$$\begin{aligned} &\mathbf{E}\left[M_{n-K^+_{p,\epsilon}}\mathbf{1}_{\left\{M_{n-K^+_{p,\epsilon}} \ge b\right\}}\right] \\ &\le \sum_{j=b}^{\binom{K^+_{p,\epsilon}}{2}} j\mathbf{P}\left(M_{n-K^+_{p,\epsilon}} = j\right) + \sum_{j=K^+_{p,\epsilon}+1}^{n} \binom{j}{2}\mathbf{P}\left(M_{n-K^+_{p,\epsilon}} = \binom{j}{2}\right). \end{aligned} \tag{20}$$

The event $\{M_{n-K^+_{p,\epsilon}} = \binom{j}{2}\}$ is exactly the event that the edge reveal process terminates between the $n-j$ coalescing time and $n-j-1$ coalescing time. Then, since

$$\binom{j}{2} \ge \left(1 + \frac{\epsilon\delta}{2}\right)\frac{(1-p)j}{p}$$

for all $j \ge K^+_{p,\epsilon\delta/2} + 1$, we can bound each of the terms in the second sum in (20) to obtain,

$$
\begin{aligned}
&\mathbf{E}\left[M_{n-K^+_{p,\epsilon}}\mathbf{1}_{\left\{M_{n-K^+_{p,\epsilon}}\geq b\right\}}\right]\\
\leq&\binom{K^+_{p,\epsilon}}{2}\mathbf{P}\left(M_{n-K^+_{p,\epsilon}}\geq b\right)+\sum_{j=K^+_{p,\epsilon}+1}^{n}\binom{j}{2}\mathbf{P}\left(M_{n-j}\geq\left(1+\frac{\epsilon\delta}{2}\right)\frac{(1-p)j}{p}\right)\\
\leq&\sum_{j=K^+_{p,\epsilon}}^{n}\binom{j}{2}\mathbf{P}\left(M_{n-j}\geq\left(1+\frac{\epsilon\delta}{2}\right)\frac{(1-p)j}{p}\right).
\end{aligned}
$$

Then, using (19) we get,

$$
\mathbf{E}\left[M_{n-K^+_{p,\epsilon}}\mathbf{1}_{\left\{M_{n-K^+_{p,\epsilon}}\geq b\right\}}\right]\leq C\sum_{j=K^+_{p,\epsilon}}^{\infty}j^2e^{-cj}.
$$

Since the upper bound above tends to zero as $K^+_{p,\epsilon}\to\infty$, by potentially increasing L_2 if needed we obtain that, when $K^+_{p,\epsilon}\geq L_2$,

$$
\mathbf{E}\left[M_{n-K^+_{p,\epsilon}}\mathbf{1}_{\left\{M_{n-K^+_{p,\epsilon}}\geq b\right\}}\right]\leq\frac{\epsilon\delta}{2}\frac{(1-p)K^+_{p,\epsilon}}{p}. \tag{21}
$$

Applying Lemma 6.5 with the bounds (18) and (21) we get, whenever $K^+_{p,\epsilon}\geq L_1\vee L_2$,

$$
\mathbf{P}(M_{n-K^+_{p,\epsilon}}\leq a)\leq\frac{1}{\epsilon(1+\delta)}\left((1+\epsilon\delta)-\left(1-\frac{\epsilon\delta}{2}\right)\left(1-\frac{K^+_{p,\epsilon}-1}{n-1}\right)+\frac{\epsilon\delta}{2}\right).
$$

Since $\delta,\epsilon\in(0,1)$, we can bound the expression on the right to get

$$
\mathbf{P}(M_{n-K^+_{p,\epsilon}}\leq a)\leq\delta+\frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right).
$$

□

Lemma 6.6. *Let* $\epsilon,\eta,\delta\in(0,1)$. *There exists an* $L\geq 0$ *such that, for any* $n\geq 0$ *and any* $p\in(0,\eta)$ *such that* $K^-_{p,\epsilon}\geq L$, *we have,*

$$
\mathbf{P}\left(M_{n-K^-_{p,\epsilon}}\leq\binom{K^-_{p,\epsilon}}{2}\right)\leq\delta+\frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right).
$$

Proof. By the previous lemma, we have some $L\geq 0$ such that, whenever satisfies

$K^+_{p,\epsilon} \geq L$,

$$\mathbf{P}\left(M_{n-K^+_{p,\epsilon}} \leq \frac{(1-\epsilon)(1-p)K^+_{p,\epsilon}}{p}\right) \leq \delta/2 + \frac{1}{2\epsilon}\left(\frac{K^+_{p,\epsilon}-1}{n-1}\right).$$

Now, consider the event

$$E = \left\{M_{n-K^-_{p,\epsilon}} \leq \binom{K^-_{p,\epsilon}}{2}\right\} \bigcap \left\{M_{n-K^+_{p,\epsilon}} \geq \frac{(1-\epsilon)(1-p)K^+_{p,\epsilon}}{p}\right\}.$$

To finish the proof it suffices to show that there is some $L^* \geq 0$ such that, when $K^-_{p,\epsilon} \geq L^*$, we have $\mathbf{P}(E) \leq \delta/2$.

If $M_{n-k} + X_{n-k}$ first exceeds $\binom{k}{2}$ at some step k with $K^-_{p,\epsilon} \leq k \leq K^+_{p,\epsilon}$, then the process freezes at this value, and so

$$M_{n-K^-_{p,\epsilon}} \geq \binom{K^-_{p,\epsilon}}{2},$$

implying that E cannot occur. One consequence of this is that we may replace $(X_{n-k} : K^-_{p,\epsilon} \leq k \leq K^+_{p,\epsilon})$ with a collection of independent (untruncated) geometric random variables $(X^*_{n-k} : K^-_{p,\epsilon} \leq k \leq K^+_{p,\epsilon})$ without decreasing the probability of E occurring E. Note also that $Y_k \leq n-k-2$ deterministically for all $0 \leq k \leq n-2$. Thus, defining E^* to be the event,

$$\left\{M_{n-K^+_{p,\epsilon}} + \sum_{K^-_{p,\epsilon}+1}^{K^+_{p,\epsilon}} (X^*_{n-k} - k) \leq \binom{K^-_{p,\epsilon}}{2}\right\} \bigcap \left\{M_{n-K^+_{p,\epsilon}} \geq \frac{(1-\epsilon)(1-p)K^+_{p,\epsilon}}{p}\right\},$$

we have that $\mathbf{P}(E) \leq \mathbf{P}(E^*)$. Next observe that

$$\binom{K^-_{p,\epsilon}}{2} \leq \frac{(1-\epsilon)(1-p)K^-_{p,\epsilon}}{p},$$

and so

$$E^* \subseteq \left\{\sum_{k=K^-_{p,\epsilon}+1}^{K^+_{p,\epsilon}} (X^*_k - k) \leq \frac{(1-\epsilon)\left(K^-_{p,\epsilon} - K^+_{p,\epsilon}\right)(1-p)}{p}\right\}. \tag{22}$$

A quick computation gives that

$$\sum_{k=K^-_{p,\epsilon}+1}^{K^+_{p,\epsilon}} k = \binom{K^+_{p,\epsilon}}{2} - \binom{K^-_{p,\epsilon}}{2}.$$

Moreover, for any $\beta \in (0,1)$, there exists $L(\beta) \geq 0$ such that

$$\left| \frac{(1-\epsilon)K^+_{p,\epsilon}(1-p)}{p\binom{K^+_{p,\epsilon}}{2}} - 1 \right| \leq \beta \text{ and } \left| \frac{(1-\epsilon)K^-_{p,\epsilon}(1-p)}{p\binom{K^-_{p,\epsilon}}{2}} - 1 \right| \leq \beta, \tag{23}$$

when $K^-_{p,\epsilon} \geq L_2(\beta)$. Hence, for $K^-_{p,\epsilon} \geq L(1/2)$, we have

$$\sum_{k=K^-_{p,\epsilon}+1}^{K^+_{p,\epsilon}} k + \frac{\left((1-\epsilon)K^-_{p,\epsilon} - (1-\epsilon)K^+_{p,\epsilon}\right)(1-p)}{p} \leq \frac{1}{2}\left(\binom{K^+_{p,\epsilon}}{2} - \binom{K^-_{p,\epsilon}}{2}\right). \tag{24}$$

Moreover, (23) also gives, for $K^-_{p,\epsilon} \geq L(3/4)$,

$$\mathbf{E}\left[\sum_{k=K^-_{p,\epsilon}+1}^{K^+_{p,\epsilon}} X^*_k\right] \geq \frac{3}{4}\left(\frac{1}{1-\epsilon}\binom{K^+_{p,\epsilon}}{2} - \frac{1}{1-\epsilon}\binom{K^-_{p,\epsilon}}{2}\right) \geq \frac{3}{4}\left(\binom{K^+_{p,\epsilon}}{2} - \binom{K^-_{p,\epsilon}}{2}\right). \tag{25}$$

Using (22) and (24) we have

$$\mathbf{P}(E) \leq \mathbf{P}\left(\sum_{k=K^-_{p,\epsilon}+1}^{K^+_{p,\epsilon}} X^*_k \leq \frac{1}{2}\left(\binom{K^+_{p,\epsilon}}{2} - \binom{K^-_{p,\epsilon}}{2}\right)\right).$$

Finally, applying the negative binomial bounds in Lemma 6.1 along with (25) to the right side of the above inequality we obtain a constant $C(\eta) > 0$ such that, for $K^-_{p,\epsilon} \geq L(1/2) \vee L(3/4)$,

$$\mathbf{P}(E) \leq \exp\left(-C(\eta)\left(\binom{K^+_{p,\epsilon}}{2} - \binom{K^-_{p,\epsilon}}{2}\right)\right). \tag{26}$$

By making $K^-_{p,\epsilon}$ large, the upper bound in (26) can be made arbitrarily small, and so less than $\delta/2$, which completes the proof. □

7 Questions and future research

The work here answers some of the basic questions one may have about the Kingman coalescent on $G_{n,p}$, however, there are still many questions worthy of study. One direction for future work could be to focus on the Kingman process on general graphs. We have two quite concrete questions concerning this topic.

(i) Does there exist a sequence of graphs $(G_n)_{n=1}^{\infty}$, with G_n a graph on n vertices, and a function $f : \mathbb{N} \to \mathbb{N}$ with $f(n)/\log(n) \to \infty$ as $n \to \infty$ such that, with probability tending to 1 as $n \to \infty$, $\text{height}(F(G_n))) \geq f(n)$?

(ii) Let $G = ([n], E)$ be a graph, and let $H = ([n], E')$ be such that $E' \subseteq E$. Does the number of components in $F(H)$ stochastically dominate $F(G)$? Does $\mathrm{height}(F(G))$ stochastically dominate $\mathrm{height}(F(H))$?

We note that answering (ii) in the affirmative would also answer (i) in the negative, as it would imply that $\mathrm{height}(F(K_n))$ is the largest possible height that can be attained from the Kingman process on any graph on the vertex set $[n]$.

A second direction could be to study the Kingman coalescent on other random graphs. A crucial part of our analysis here was the fact that vertex degrees in $G_{n,p}$ are all identically distributed and exchangeable. What if our underlying graph did not have this property? There are many examples of inhomogeneous random graphs one could choose, but for a concrete example, consider a random graph G_n with some fixed degree sequence $(d_1, ..., d_n)$. What can be said about the structure of $\textsc{Kingman}(G_n)$? When the degree sequence is not regular (i.e., we do not have $d_1 = \ldots = d_n$), how do the statistics of the height of a given vertex in $F(G_n)$ depend on its degree. We expect that high degree vertices are typically found at larger heights than low degree vertices.

It would also be interesting to study other coalescent rules on random graphs. As was mentioned in Section 2, Kingman's coalescent is not the only commonly studied coalescent process. Deriving results similar to those of this article when we consider a generalization of the additive coalescent rather than Kingman's would be interesting. Moreover, the inhomogeneous random graph case should be just as interesting for the additive coalescent as it is for Kingman's coalescent. Some results concerning the multiplicative coalescent on $G_{n,p}$ are already known, but there is much yet to be done [ABBGM17, ABS21].

Acknowledgements

CA acknowledges the support of the NSERC Postgraduate Scholarship-Doctoral. LAB acknowledges the support of the NSERC Discovery Grants program and the Canada Research Chairs program.

References

[AB15] Louigi Addario-Berry. Partition functions of discrete coalescents: from Cayley's formula to Frieze's ζ (3) limit theorem. In *XI Symposium on Probability and Stochastic Processes: CIMAT, Mexico, November 18-22, 2013*, pages 1–45. Springer, 2015.

[ABBGM17] Louigi Addario-Berry, Nicolas Broutin, Christina Goldschmidt, and Grégory Miermont. The scaling limit of the minimum spanning tree of the complete graph. *The Annals of Probability*, 45(5):3075 – 3144, 2017.

[ABBR09] Louigi Addario-Berry, Nicolas Broutin, and Bruce Reed. Critical random graphs and the structure of a minimum spanning tree. *Random Structures & Algorithms*, 35(3):323–347, 2009.

[ABE18] Louigi Addario-Berry and Laura Eslava. High degrees in random recursive trees. *Random Structures & Algorithms*, 52(4):560–575, 2018.

[ABF13] Louigi Addario-Berry and Kevin Ford. Poisson-Dirichlet branching random walks. *Ann. Appl. Probab.*, 23(1):283–307, 2013.

[ABS21] Louigi Addario-Berry and Sanchayan Sen. Geometry of the minimal spanning tree of a random 3-regular graph. *Probability Theory and Related Fields*, 180(3):553–620, 2021.

[Ald97] David Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *The Annals of Probability*, pages 812–854, 1997.

[BBRKK25] Étienne Bellin, Arthur Blanc-Renaudie, Emmanuel Kammerer, and Igor Kortchemski. Uniform attachment with freezing. *The Annals of Applied Probability*, 35(4):2882–2922, 2025.

[Ber09] Nathanaël Berestycki. Recent progress in coalescent theory. *Ensaios Matemáticos*, 16:1–193, 2009.

[Bil13] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[BS98] Erwin Bolthausen and A-S Sznitman. On Ruelle's probability cascades and an abstract cavity method. *Communications in mathematical physics*, 197:247–276, 1998.

[CSD25] Trevor Cousins, Aylwyn Scally, and Richard Durbin. A Structured Coalescent Model Reveals Deep Ancestral Structure Shared by All Modern Humans. *Nature Genetics*, 57(4):856–864, April 2025.

[Dev87] Luc Devroye. Branching processes in the analysis of the heights of trees. *Acta Informatica*, 24(3):277–298, 1987.

[DF99] Robert P Dobrow and James Allen Fill. Total path length for random recursive trees. *Combinatorics, Probability and Computing*, 8(4):317–333, 1999.

[DL95] Luc Devroye and Jiang Lu. The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures & Algorithms*, 7(1):1–14, 1995.

[DR76] Jon Doyle and Ronald L. Rivest. Linear expected time of a simple union-find algorithm. *Inf. Process. Lett.*, 5(5):146–148, 1976.

[Esl22] Laura Eslava. Depth of vertices with high degree in random recursive trees. *ALEA. Latin American Journal of Probability & Mathematical Statistics*, 19(1), 2022.

[FHN06] Michael Fuchs, Hsien-Kuei Hwang, and Ralph Neininger. Profiles of random trees: Limit theorems for random recursive trees and binary search trees. *Algorithmica*, 46:367–407, 2006.

[GM05] Christina Goldschmidt and James Martin. Random recursive trees and the Bolthausen-Sznitman coalesent. *Electronic Journal of Probability*, 10:718–745, 2005.

[GPS18] Christophe Garban, Gábor Pete, and Oded Schramm. The scaling limits of the minimal spanning tree and invasion percolation in the plane. *Ann. Probab.*, 46(6):3501–3557, 2018.

[GS02] William Goh and Eric Schmutz. Limit distribution for the maximum degree of a random recursive tree. *Journal of computational and applied mathematics*, 142(1):61–82, 2002.

[Hoe94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[Jan05] Svante Janson. Asymptotic degree distribution in random recursive trees. *Random Structures & Algorithms*, 26(1-2):69–83, 2005.

[Kin82a] John FC Kingman. On the genealogy of large populations. *Journal of applied probability*, 19(A):27–43, 1982.

[Kin82b] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[Mah08] Hosam Mahmoud. *Pólya urn models*. Chapman and Hall/CRC, 2008.

[NVD25] Rasmus Nielsen, Andrew H. Vaughn, and Yun Deng. Inference and Applications of Ancestral Recombination Graphs. *Nature Reviews Genetics*, 26(1):47–58, January 2025.

[Pem07] Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4(none):1 – 79, 2007.

[Pit94] Boris Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures & Algorithms*, 5(2):337–347, 1994.

[Pit99a] Jim Pitman. Coalescent random forests. *Journal of Combinatorial Theory, Series A*, 85(2):165–193, 1999.

[Pit99b] Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902, 1999.

[Zha15] Yazhe Zhang. On the number of leaves in a random recursive tree. *Brazilian Journal of Probability and Statistics*, 29(4):897 – 908, 2015.

LOUIGI ADDARIO-BERRY
EMAIL: louigi@gmail.com
DEPARTMENT OF MATHEMATICS AND STATISTICS
MCGILL UNIVERSITY

CAELAN ATAMANCHUK
EMAIL: caelan.atamanchuk@gmail.com
DEPARTMENT OF MATHEMATICS AND STATISTICS
MCGILL UNIVERSITY

MAXWELL KAYE
EMAIL: maxwell.kaye@mail.mcgill.ca
DEPARTMENT OF MATHEMATICS AND STATISTICS
MCGILL UNIVERSITY