Electron. J. Probab. **27** (2022), article no. 149, 1–45.
ISSN: 1083-6489 https://doi.org/10.1214/22-EJP877

On joint properties of vertices with a given degree or label in the random recursive tree*

Bas Lodewijks†

Abstract

In this paper, we study the joint behaviour of the degree, depth, and label of and graph distance between high-degree vertices in the random recursive tree. We generalise the results obtained by Eslava [12] and extend these to include the labels of and graph distance between high-degree vertices. The analysis of both these two properties of high-degree vertices is novel, in particular in relation to the behaviour of the depth of such vertices.

In passing, we also obtain results for the joint behaviour of the degree and depth of and graph distance between any fixed number of vertices with a prescribed label. This combines several isolated results on the degree [22], depth [7, 24], and graph distance [9, 15] of vertices with a prescribed label already present in the literature. Furthermore, we extend these results to hold jointly for any number of fixed vertices and improve these results by providing more detailed descriptions of the distributional limits.

Our analysis is based on a correspondence between the random recursive tree and a representation of the Kingman n-coalescent.

Keywords: random recursive tree; Kingman coalescent; depth; label; graph distance; high degrees.
MSC2020 subject classifications: Primary 05C80, Secondary 05C05; 05C12.
Submitted to EJP on May 9, 2022, final version accepted on November 3, 2022.
Supersedes arXiv:2204.09032.

1 Introduction

The random recursive tree model has, since its introduction by Na and Rapoport [30], received a wealth of interest and many properties have been studied. This wide range of topics includes, among others, the degree distribution [20, 26, 27], the degree of vertices with a prescribed label [7, 22], the maximum degree [1, 3, 8, 17, 34], the height of the tree [32], the insertion depth of the tree [7, 24], and the graph distance between

*Bas Lodewijks has been supported by grant GrHyDy ANR-20-CE40-0002.
†Institut Camille Jordan, Univ. Lyon 1, Lyon, France, Univ. Jean Monnet, Saint-Étienne, France.
E-mail: bas.lodewijks@univ-st-etienne.fr

vertices [9, 15]. Beyond these statistics, real-world applications of random recursive trees have been considered as well [16, 28, 31]. See also [10, 25] for two surveys on random trees that include a more extensive overview of the research literature on random recursive trees.

Different approaches for studying the random recursive tree model have been considered throughout the literature. Using the recursive definition of the model and the fact that the random recursive tree with n vertices is defined to be a uniform tree among all increasing trees with n vertices (labelled trees where the vertices on a path from the root to any vertex have increasing labels) are among the most prevalent. Other methods include using continuous-time embedding in Crump-Mode-Jagers branching processes, first introduced by Athreya and Karlin for Pólya urns in [2] and later used for a wide range of recursive tree models such as the random recursive tree (see e.g. [4, 18, 19, 32]), Pólya urns [20] and a representation of Kingman's coalescent [1, 12, 32].

In most studies found in the literature regarding the random recursive tree model, statistics like those mentioned above are considered in *isolation*, rather then studying their *joint behaviour*. As far as the author is aware, only a handful of papers consider the joint behaviour of different statistics for the random recursive tree. In [12], Eslava studies the depth of high-degree vertices, Banerjee and Bhamidi study the label size of the vertex attaining the maximum degree in [3], and the author studies the labels of high-degree vertices in the more general weighted recursive tree model [23], of which the random recursive tree model is a particular example.

The aim of this paper is to extend what is known about the joint behaviour of several statistics of the random recursive tree. We consider, in particular, two settings. First, we study the joint behaviour of the depth and label of and graph distance between any fixed number of vertices selected uniformly at random, conditionally on having a degree that exceeds a certain quantity. We combine, extend, improve and recover the results of the author [23] (in the particular case of the random recursive tree) and Eslava [12]. We also recover the results of Addario-Berry and Eslava [1] and Eslava, the author, and Ortgiese [14] (again, in the particular case of the random recursive tree).

Let T_n denote the random recursive tree with n vertices. Eslava considers in [12] the vector $(d_n^i - \lfloor \log_2 n \rfloor, (h_n^i - \mu \log n)/\sqrt{\sigma^2 \log n})_{i\in[n]}$, where d_n^i and h_n^i denote the degree and depth of the vertex with the i^{th} largest degree (ties broken uniformly at random), respectively, and $\mu := 1 - 1/(2\log 2), \sigma^2 := 1 - 1/(4\log 2)$. Eslava shows this vector converges in distribution along suitable subsequences $(n_t)_{t\in\mathbb{N}}$ to a marked point process on $(\mathbb{Z}\cup\{\infty\})\times\mathbb{R}$, where the marks are independent standard normal random variables. The author proves a similar result for the vector

$$(d_n^i - \lfloor \log_2 n \rfloor, (\ell_n^i - \mu \log n)/\sqrt{(1-\sigma^2)\log n})_{i\in[n]}$$

in [23], where ℓ_n^i denotes the label of the vertex with degree d_n^i (ties broken uniformly at random). Again, along suitable subsequences, this vector converges in distribution to a marked point process on $(\mathbb{Z}\cup\{\infty\})\times\mathbb{R}$, where the marks are independent standard normal random variables. Our results here combine these results to show that the vector

$$(d_n^i - \lfloor \log_2 n \rfloor, (h_n^i - \mu \log n)/\sqrt{\sigma^2 \log n}, (\ell_n^i - \mu \log n)/\sqrt{(1-\sigma^2)\log n})_{i\in[n]}$$

converges along suitable subsequences to a marked point process on $(\mathbb{Z}\cup\{\infty\})\times\mathbb{R}^2$, where the marks are i.i.d. copies of $(M\sqrt{1-\mu/\sigma^2} + N\sqrt{\mu/\sigma^2}, M)$, with M, N, two i.i.d. standard normal random variables. This recovers both results and, additionally, provides a novel and interesting dependence between the scaling limit of the depth and label of high-degree vertices. It describes exactly *how large* the largest degrees in the tree are, as well as *where* and *when* they appear in the tree. This natural extension of the current knowledge provides a rather complete picture of the behaviour of high-degree vertices.

Moreover, we also obtain the distributional convergence of the (properly rescaled) depth and label of and graph distance between any finite number of vertices selected uniformly at random, conditionally on their degrees growing infinitely large as $n \to \infty$. The graph distance between such high-degree vertices has not been studied previously, and we are, in particular, able to characterise the limiting law of the graph distance in terms of the limiting law of the depth of these vertices.

Second, we study the joint behaviour of the degree and depth of and graph distance between any fixed number of vertices with a prescribed label. This combines, extends, improves and recovers a range of results on the degree [7, 22] and depth [7, 24] of and graph distance [9, 15] between vertices with a prescribed label. Given any fixed $k \geq 2$ vertices with labels $(v_{i,n})_{i\in[k]}$ such that $v_{i,n}$ diverges with n, we obtain the joint distributional convergence of the degree and depth of and graph distance between vertices $v_{1,n}, \dots, v_{k,n}$. Again, we characterise the limiting law of the graph distances in terms of those of the depths of vertices $v_{1,n}, \dots, v_{k,n}$, which is novel.

Our extensions of the aforementioned results arise mainly due to two contributions. First, we are able to analyse the joint behaviour of multiple statistics beyond what was known already in the literature. Second, we obtain these results for any finite number of vertices, whereas only a single vertex or single pair of vertices is considered in most results available to date. It is exactly the correlations that arise due to considering several statistics and many vertices at once that prove to be the most challenging aspects of the analysis. The improvement of the existing results is mostly due to the fact that considering the joint behaviour of several statistics allows us, in certain cases, to obtain more detailed descriptions of their limiting laws beyond what was known previously.

The analysis in this paper is based on the Kingman n-coalescent construction of the random recursive tree. This construction was first observed by Pittel in [32] and later recovered and used by Addario-Berry and Eslava [1], and Eslava [12, 13]. This construction provides several advantages compared to the more common recursive construction of the random recursive tree. First, rather than in the recursive construction in which distinct vertices have different arrival times (which influence their degree, depth, label, and graph distance), the coalescent construction allows for a perspective in which all vertices are exchangeable. Second, the coalescent construction enables a more natural decoupling of the statistics of distinct vertices, which provides us with tools to tackle the correlations between these statistics in a more refined manner. Finally, in particular the degree, label and depth of a vertex can be expressed in terms of random numbers of coin flips, simplifying the analysis of these statistics. The degree of a vertex equals the length of the first streak of heads, the label equals the step at which the first tails occurs and the depth equals the total number of tails thrown.

Notation. Throughout the paper we use the following notation: we let $\mathbb{N} := \{1, 2, \dots\}$ denote the natural numbers, set $\mathbb{N}_0 := \{0, 1, \dots\}$ and let $[t] := \{i \in \mathbb{N} : i \leq t\}$ for any $t \geq 1$. For $x \in \mathbb{R}$, we let $\lceil x \rceil := \inf\{n \in \mathbb{Z} : n \geq x\}$ and $\lfloor x \rfloor := \sup\{n \in \mathbb{Z} : n \leq x\}$. For $x \in \mathbb{R}, k \in \mathbb{N}$, we let $(x)_k := x(x-1)\cdots(x-(k-1))$ and $(x)_0 := 1$ and use the notation $\bar{d}$ to denote a k-tuple $d = (d_1, \dots, d_k)$ (the size of the tuple will be clear from the context), where the $d_1, \dots, d_k$ are either numbers or sets. For sequences $(a_n, b_n)_{n\in\mathbb{N}}$ such that b_n is positive for all n we say that $a_n = o(b_n), a_n = \omega(b_n), a_n \sim b_n, a_n = \mathcal{O}(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0, \lim_{n\to\infty} |a_n|/b_n = \infty, \lim_{n\to\infty} a_n/b_n = 1$ and if there exists a constant $C > 0$ such that $|a_n| \leq Cb_n$ for all $n \in \mathbb{N}$, respectively. For random variables $X, (X_n)_{n\in\mathbb{N}}$ we let $X_n \xrightarrow{d} X, X_n \xrightarrow{\mathbb{P}} X$ and $X_n \xrightarrow{a.s.} X$ denote convergence in distribution, probability and almost sure convergence of X_n to X, respectively. Also, let $\Phi : \mathbb{R} \to (0, 1)$ denote the cumulative density function of a standard normal random variable.

We also provide a table with the most important symbols used throughout the paper and their definitions, in order of appearance.

Symbol	Definition
T_n	Random recursive tree on n vertices
$\mathrm{d}_{T_n}(u)$	In-degree of vertex u in T_n
$\mathrm{dist}_{T_n}(u,v)$	Graph distance between vertices u, v in T_n
$h_{T_n}(u)$	Depth of vertex u in T_n (graph distance to the root, $\mathrm{dist}_{T_n}(u,1)$)
v^j	j^{th} vertex in T_n, in decreasing order of in-degree
d_n^j	In-degree of v^j, $\mathrm{d}_{T_n}(v^j)$
h_n^j	Depth of v^j, $h_{T_n}(v^j)$
μ	$1-\frac{1}{2\log 2}$
σ^2	$1-\frac{1}{4\log 2}$
$(v_i)_{i\in[k]}$	k distinct vertices in T_n selected uniformly at random
$T^{(n)}$	Kingman n-coalescent tree
$\mathrm{d}_{T^{(n)}}(i)/\mathrm{d}_n(i)$	In-degree of vertex i in $T^{(n)}$
$h_{T^{(n)}}(i)/h_n(i)$	Depth of vertex i in $T^{(n)}$
$\ell_{T^{(n)}}(i)/\ell_n(i)$	Label of vertex i in $T^{(n)}$ after relabelling (as in (3.4))
$\mathcal{S}_n(i)$	Selection set of vertex i in $T^{(n)}$
$S_n(i)$	$\lvert\mathcal{S}_n(i)\rvert$
$\overline{\mathcal{S}_n}$	$(\mathcal{S}_n(i))_{i\in[k]}$
τ_k	$\max\cup_{1\le i<j\le k}\big(\mathcal{S}_n(i)\cap\mathcal{S}_n(j)\big)$, the first coalescence of vertices $1,\dots,k$
$\mathcal{S}_{n,1}(i)$	Truncated selection set of vertex i in $T^{(n)}$
$\overline{\mathcal{S}}_{n,1}$	$(\mathcal{S}_{n,1}(i))_{i\in[k]}$
$\overline{\mathcal{R}}_{n,1}$	$(\mathcal{R}_{n,1}(i))_{i\in[k]}$, where each element is an independent copy of $\mathcal{S}_{n,1}(1)$
$h_{n,1}(i)$	Truncated depth of vertex i in $T^{(n)}$
$h_{n,2}(i)$	$h_n(i)-h_{n,1}(i)$, the remaining depth

2 Definitions and main results

The random recursive tree model is defined as follows:

Definition 2.1 (Random recursive tree model)**.** *Let $(T_n)_{n\in\mathbb{N}}$ be a sequence of trees. Initialise T_1 by a root with label 1. For every $n\in\mathbb{N}$, construct T_{n+1} from T_n by adding a vertex with label $n+1$ to T_n and connecting it by a directed edge to a vertex $v\in[n]$ which is selected uniformly at random.*

Due to the temporal nature of the random recursive tree model, it is natural to think of the edges as directed towards the root. Throughout, for any $n\in\mathbb{N}$ and $u,v\in[n]$, we write

$$\begin{aligned}\mathrm{d}_{T_n}(u) &:= \text{in-degree of vertex } u \text{ in } T_n,\\ \mathrm{dist}_{T_n}(u,v) &:= \text{graph distance between vertices } u,v \text{ in } T_n,\\ h_{T_n}(u) &:= \text{depth of vertex } u \text{ in } T_n = \mathrm{dist}_{T_n}(u,1).\end{aligned}$$

The graph distance between vertices u and v denotes the number of edge on the unique path between vertices. Here we do not take the direction of the edges into account. This only matters for the in-degree.

Addario-Berry and Eslava study behaviour of high-degree vertices in the RRT in [1] and Eslava extends this to the joint convergence of the degree and depth of such high-degree vertices in [12]. We further extend this joint convergence by including the rescaled label of the vertices as well in the following result.

Theorem 2.2 (Degree, depth and label of high-degree vertices in the RRT)**.** *Consider the random recursive tree (RRT) model as in Definition 2.1. Let $v^1, v^2, \dots, v^n$ be the vertices in the RRT in decreasing order of their in-degree (where ties are split uniformly at random) and let $(d_n^s, h_n^s, \ell_n^s)_{s\in[n]}$ denote their in-degree, depth, and label, respectively.*

Fix $\epsilon \in [0,1]$, define $\epsilon_n := \log_2 n - \lfloor \log_2 n \rfloor$, and let $(n_t)_{t\in\mathbb{N}}$ be a positive, diverging, integer-valued sequence such that $\epsilon_{n_t} \to \epsilon$ as $t \to \infty$. Finally, let $(P_s)_{s\in\mathbb{N}}$ be the points of the Poisson point process $\mathcal{P}$ on $\mathbb{R}$ with intensity measure $\lambda(\mathrm{d}x) = 2^{-x}\log 2\,\mathrm{d}x$, ordered in decreasing order, let $(M_s, N_s)_{s\in\mathbb{N}}$ be two sequences of i.i.d. standard normal random variables and define $\mu := 1 - 1/(2\log 2)$ and $\sigma^2 := 1 - 1/(4\log 2)$. Then, as $t \to \infty$,

$$\begin{aligned}&\Big(d^s_{n_t} - \lfloor \log_2 n_t \rfloor, \frac{h^s_{n_t} - \mu \log n_t}{\sqrt{\sigma^2 \log n_t}}, \frac{\log(\ell^s_{n_t}) - \mu \log n_t}{\sqrt{(1-\sigma^2)\log n_t}}, s \in [n_t]\Big)\\ &\xrightarrow{d} \Big(\lfloor P_s + \epsilon \rfloor, M_s\sqrt{1 - \frac{\mu}{\sigma^2}} + N_s\sqrt{\frac{\mu}{\sigma^2}}, M_s, s \in \mathbb{N}\Big).\end{aligned}$$

Remark 2.3. Theorem 2.2 extends both [23, Theorem 2.6] in the case of the random recursive tree, as well as [12, Theorem 1.2] (since, for each $s \in \mathbb{N}$, we have that $M_s\sqrt{1-\mu/\sigma^2} + N_s\sqrt{\mu/\sigma^2} \sim \mathcal{N}(0,1)$). Moreover, it provides the relation and dependence between the depth of a high-degree vertex and its label, which only becomes apparent in the second-order scaling and the limit.

Beyond studying the behaviour of vertices with 'near-maximum' degree, we are also interested in a more general setting. Here, we select $k \in \mathbb{N}$ many vertices uniformly at random from T_n and condition on their degree. We can then provide the following detailed results on the joint behaviour of their depths, labels and the graph distances between them. The following result is instrumental in proving Theorem 2.2 as well.

Theorem 2.4. *Consider the random recursive tree model as in Definition 2.1. Fix $k \in \mathbb{N}$, $(a_i)_{i\in[k]} \in [0,2)^k$ and let $(v_i)_{i\in[k]}$ be k distinct vertices chosen uniformly at random from $[n]$. Let $(d_i)_{i\in[k]}$ be k integer-valued sequences such that*

$$\lim_{n\to\infty} \frac{d_i}{\log n} = a_i,$$

for each $i \in [k]$. The tuple

$$\Big(\Big(\frac{h_{T_n}(v_i) - (\log n - d_i/2)}{\sqrt{\log n - d_i/4}}\Big)_{i\in[k]}, \Big(\frac{\mathrm{dist}_{T_n}(v_i,v_j) - (2\log n - (d_i+d_j)/2)}{\sqrt{2\log n - (d_i+d_j)/4}}\Big)_{1\le i<j\le k}\Big), \quad (2.1)$$

conditionally on the event $\mathrm{d}_{T_n}(v_i) \ge d_i$ for all $i \in [k]$, converges in distribution to

$$\Big((H_i)_{i\in[k]}, \Big(\frac{\sqrt{4-a_i}H_i + \sqrt{4-a_j}H_j}{\sqrt{8-(a_i+a_j)}}\Big)_{1\le i<j\le k}\Big),$$

where the $(H_i)_{i\in[k]}$ are independent standard normal random variables. Additionally assume that for all $i \in [k]$, d_i diverges as $n \to \infty$. Then, the tuple

$$\begin{aligned}&\Big(\Big(\frac{h_{T_n}(v_i) - (\log n - d_i/2)}{\sqrt{\log n - d_i/4}}, \frac{\log v_i - (\log n - d_i/2)}{\sqrt{d_i/4}}\Big)_{i\in[k]},\\ &\Big(\frac{\mathrm{dist}_{T_n}(v_i,v_j) - (2\log n - (d_i+d_j)/2)}{\sqrt{2\log n - (d_i+d_j)/4}}\Big)_{1\le i<j\le k}\Big),\end{aligned} \quad (2.2)$$

conditionally on the event $\mathrm{d}_{T_n}(v_i) \ge d_i$ for all $i \in [k]$, converges in distribution to

$$\begin{aligned}&\Big(\Big(M_i\sqrt{\frac{a_i}{4-a_i}} + N_i\sqrt{1-\frac{a_i}{4-a_i}}, M_i\Big)_{i\in[k]},\\ &\Big(\frac{M_i\sqrt{a_i} + N_i\sqrt{4-2a_i} + M_j\sqrt{a_j} + N_j\sqrt{4-2a_j}}{\sqrt{8-(a_i+a_j)}}\Big)_{1\le i<j\le k}\Big),\end{aligned}$$

where the $(M_i, N_i)_{i\in[k]}$ are independent standard normal random variables.

Remark 2.5. (i) With an almost identical proof, the same results can be obtained when using the conditional event $\{\mathrm{d}_{T_n}(v_i) = d_i, i \in [k]\}$ rather than $\{\mathrm{d}_{T_n}(v_i) \geq d_i, i \in [k]\}$.

(ii) When $a_i = 0$ for all $i \in [k]$, we obtain the behaviour of the *insertion depth* of k uniform vertices, as well as the graph distance between them.

(iii) The conditional convergence of the tuple in (2.1) recovers, improves, and extends the result of Eslava in [12, Theorem 1.1]. When we omit the distance between the vertices v_i, v_j and set $d_i := \lfloor a_i \log n \rfloor + b_i$ for some $a_i \in [0, 2), b_i \in \mathbb{Z}$ for all $i \in [k]$, we obtain [12, Theorem 1.1]. Our result allows for a greater freedom in the choice of the degrees d_i rather than the parametrised setting used by Eslava. We extend Eslava's result even further by including the graph distance between any pair of vertices and, in (2.2), by also including the label of the vertices $v_1, \dots, v_k$. The latter also allows for a more precise description of the limiting distribution of the depth compared to [12, Theorem 1.1]. We observe that the scaling of the graph distance suggests that the graph distance between vertices v_i and v_j, for any distinct $i, j \in [k]$, is the sum of their depths. Though this sum is a trivial upper bound, we show that it is of the correct order by using the fact that the largest common ancestor of v_i and v_j, $\mathrm{LCA}_{i,j}$, forms a tight sequence of random variables (in $n \in \mathbb{N}$).

Next to conditioning on the degree of vertices selected uniformly at random, we also have the following result on the degree and depth of and graph distance between vertices with a fixed label. Though the marginal convergence of the degree and depth of a vertices and graph distance of a pair of vertices with a fixed label has been studied previously (see [22, 7, 24, 9, 15]), we combine, extend, and improve these results by considering the joint convergence and by allowing for any number of (pairs of) vertices.

Theorem 2.6. *Consider the random recursive tree model as in Definition 2.1 Fix $k \in \mathbb{N}$ and let $(v_{i,n})_{i\in[k]} \in [n]^k$ be k distinct integer-valued sequences such that $v_{i,n}$ increases with n, diverges as $n \to \infty$ and such that*

$$c_{i,j} := \lim_{n\to\infty} \sqrt{\frac{\log v_{i,n}}{\log v_{i,n} + \log v_{j,n}}}$$

exists for all $1 \leq i < j \leq k$. Let $(N_i)_{i\in[k]}$ be k independent standard normal random variables. We also define for each $i \in [k]$,

$$\mathrm{d}^*_{T_n}(v_{i,n}) := \begin{cases} \frac{\mathrm{d}_{T_n}(v_{i,n}) - \log(n/v_{i,n})}{\sqrt{\log(n/v_{i,n})}} & \text{if } v_{i,n} = o(n), \\ \mathrm{d}_{T_n}(v_{i,n}), & \text{otherwise,} \end{cases},$$

and let $(Z_i)_{i\in[k]}$ be k independent random variables (also independent of $(N_i)_{i\in[k]}$) such that, for $(\rho_i)_{i\in[k]} \in (0, 1)^k$,

$$Z_i \sim \begin{cases} \mathcal{N}(0, 1) & \text{if } v_{i,n} = o(n), \\ \mathrm{Poi}(\log(1/\rho_i)) & \text{if } v_{i,n} = (1 + o(1))\rho_i n, \\ 0 & \text{if } v_{i,n} = n - o(n). \end{cases}$$

Then,

$$\left(\left(\mathrm{d}^*_{T_n}(v_{i,n}), \frac{h_{T_n}(v_{i,n}) - \log v_{i,n}}{\sqrt{\log v_{i,n}}} \right)_{i\in[k]}, \left(\frac{\mathrm{dist}_{T_n}(i, j) - (\log v_{i,n} + \log v_{j,n})}{\sqrt{\log v_{i,n} + \log v_{j,n}}} \right)_{1\leq i<j\leq k} \right)$$
$$\xrightarrow{d} \big((Z_i, N_i)_{i\in[k]}, (c_{i,j}N_i + c_{j,i}N_j)_{1\leq i<j\leq k} \big).$$

Remark 2.7. (i) The theorem partially recovers a result from Feng, Lui, and Su [15, Theorem 1], where the distance between vertices i_n and n for any integer sequence

$(i_n)_{n\in\mathbb{N}}$ such that $i_n \in [n-1]$ holds is covered. In our setting, we require the labels $v_{i,n}$ to be increasing in n and to diverge with n, as we are unable to characterise the limiting distributions of the depth and degree otherwise. We also recover the less general results (compared to Feng *et al.*) of Dobrow [9, Theorems 3 and 4] on the graph distance between vertices i_n and n with $i_n = n-1$ or $i_n = \lfloor \lambda n \rfloor, \lambda \in (0,1)$. Moreover, we are able to provide a more detailed description of the scaling limit of the distance between the vertices $v_{1,n}, \dots, v_{k,n}$ in relation to their depth, which is not present in [15] or [9].

(ii) The theorem recovers the results of Devroye [7] and Mahmoud [24] on the insertion depth.

(iii) The theorem recovers a result of Kuba and Panholzer [22, Theorem 2] regarding the degree of a vertex with a prescribed label.

(iv) In all cases described in points (i), (ii) and (iii), we extend the results of Feng et al., Devroye, Mahmoud, and Kuba and Panholzer to k vertices and $\binom{k}{2}$ pairs of vertices for any $k \geq 2$.

(v) The constraint that all $v_{i,n}$ are increasing in n arises due a technicality, which we illustrate with the following example. Suppose $k = 2$ and

$$v_{1,n} = \lfloor n/2 \rfloor \mathbb{1}_{\{n \text{ is even}\}} + \lfloor n/3 \rfloor \mathbb{1}_{\{n \text{ is odd}\}}, \qquad v_{2,n} = \lfloor n/3 \rfloor \mathbb{1}_{\{n \text{ is even}\}} + \lfloor n/2 \rfloor \mathbb{1}_{\{n \text{ is odd}\}}.$$

In this case, $c_{1,2} = c_{2,1} = 1/\sqrt{2}$ both exist, so that the limiting law of the graph distance can be obtained, but the limiting laws of $d^*_{T_n}(v_{1,n})$ and $d^*_{T_n}(v_{2,n})$ do not exist. Indeed, $d^*_{T_{2n}}(v_{1,2n}) \xrightarrow{d} \text{Poi}(\log 2)$ and $d^*_{T_{2n+1}}(v_{1,2n+1}) \xrightarrow{d} \text{Poi}(\log 3)$. Such cases are circumvented when the $v_{i,n}$ are increasing with n. When omitting the degree, *any* diverging sequences $(v_{i,n})_{i\in[k]}$ such that the $(c_{i,j})_{1\leq i<j\leq k}$ exist can be considered.

The main approach to proving Theorems 2.2, 2.4, and 2.6 is to use a 'reversed-time' construction or coalescent construction of the random recursive tree, known as the Kingman n-coalescent construction (see Section 3). This construction has several advantages compared to the construction in Definition 2.1. First, the depth, degree, and label of vertices in the Kingman n-coalescent are exchangeable, which simplifies the analysis of their joint behaviour. Second, the coalescent construction simplifies dealing with correlations that appear when considering the depth, degree, and label of multiple vertices at once. In particular, it provides an elegant way to decouple the degree, depth, and label of distinct vertices. Finally, the size of the depth, degree, and label of a vertex can be understood in terms of sums of independent indicator random variables and independent fair coin flips. As a result, standard central limit theorem results can be applied to obtain the desired results.

Outline of the paper

The paper is organised as follows: We first provide some theoretical preparations, necessary to prove the Theorems stated in Section 2. We provide a perspective for Theorem 2.2 in terms of marked point processes, and provide a construction of the random recursive tree, called the Kingman n-coalescent construction, that aids in the analysis of the properties of interest here. In particular, we rephrase Theorems 2.4 and 2.6 in terms of the Kingman n-coalescent in Theorems 3.5 and 3.7, respectively. Section 4 is then dedicated to developing some preliminary results based on the Kingman n-coalescent construction. These preliminary results are used in Sections 5 and 6 to obtain intermediate results on the behaviour of high-degree vertices and vertices with a given label, respectively. Finally, these intermediate results are used in Section 7 to prove Theorem 2.2 and in Section 8 to prove Theorems 2.4 and 2.6.

3 The degree, depth, and label of high-degree vertices in the random recursive tree: theoretical preparations

In this section we provide a new perspective of Theorem 2.2, alongside a different construction of the random recursive tree compared to Definition 2.1. The latter will be of aid in proving all results presented in Section 2.

To prove Theorem 2.2, we use the convergence of marked point processes. Recall that d_n^s, h_n^s and ℓ_n^s denote the degree, depth, and label of the vertex with the s^{th} largest degree in the random recursive tree, respectively, with $s \in [n]$, where ties are split uniformly at random. Let $\mu := 1 - 1/(2\log 2)$ and $\sigma^2 := 1 - 1/(4\log 2)$. We view the tuples

$$\Big(d_n^s - \lfloor \log_2 n \rfloor, \frac{h_n^s - \mu \log n}{\sqrt{(1-\sigma^2)\log n}}, \frac{\log \ell_n^s - \mu \log n}{\sqrt{\sigma^2 \log n}}, s \in [n]\Big),$$

as a marked point process, where the rescaled degrees form the points and the rescaled depth and label form the marks of the points. Let $\mathbb{Z}^* := \mathbb{Z} \cup \{\infty\}$ and endow $\mathbb{Z}^*$ with the metric $d(s,t) := |2^{-s} - 2^{-t}|, d(s,\infty) = 2^{-s}, s,t \in \mathbb{Z}$. We work with $\mathbb{Z}^*$ rather than $\mathbb{Z}$, as sets $[s,\infty]$ for $s \in \mathbb{Z}$ are now compact. Let $\mathcal{P}$ be a Poisson point process on $\mathbb{R}$ with intensity $\lambda(\mathrm{d}x) := 2^{-x}\log 2\,\mathrm{d}x$ and let $(\xi_x^{(1)}, \xi_x^{(2)})_{x\in\mathcal{P}}$ be independent standard normal random variables. For $\epsilon \in [0,1]$, we define the ground process $\mathcal{P}^\epsilon$ on $\mathbb{Z}^*$ and the marked process $\mathcal{MP}^\epsilon$ on $\mathbb{Z}^* \times \mathbb{R}^2$ by

$$\mathcal{P}^\epsilon := \sum_{x\in\mathcal{P}} \delta_{\lfloor x+\epsilon \rfloor}, \qquad \mathcal{MP}^\epsilon := \sum_{x\in\mathcal{P}} \delta_{(\lfloor x+\epsilon\rfloor, \sqrt{\mu/\sigma^2}\xi_x^{(1)} + \sqrt{1-\mu/\sigma^2}\xi_x^{(2)}, \xi_x^{(2)})}, \tag{3.1}$$

where δ is a Dirac measure. Similarly, we define

$$\begin{aligned} \mathcal{P}^{(n)} &:= \sum_{v=1}^n \delta_{\mathrm{d}_{T_n}(v) - \lfloor \log_2 n\rfloor}, \\ \mathcal{MP}^{(n)} &:= \sum_{v=1}^n \delta_{(\mathrm{d}_{T_n}(v) - \lfloor \log_2 n\rfloor, (h_{T_n}(v) - \mu\log n)/\sqrt{\sigma^2 \log n}, (\log v - \mu \log n)/\sqrt{(1-\sigma^2)\log n})}. \end{aligned} \tag{3.2}$$

We then let $\mathcal{M}^\#_{\mathbb{Z}^*}$ and $\mathcal{M}^\#_{\mathbb{Z}^*\times\mathbb{R}^2}$ be the spaces of boundedly finite measures on $\mathbb{Z}^*$ and $\mathbb{Z}^* \times \mathbb{R}^2$, respectively, and observe that $\mathcal{P}^{(n)}, \mathcal{P}^\epsilon$ and $\mathcal{MP}^{(n)}, \mathcal{MP}^\epsilon$ are elements of $\mathcal{M}^\#_{\mathbb{Z}^*}$ and $\mathcal{M}^\#_{\mathbb{Z}^*\times\mathbb{R}^2}$, respectively. Theorem 2.2 is then equivalent to the weak convergence of $\mathcal{MP}^{(n_t)}$ to $\mathcal{MP}^\epsilon$ in $\mathcal{M}^\#_{\mathbb{Z}^*\times\mathbb{R}^2}$ along suitable subsequences $(n_t)_{t\in\mathbb{N}}$, as we can order the points in the definition of $\mathcal{MP}^{(n)}$ (resp. $\mathcal{MP}^\epsilon$) in decreasing order of their degrees (resp. of the points $x \in \mathcal{P}$). We remark that the weak convergence of $\mathcal{P}^{(n_t)}$ to $\mathcal{P}^\epsilon$ in $\mathcal{M}^\#_{\mathbb{Z}^*}$ along subsequences has been established by Addario-Berry and Eslava in [1] (later generalised to weighted recursive trees by Eslava, the author, and Ortgiese in [14] and extended to marked point processes by the author in [23]) and that Eslava established the weak convergence of $\widetilde{\mathcal{MP}}^{(n_t)}$ along subsequences, which is $\mathcal{MP}^{(n_t)}$ with each mark restricted to the first element (i.e. not considering the label), in [12]. We extend these results here to the tuple of degree, depth, and label, which also shows an interesting dependence in the limit of the rescaled depth and rescaled labels.

Recall the Poisson point process $\mathcal{P}$ used in the definition of $\mathcal{P}^\epsilon$ in (3.1) and enumerate its points in decreasing order. That is, P_v denotes the v^{th} largest point of $\mathcal{P}$ (ties broken uniformly at random). We observe that this is well-defined, since $\mathcal{P}([x,\infty)) < \infty$ almost surely for any $x \in \mathbb{R}$. Also, let $(M_v, N_v)_{v\in\mathbb{N}}$ be two sequences of i.i.d. standard normal random variables. To prove the weak convergence of the marked point process $\mathcal{MP}^{(n)}$,

we define, for $s \in \mathbb{Z}, B \in \mathcal{B}(\mathbb{R}^2)$, the counting measures

$$
\begin{aligned}
X_s^{(n)}(B) &:= \Big|\Big\{v \in [n] : \mathrm{d}_{T_n}(v) = \lfloor \log_2 n \rfloor + s, \Big(\frac{h_{T_n}(v) - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{\log n - (\lfloor \log_2 n \rfloor + s)/4}}, \\
&\qquad\qquad \frac{\log v - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{(\lfloor \log_2 n \rfloor + s)/4}}\Big) \in B\Big\}\Big|, \\
X_{\geq s}^{(n)}(B) &:= \Big|\Big\{v \in [n] : \mathrm{d}_{T_n}(v) \geq \lfloor \log_2 n \rfloor + s, \Big(\frac{h_{T_n}(v) - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{\log n - (\lfloor \log_2 n \rfloor + s)/4}}, \\
&\qquad\qquad \frac{\log v - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{(\lfloor \log_2 n \rfloor + s)/4}}\Big) \in B\Big\}\Big|, \\
\widetilde{X}_s^{(n)}(B) &:= \Big|\Big\{v \in [n] : \mathrm{d}_{T_n}(v) = \lfloor \log_2 n \rfloor + s, \Big(\frac{h_{T_n}(v) - \mu \log n}{\sqrt{\sigma^2 \log n}}, \frac{\log v - \mu \log n}{\sqrt{(1-\sigma^2)\log n}}\Big) \in B\Big\}\Big|, \\
\widetilde{X}_{\geq s}^{(n)}(B) &:= \Big|\Big\{v \in [n] : \mathrm{d}_{T_n}(v) \geq \lfloor \log_2 n \rfloor + s, \Big(\frac{h_{T_n}(v) - \mu \log n}{\sqrt{\sigma^2 \log n}}, \frac{\log v - \mu \log n}{\sqrt{(1-\sigma^2)\log n}}\Big) \in B\Big\}\Big|, \\
X_s(B) &:= \Big|\Big\{v \in \mathbb{N} : \lfloor P_v + \epsilon \rfloor = s, \Big(M_v\sqrt{1 - \frac{\mu}{\sigma^2}} + N_v\sqrt{\frac{\mu}{\sigma^2}}, M_v\Big) \in B\Big\}\Big|, \\
X_{\geq s}(B) &:= \Big|\Big\{v \in \mathbb{N} : \lfloor P_v + \epsilon \rfloor \geq s, \Big(M_v\sqrt{1 - \frac{\mu}{\sigma^2}} + N_v\sqrt{\frac{\mu}{\sigma^2}}, M_v\Big) \in B\Big\}\Big|.
\end{aligned}
\tag{3.3}
$$

We note that, when $s = o(\sqrt{\log n})$, $X_s^{(n)}(B) \approx \widetilde{X}_s^{(n)}(B)$ and $X_{\geq s}^{(n)}(B) \approx \widetilde{X}_{\geq s}^{(n)}(B)$ for any fixed $B \subseteq \mathbb{R}$. For the result in Theorem 2.2 we are interested in the distributional convergence of $\widetilde{X}_s^{(n)}(B), \widetilde{X}_{\geq s}^{(n)}(B)$ to $X_s(B), X_{\geq s}(B)$, which we obtain in a more general setting for the random variables $X_s^{(n)}(B), X_{\geq s}^{(n)}(B)$. The following intermediate result related to these counting measures aids us in obtaining this distributional convergence.

Proposition 3.1 (Factorial moments of counting measures)**.** *Fix constants $K \in \mathbb{N}$ and $(a_m)_{m\in[K]} \in [0,2)^K$. Let $(s_m)_{m\in[K]}$ be a non-decreasing integer-valued sequence with $0 \leq K' := \min\{m : s_{m+1} = s_K\}$ such that $s_1 + \log_2 n = \omega(1)$ and*

$$\lim_{n\to\infty} \frac{s_m + \log_2 n}{\log n} = a_m,$$

for all $m \in [K]$. Let $(B_m)_{m\in[K]}$ be a sequence of sets $B_m \subset \mathcal{B}(\mathbb{R}^2)$ such that $B_m \cap B_\ell = \varnothing$ when $s_m = s_\ell$ and $m \neq \ell$, let $(c_m)_{m\in[K]} \in \mathbb{N}_0^K$ and let M and N be two independent standard normal random variables. Recall the random variables $X_s^{(n)}(B), X_{\geq s}^{(n)}(B)$ and $\widetilde{X}_s^{(n)}(B), \widetilde{X}_{\geq s}^{(n)}(B)$ from (3.3)*, and define $\epsilon_n := \log_2 n - \lfloor \log_2 n \rfloor$. Then,*

$$
\begin{aligned}
&\mathbb{E}\Big[\prod_{m=1}^{K'} \Big(X_{s_m}^{(n)}(B_m)\Big)_{c_m} \prod_{m=K'+1}^{K} \Big(X_{\geq s_m}^{(n)}(B_m)\Big)_{c_m}\Big] \\
&\quad = (1+o(1)) \prod_{m=1}^{K'} \Big(2^{-(s_m+1)+\epsilon_n}\mathbb{P}\Big(\Big(M\sqrt{\frac{a_m}{4-a_m}} + N\sqrt{1-\frac{a_m}{4-a_m}}, M\Big) \in B_m\Big)\Big)^{c_m} \\
&\qquad \times \prod_{m=K'+1}^{K} \Big(2^{-s_K+\epsilon_n}\mathbb{P}\Big(\Big(M\sqrt{\frac{a_m}{4-a_m}} + N\sqrt{1-\frac{a_m}{4-a_m}}, M\Big) \in B_m\Big)\Big)^{c_m}.
\end{aligned}
$$

Moreover, when $s_1, \dots, s_K = o(\sqrt{\log n})$ *and* $a_m = 1/\log 2$ *for all* $m \in [K]$,

$$\begin{aligned}
&\mathbb{E}\Big[\prod_{m=1}^{K'} \Big(\widetilde{X}^{(n)}_{s_m}(B_m)\Big)_{c_m} \prod_{m=K'+1}^{K} \Big(\widetilde{X}^{(n)}_{\geq s_m}(B_m)\Big)_{c_m}\Big] \\
&\quad = (1+o(1)) \prod_{m=1}^{K'} \left(2^{-(s_m+1)+\epsilon_n} \mathbb{P}\left(\left(M\sqrt{1-\frac{\mu}{\sigma^2}} + N\sqrt{\frac{\mu}{\sigma^2}}, M\right) \in B_m\right)\right)^{c_m} \\
&\qquad \times \prod_{m=K'+1}^{K} \left(2^{-s_K+\epsilon_n} \mathbb{P}\left(\left(M\sqrt{1-\frac{\mu}{\sigma^2}} + N\sqrt{\frac{\mu}{\sigma^2}}, M\right) \in B_m\right)\right)^{c_m}.
\end{aligned}$$

As the counting measures defined in (3.3) are sums of indicator random variables, their factorial moments can be expressed in terms of probabilities

$$\begin{aligned}
&\mathbb{P}(\mathrm{d}_{T_n}(v_i) \geq d_i, (h_{T_n}(v_i), \log v_i) \in B_i, i \in [k]) \\
&\quad = \mathbb{P}(\mathrm{d}_{T_n}(v_i) \geq d_i, i \in [k])\, \mathbb{P}((h_{T_n}(v_i), \log v_i) \in B_i, i \in [k] \,|\, \mathrm{d}_{T_n}(i) \geq d_i, i \in [k])\,.
\end{aligned}$$

Here, we let $(d_i)_{i\in[k]} \in \mathbb{N}_0^k$ such that $d_i < 2\log n$, $(B_i)_{i\in[k]} \in \mathcal{B}(\mathbb{R}^2)^k$, $(v_i)_{i\in[k]}$ distinct vertices selected uniformly at random, and $k \in \mathbb{N}$. The first probability on the right-hand side is studied by Addario-Berry and Eslava in [1], and the latter is the subject of Theorem 2.4. This can in turn be used to prove Proposition 3.1, which finally leads to Theorem 2.2. We provide more details alongside the proof of Proposition 3.1 and Theorem 2.2 in Section 7.

3.1 The Kingman n-coalescent

We now provide an alternative construction of the random recursive tree (RRT), which we use to prove Theorems 2.2, 2.4 and 2.6.

This alternative construction of the RRT, (a variant of) the Kingman n-coalescent construction, was first discussed by Pittel in [32] and recovered and used by Addario-Berry and Eslava to study high degrees in RRTs [1]. Later, Eslava extended this to the joint convergence of the depth and degree of vertices with large degree [12] and also provides a more general coupled recursive construction of a tree T and a permutation σ on the labels of the vertices of T, coined Robin-Hood pruning [13]. Here, we further extend Eslava's results from [12] on the depth and degree of high-degree vertices to also include the label of and graph distance between such high-degree vertices. We also obtain results on the joint behaviour of the degree and depth of and graph distance between vertices with a given label, which combine, extend and improve several known results from the literature on the degree [22] and depth [7] of a vertex with a given label and the graph distance between vertices n and i_n, for any sequence i_n [15].

The variant of the Kingman n-coalescent we use here is a process which starts with n trees, each consisting of only a single root. At every step n through 2 (counting backwards), a pair of roots is selected uniformly at random and independently of this selection a directed edge is formed between the two roots, each direction being equiprobable. This reduces the number of trees by one and, after completing step 2, yields a directed tree. It turns out that a particular relabelling of this directed tree yields a tree equal in law to the random recursive tree. Moreover, using the Kingman n-coalescent construction simplifies the analysis of degrees, depths, and labels in the RRT model, among other reasons because the degree, depth, and label of the vertices are exchangeable random variables in the Kingman n-coalescent.

We now formally introduce the Kingman n-coalescent construction of the random recursive tree. Let $\mathcal{CF}_n := \{f : V(f) = [n]\}$ denote the set of all forests with exactly n vertices. An n-chain is a sequence $(f_n, \dots, f_1)$ of elements of $\mathcal{CF}_n$, where for each integer

$1 < j \leq n$, f_{j-1} is obtained from f_j by adding a directed edge between the roots of two trees in f_j. We write $f_j = \{t_1^{(j)}, \dots, t_j^{(j)}\}$, ordering the trees in increasing order of their smallest-labelled vertex. In particular, f_n consists of n trees, each of which is a root with no edges, and f_1 consists of exactly one tree. Also, we let $r(T)$ denote the root of the tree T and write $F_j = \{T_1^{(j)}, \dots, T_j^{(j)}\}$ for a random element in $\mathcal{CF}_n$ for any $j \in [n]$.

Definition 3.2 (Kingman n-coalescent)**.** *For each $1 < j \leq n$, choose a pair $\{a_j, b_j\} \subseteq \{\{a, b\} : 1 \leq a < b \leq j\}$ independently and uniformly at random; also let $(\xi_j)_{1<j\leq n}$ be a sequence of independent* $\mathrm{Bernoulli}(1/2)$ *random variables. Initialise the coalescent by F_n: a forest of n trees, each consisting of a root and no edges. For $1 < j \leq n$, F_{j-1} is obtained from F_j as follows: Add an edge e_{j-1} between the roots $r(T_{a_j}^{(j)})$ and $r(T_{b_j}^{(j)})$; direct e_{j-1} towards $r(T_{a_j}^{(j)})$ if $\xi_j = 1$ and towards $r(T_{b_j}^{(j)})$ if $\xi_j = 0$. Then, F_{j-1} consists of the new tree and the remaining $j-1$ unaltered trees from F_j.*

Finally, let $T^{(n)} := T_1^{(1)} = F_1$ denote the final tree in the coalescent $\mathbf{C} = (F_n, \dots, F_1)$.

See Figure 1 for an example of the process. When at step j the edge $e_j = v_j u_j$ is directed towards u_j, we say that the associated random variable ξ_j (which we can interpret as flipping a fair coin) favours the root u_i. Similarly, we might also say that ξ_j favours w or that the associated coin flip at step j favours w, where w is any vertex in the tree that contains u_j.

The link between the final tree in the coalescent and the RRT is as follows. Let us define the mapping $\sigma_C : V(T^{(n)}) \to [n]$ by $\sigma_C(r(T^{(n)})) := 1$ and for each edge $e_j = v_j u_j \in E(T^{(n)})$, $j \in [n-1]$,

$$\sigma_C(v_j) := j + 1. \tag{3.4}$$

As all edges are directed towards the root, $v_j \neq v_{j'}$ for all $j \neq j' \in [n-1]$, so that σ_C is well-defined. σ_C is the relabelling of $T^{(n)}$ into an increasing tree. If we let $\mathcal{I}_n$ denote the set of all increasing trees on n vertices, then it is clear that the RRT is a uniform element in $\mathcal{I}_n$. The most important attribute of the n-chain in the Kingman n-coalescent is that it has a uniform distribution over all possible n-chains and that the relabelling of $T^{(n)}$ by σ_C yields a uniform element of $\mathcal{I}_n$, as outlined in the following proposition.

Proposition 3.3 (Lemma 7.1 and Proposition 7.2 in [12])**.** *The Kingman n-coalescent $\mathbf{C}$ is uniformly random in $\mathcal{CF}_n$, the set of n-chains. Moreover, for each $C = (f_n, \dots, f_1) \in \mathcal{CF}_n$, relabel the vertices in f_1 with σ_C to obtain a tree $\phi(C) \in \mathcal{I}_n$. Then the law of $\phi(\mathbf{C})$ is that of a random recursive tree of size n.*

Recall that $\mathrm{d}_{T_n}(u)$, $h_{T_n}(u)$ and $\mathrm{dist}_{T_n}(u, v)$ denote the in-degree and depth of vertex $u \in [n]$ and the graph distance between vertices $u, v \in [n]$ in the random recursive tree T_n, respectively. Similarly, for a realisation of the final tree $T^{(n)}$ in the coalescent $\mathbf{C}$, let $\mathrm{d}_{T^{(n)}}(i), h_{T^{(n)}}(i)$ and $\mathrm{dist}_{T^{(n)}}(i, j)$ denote the in-degree and depth of vertex i and the graph distance between i and j, respectively, and let $\ell_{T^{(n)}}(i) := \sigma_C(i)$ denote the relabelling of vertex i, $i \in [n]$. That is, $\ell_{T^{(n)}}(i)$ denotes the label that vertex i in $\mathbf{C}$ obtains in the random recursive tree $\phi(\mathbf{C})$. We can then formulate the following corollary.

Corollary 3.4. *Let T_n be a random recursive tree and let $T^{(n)}$ be the resulting tree in the Kingman n-coalescent. Let $\sigma : [n] \to [n]$ be a uniform random permutation on $[n]$. Then,*

$$\begin{aligned}&\big((\mathrm{d}_{T^{(n)}}(i), h_{T^{(n)}}(i), \ell_{T^{(n)}}(i))_{i\in[n]}, (\mathrm{dist}_{T^{(n)}}(i,j))_{1\leq i<j\leq n}\big)\\ &\quad\stackrel{d}{=} \big((\mathrm{d}_{T_n}(\sigma(i)), h_{T_n}(\sigma(i)), \sigma(i))_{i\in[n]}, (\mathrm{dist}_{T_n}(\sigma(i),\sigma(j)))_{1\leq i<j\leq n}\big).\end{aligned}$$

Moreover, jointly for all $i, j \in \mathbb{N}$ and all sets $B \subseteq [n]$, we have

$$|\{v \in B : \mathrm{d}_{T^{(n)}}(v) = i, h_{T^{(n)}}(v) = j\}| \stackrel{d}{=} |\{v \in [n] : \sigma(v) \in B, \mathrm{d}_{T_n}(\sigma(v)) = i, h_{T_n}(\sigma(v)) = j\}|.$$

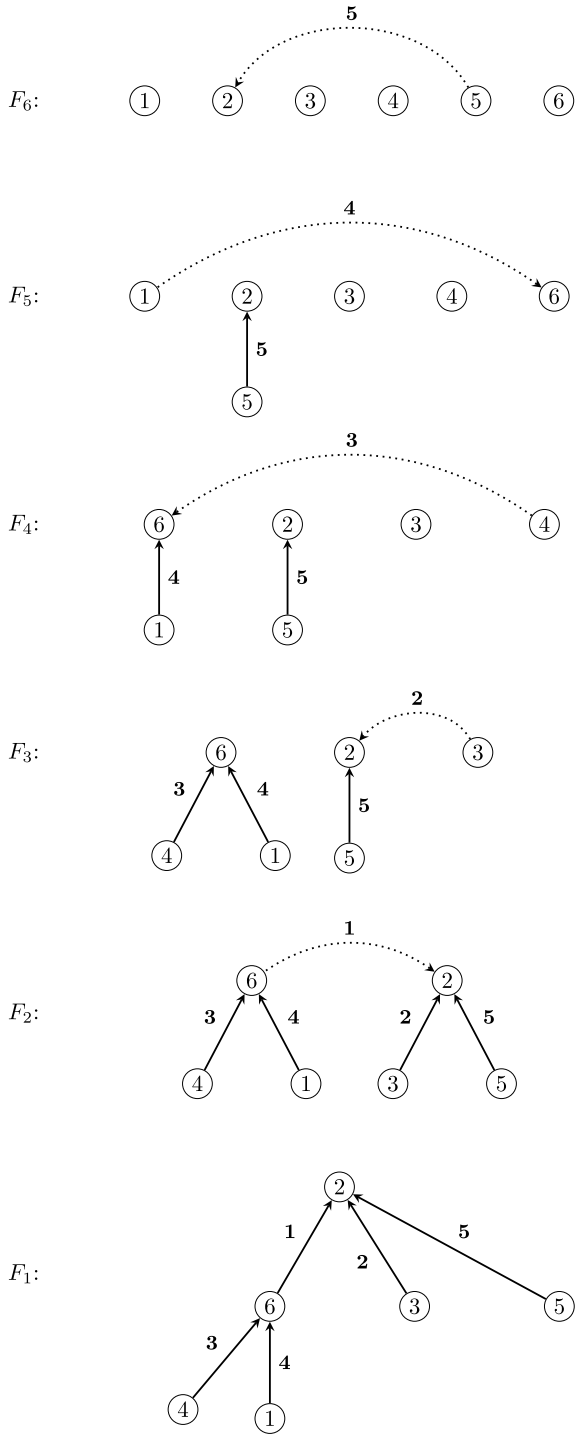

Figure 1: An example of the Kingman n-coalescent $\mathbf{C} = (F_n, \dots, F_1)$ for $n = 6$. For $2 \le j \le 6$, we represent the edge in $E(F_{j-1}) \backslash E(F_j)$ with a dotted line in F_j. In this case, $\xi_6 = \xi_4 = \xi_3 = 1, \xi_5 = \xi_2 = 0$ and $\{a_6, b_6\} = \{2, 5\}$, $\{a_5, b_5\} = \{1, 5\}$, $\{a_4, b_4\} = \{1, 4\}$, $\{a_3, b_3\} = \{2, 3\}$, $\{a_2, b_2\} = \{1, 2\}$. From [12].

In what follows, we replace the subscript $T^{(n)}$ with n for ease of writing, since we work with the coalescent from now on instead of the RRT. As a direct result from Corollary 3.4, Theorem 2.4 follows from the following result (which is a reformulation of Theorem 2.4 in terms of the Kingman n-coalescent).

Theorem 3.5. *Consider the Kingman n-coalescent as in Definition 3.2. Fix $k \in \mathbb{N}$, and $(a_i)_{i\in[k]} \in [0,2)^k$ and let $(d_i)_{i\in[k]}$ be k integer-valued sequences such that*

$$\lim_{n\to\infty} \frac{d_i}{\log n} = a_i,$$

for each $i \in [k]$. The tuple

$$\Big(\Big(\frac{h_n(i) - (\log n - d_i/2)}{\sqrt{\log n - d_i/4}} \Big)_{i\in[k]} \Big(\frac{\mathrm{dist}_n(i,j) - (2\log n - (d_i+d_j)/2)}{\sqrt{2\log n - (d_i+d_j)/4}} \Big)_{1\le i<j\le k} \Big),$$

conditionally on the event $d_n(i) \ge d_i$ for all $i \in [k]$, converges in distribution to

$$\Big((H_i)_{i\in[k]}, \Big(\frac{\sqrt{4-a_i}H_i + \sqrt{4-a_j}H_j}{\sqrt{8-(a_i+a_j)}} \Big)_{1\le i<j\le k} \Big), \tag{3.5}$$

where the $(H_i)_{i\in[k]}$ are independent standard normal random variables. Additionally assume that for all $i \in [k]$, d_i diverges as $n \to \infty$. Then, the tuple

$$\begin{aligned} \Big(\Big(\frac{h_n(i) - (\log n - d_i/2)}{\sqrt{\log n - d_i/4}}, \frac{\log(\ell_n(i)) - \log n - d_i/4}{\sqrt{d_i/4}} \Big)_{i\in[k]}, \\ \Big(\frac{\mathrm{dist}_n(i,j) - (2\log n - (d_i+d_j)/2)}{\sqrt{2\log n - (d_i+d_j)/4}} \Big)_{1\le i<j\le k} \Big), \end{aligned}$$

conditionally on the event $d_n(i) \ge d_i$ for all $i \in [k]$, converges in distribution to

$$\begin{aligned} \Big(\Big(M_i\sqrt{\frac{a_i}{4-a_i}} + N_i\sqrt{1-\frac{a_i}{4-a_i}}, M_i \Big)_{i\in[k]}, \\ \Big(\frac{M_i\sqrt{a_i} + N_i\sqrt{4-2a_i} + M_j\sqrt{a_j} + N_j\sqrt{4-2a_j}}{\sqrt{8-(a_i+a_j)}} \Big)_{1\le i<j\le k} \Big), \end{aligned} \tag{3.6}$$

where the $(M_i, N_i)_{i\in[k]}$ are independent standard normal random variables.

Remark 3.6. As is the case in Remark 2.5, the same results in Theorem 3.5 can be obtained when working with the conditional event $\{\mathrm{d}_n(v_i) = d_i, i \in [k]\}$ rather than $\{\mathrm{d}_n(v_i) \ge d_i, i \in [k]\}$, with an almost identical proof.

Moreover, Theorem 3.5 can be used to prove Proposition 3.1. By Corollary 3.4, we can redefine the random variables $X_s^{(n)}(B), X_{\ge s}^{(n)}(B)$ and $\widetilde{X}_s^{(n)}(B), \widetilde{X}_{\ge s}^{(n)}(B)$, as defined

in (3.3), in terms of the Kingman n-coalescent, by writing, for $s \in \mathbb{Z}, B \in \mathcal{B}(\mathbb{R}^2)$,

$$\begin{aligned}
X_s^{(n)}(B) &:= \Big|\Big\{i \in [n] : \mathrm{d}_n(i) = \lfloor \log_2 n \rfloor + s, \Big(\frac{h_n(i) - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{\log n - (\lfloor \log_2 n \rfloor + s)/4}}, \\
&\qquad\qquad \frac{\log \ell_n(i) - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{(\lfloor \log_2 n \rfloor + s)/4}}\Big) \in B\Big\}\Big|, \\
X_{\geq s}^{(n)}(B) &:= \Big|\Big\{i \in [n] : \mathrm{d}_n(i) \geq \lfloor \log_2 n \rfloor + s, \Big(\frac{h_n(i) - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{\log n - (\lfloor \log_2 n \rfloor + s)/4}}, \\
&\qquad\qquad \frac{\log \ell_n(i) - (\log n - (\lfloor \log_2 n \rfloor + s)/2)}{\sqrt{(\lfloor \log_2 n \rfloor + s)/4}}\Big) \in B\Big\}\Big|, \\
\widetilde{X}_s^{(n)}(B) &:= \Big|\Big\{i \in [n] : \mathrm{d}_n(i) = \lfloor \log_2 n \rfloor + s, \Big(\frac{h_n(i) - \mu \log n}{\sqrt{\sigma^2 \log n}}, \frac{\log \ell_n(i) - \mu \log n}{\sqrt{(1-\sigma^2)\log n}}\Big) \in B\Big\}\Big|, \\
\widetilde{X}_{\geq s}^{(n)}(B) &:= \Big|\Big\{i \in [n] : \mathrm{d}_n(i) \geq \lfloor \log_2 n \rfloor + s, \Big(\frac{h_n(i) - \mu \log n}{\sqrt{\sigma^2 \log n}}, \frac{\log \ell_n(i) - \mu \log n}{\sqrt{(1-\sigma^2)\log n}}\Big) \in B\Big\}\Big|.
\end{aligned} \tag{3.7}$$

We can also reformulate Theorem 2.6 in terms of the Kingman n-coalescent. As is the case with Theorem 2.4, combining Corollary 3.4 with the following theorem immediately implies Theorem 2.6.

Theorem 3.7. *Consider the Kingman n-coalescent as in Definition 3.2. Fix $k \in \mathbb{N}$ and let $(\ell_i)_{i\in[k]} \in [n]^k$ be k distinct integer-valued sequences such that ℓ_i increases with n, diverges as $n \to \infty$ and such that*

$$c_{i,j} := \lim_{n\to\infty} \sqrt{\frac{\log \ell_i}{\log \ell_i + \log \ell_j}} \tag{3.8}$$

exists for all $1 \leq i < j \leq k$. Let $(N_i)_{i\in[k]}$ be k independent standard normal random variables. We also define for $(\rho_i)_{i\in[k]} \in (0,1)^k$ and each $i \in [k]$,

$$\begin{aligned}
d_n^*(i) &:= \begin{cases} \frac{\mathrm{d}_n(i) - \log(n/\ell_i)}{\sqrt{\log(n/\ell_i)}}, & \text{if } \ell_i = o(n), \\ \mathrm{d}_n(i), & \text{otherwise}, \end{cases} \\
Z_i &\sim \begin{cases} \mathcal{N}(0,1) & \text{if } \ell_i = o(n), \\ \mathrm{Poi}(\log(1/\rho_i)) & \text{if } \ell_i = (1+o(1))\rho_i n, \\ 0 & \text{if } \ell_i = n - o(n). \end{cases}
\end{aligned} \tag{3.9}$$

where the Z_i are independent and also independent of the $(N_i)_{i\in[k]}$. The tuple

$$\Big(\Big(d_n^*(i), \frac{h_n(i) - \log \ell_i}{\sqrt{\ell_i}}\Big)_{i\in[k]}, \Big(\frac{\mathrm{dist}_n(i,j) - (\log \ell_i + \log \ell_j)}{\sqrt{\log \ell_i + \log \ell_j}}\Big)_{1\leq i<j\leq k}\Big),$$

conditionally on the event $\ell_n(i) = \ell_i$ for all $i \in [k]$, converges in distribution to

$$\big((Z_i, N_i)_{i\in[k]}, (c_{i,j}N_i + c_{j,i}N_j)_{1\leq i<j\leq k}\big).$$

Remark 3.8. It is necessary to work on the conditional event $\{\ell_n(i) = \ell_i, i \in [k]\}$ in Theorem 3.7, despite this not being the case in Theorem 2.6. Since vertices $1, \ldots, k$ in the Kingman n-coalescent obtain a *random* label in the relabelled tree $\phi(\mathbf{C})$ (which is equal in law to the random recursive tree by Proposition 3.3), the need to condition on their relabelling $\ell_n(i) = \ell_i, i \in [k]$, arises.

In the next sections we analyse the Kingman n-coalescent construction to prove Theorems 3.5 and 3.7 and Proposition 3.1.

4 Preliminary results

In this section we provide some important intermediate results related to the Kingman n-coalescent construction, provided in Section 3. We focus on two things in this section. First, we study the evolution of the degree, depth, and label of vertices $1, \dots, k$ in the Kingman n-coalescent, which is an important first step in proving the theorems in Section 3. Second, we investigate the correlations between the steps $j \in [2, n]$ at which vertices $1, \dots, k$ are selected in the coalescent.

Though the theorems presented in Section 3 are concerned with the graph distance between vertices $1, \dots, k$ as well as their degree, depth, and label, we do not include this in our analysis yet. While the latter quantities are easier to explicitly understand in terms of the Kingman n-coalescent, the graph distance does not lend itself to an equally elegant analysis. As it turns out, though, there is a close relation between the depth of and graph distance between the vertices $1, \dots, k$ which allows us to infer the scaling limit of the graph distances from the results on the depth. We make use of this relation in Section 8 when proving Theorems 3.5 and 3.7.

4.1 Analysis of the Kingman n-coalescent

We start by introducing some notation related to the Kingman n-coalescent. For an n-chain $C = (f_n, \dots, f_1)$ and some $i, j \in [n]$, let $T^{(j)}(i)$ denote the tree in f_j that contains vertex i. For $i \in [n]$, let $s_{i,j}$ be the indicator that $T^{(j)}(i) \in \{T^{(j)}_{a_j}, T^{(j)}_{b_j}\}$ and let $h_{i,j}$ be the indicator that the edge e_j is directed outwards from $r(T^{(j)}(i))$, $2 \le j \le n$. That is, $s_{i,j}$ equals one if i is part of one of the two trees selected to merge at step j, and $h_{i,j}$ is one if $s_{i,j}$ is one and if the new edge e_j causes vertex i to increase its depth by one, see Figure 2.

Since the trees selected to be merged at every step are independent and uniformly distributed, the variables $(s_{i,j})_{2 \le j \le n}$ are independent Bernoulli random variables for any fixed $i \in [n]$, with $\mathbb{E}[s_{i,j}] = 2/j$. Similarly, since the direction of the edge e_j depends only on ξ_j, the variables $(h_{i,j})_{2 \le j \le n}$ are also independent Bernoulli random variables for any fixed $i \in [n]$, with $\mathbb{E}[h_{i,j}] = 1/j$.

Let us define

$$\mathcal{S}_n(i) := \{2 \le j \le n : s_{i,j} = 1\}, \qquad i \in [n],$$

and set $S_n(i) := |\mathcal{S}_n(i)|$. We refer to $\mathcal{S}_n(i)$ as the *selection set* of vertex i. We can express the quantities $\mathrm{d}_n(i), h_n(i)$ and $\ell_n(i)$ in terms of $\mathcal{S}_n(i)$ and the indicator variables $(h_{i,j})_{j \in \mathcal{S}_n(i)}$. Namely, if we write $\mathcal{S}_n(i) = \{j_{i,1}, \dots, j_{i,S_n(i)}\}$ with $j_{i,1} > j_{i,2} > \dots > j_{i,S_n(i)}$,

Figure 2: For $i \in [n]$ and $2 \le j \le n$, let $r_j := r(T^{(j)}(i))$ denote the root of the tree in f_j that contains vertex i, and suppose that $T^{(j)}(i) \in \{T^{(j)}_{a_j}, T^{(j)}_{b_j}\}$. If e_j is directed towards r_j, then the degree of r_j increases by one in F_{j-1}. If e_i is directed outwards from r_j, then the depth of each $v \in T^{(j)}(i)$ increases by one in F_{j-1}. From [12].

then

$$\begin{aligned} \mathrm{d}_n(i) &= \max\{0 \le d \le S_n(i) : h_{i,j_{i,1}} = \ldots = h_{i,j_{i,d}} = 0\}, \\ h_n(i) &= \sum_{j \in \mathcal{S}_n(i)} h_{i,j} = \sum_{j=2}^{n} h_{i,j}, \\ \ell_n(i) &= \max\{j \in \mathcal{S}_n(i) : h_{i,j} = 1\} = \max\{i \in [n] : h_{i,j} = 1\}, \end{aligned} \tag{4.1}$$

where we set $h_{i,1} = 1$ for all $i \in [n]$, so that $\max\{j \in [n] : h_{i,j} = 1\} = 1$ if there is no $2 \le j \le n$ such that $h_{i,j} = 1$ (which corresponds to vertex i being the root of $T^{(n)}$, so that its relabelling with σ_C as in (3.4) yields $\ell_n(i) = 1$). Note that there is always a unique vertex i for which $h_{i,j} = 0$ for all $2 \le j \le n$, so that $\ell_n(i) \neq \ell_n(i')$ whenever $i \neq i'$. Explaining (4.1) in words, the degree of a vertex i is equal to the length of the first streak of zeros of the indicators $(h_{i,j_\ell})_{\ell \in [S_n(i)]}$, the relabelling of vertex i in the RRT is equal to the first step directly after this streak when $h_{i,j} = 1$, and the depth equals the total number of steps j for which $h_{i,j} = 1$.

We are interested in the behaviour of the degree, depth, and label of the vertices $1, \ldots, k$ for *any* fixed $k \in \mathbb{N}$. While these quantities are easily expressed in terms of the selection sets $(\mathcal{S}_n(i))_{i \in [k]}$ and the associated coin flips, as in (4.1), considering k vertices provides some additional difficulties in terms of correlations between the selection sets of these k vertices. The main issue is the following: whenever two distinct vertices $i, i' \in [k]$ are both selected *at the same step*, say step $\lambda_{i,i'}$, there is a dependence between the outcome of the associated coin flip of vertices i and i'. Namely, $h_{i,\lambda_{i,i'}} = 1 - h_{j,\lambda_{i,i'}}$. Furthermore, for any step $2 \le j < \lambda_{i,j}$, we know that $h_{i,j} = h_{i',j}$. As these correlations between the vertices $1, \ldots, k$ are difficult to handle, we define

$$\tau_k := \max\{2 \le j \le n : s_{i,j} = s_{i',j} = 1 \text{ for distinct } i, i' \in [k]\}. \tag{4.2}$$

Since the trees in the Kingman n-coalescent are ordered based on their smallest-labelled vertex, τ_k is the first step at which two vertices $i, i' \in [k]$ are both selected (in the sense that the root of the tree they belong to is selected), and thus up to step τ_k the vertices $1, \ldots, k$ are contained in disjoint trees. As a result, this implies that the sets $[\tau_k + 1, n] \cap \mathcal{S}_n(1), \ldots, [\tau_k + 1, n] \cap \mathcal{S}_n(k)$ are disjoint, and since the associated coin flips of these disjoint sets are independent, the evolutions of the degree, depth, and label of vertices $1, \ldots, k$, up to step τ_k are independent. This helps to avoid correlations and simplifies the analysis. Eslava (implicitly) shows in the proof of [12, Lemma 3.2] that the sequence $(\tau_k)_{n \in \mathbb{N}}$ is a tight sequence of random variables. As a result, for any integer-valued sequence $(t_n)_{n \in \mathbb{N}}$ which diverges to infinity as $n \to \infty$, we know that $\mathbb{P}(\tau_k < t_n) = 1 - o(1)$. This justifies, for $t_n \le n$, the definition of the sets, for each $i \in [n]$,

$$\Omega_1 := \{t_n, \ldots, n\}, \quad \mathcal{S}_{n,1}(i) := \{j \in \Omega_1 : s_{i,j} = 1\}, \quad \mathcal{H}_{n,1}(i) := \{j \in \Omega_1 : h_{i,j} = 1\}, \tag{4.3}$$

and we let $S_{n,1}(i) := |\mathcal{S}_{n,1}(i)|$ and $h_{n,1}(i) := |\mathcal{H}_{n,1}(i)|, h_{n,2}(i) := h_n(i) - h_{n,1}(i)$. We refer to the sets $(\mathcal{S}_{n,1}(i))_{i \in [n]}$ as the *truncated selection sets*, to $h_{n,1}(i)$ as the *truncated depth* of vertex i, and to $(t_n)_{n \in \mathbb{N}}$ as the *truncation sequence*. Though $\mathcal{S}_{n,1}(i)$, $h_{n,1}(i), h_{n,2}(i)$ depend on t_n, we omit this in their notation for ease of writing. The truncated depth $h_{n,1}(i)$ and $h_{n,2}(i)$ can be described similar to $h_n(i)$ in (4.1), as

$$h_{n,1}(i) = \sum_{j \in \mathcal{S}_{n,1}(i)} h_{i,j} = \sum_{j=t_n}^{n} h_{i,j}, \qquad h_{n,2}(i) = \sum_{j \in \mathcal{S}_n(i) \setminus \mathcal{S}_{n,1}} h_{i,j} = \sum_{j=2}^{t_n - 1} h_{i,j} = h_n(i) - h_{n,1}(i).$$

The following lemma uses (4.1) to provide a description of the relation between the joint distribution of $\mathrm{d}_n(1), h_{n,1}(1)$ and $\ell_n(1)$ and the truncated selection set $\mathcal{S}_{n,1}(1)$. Since the vertices are exchangeable, as follows from Corollary 3.4, the lemma also holds for any vertex $i \in [n]$.

Lemma 4.1. *Let $G \sim \mathrm{Geo}(1/2)$ be independent from $\mathcal{S}_n(1)$. Then $\mathrm{d}_n(1) \stackrel{d}{=} \min\{G, S_n(1)\}$. Moreover, fix $h, d \in \mathbb{N}_0$ and consider a truncation sequence $(t_n)_{n,\in\mathbb{N}}$ such that $t_n \leq n$ for all $n \in \mathbb{N}$. Let $\ell \in \Omega_1$, $J \subseteq \Omega_1$, and let $X_{n,\ell,1} \sim \mathrm{Bin}(|[\ell, n] \cap J| - d, 1/2)$ and $X_{n,\ell,2} \sim \mathrm{Bin}(|[t_n, \ell-1] \cap J|, 1/2)$ be two independent binomial random variables (where we set $X_{n,\ell,1} = 0, X_{n,\ell,2} = 0$ when $|[\ell, n] \cap J| - d \leq 0, |[t_n, \ell - 1] \cap J| = 0$, respectively). Then,*

$$\begin{aligned} &\mathbb{P}(h_{n,1}(1) \leq h, \ell_n(1) \geq \ell, \mathrm{d}_n(1) \geq d \,|\, \mathcal{S}_{n,1}(1) = J) \\ &\quad = 2^{-d} \mathbb{1}_{\{|[\ell,n] \cap J| \geq d+1\}} \mathbb{P}(X_{n,\ell,1} + X_{n,\ell,2} \leq h, X_{n,\ell,1} \geq 1)\,. \end{aligned} \tag{4.4}$$

Furthermore,

$$\begin{aligned} &\mathbb{P}(h_{n,1}(1) \leq h, \ell_n(1) = \ell, \mathrm{d}_n(1) \leq d \,|\, \mathcal{S}_{n,1}(1) = J) \\ &\quad = \mathbb{1}_{\{|[\ell+1,n] \cap J| \leq d\}} \mathbb{1}_{\{\ell \in J\}} 2^{-(|[\ell+1,n] \cap J| + 1)} \mathbb{P}(X_{n,\ell,2} \leq h - 1)\,. \end{aligned} \tag{4.5}$$

Remark 4.2. In the case $\ell = 1$, the result in (4.4) simplifies to

$$\mathbb{P}(h_n(1) \leq h, \mathrm{d}_n(1) \geq d \,|\, \mathcal{S}_{n,1}(1) = J) = 2^{-d} \mathbb{1}_{\{|J| \geq d\}} \mathbb{P}(X_n \leq h)\,,$$

where $X_n \sim \mathrm{Bin}(|J| - d, 1/2)$ and we set $X_n = 0$ when $|J| - d \leq 0$. The proof follows the same approach as the proof of (4.4) and is hence omitted.

Remark 4.3. The constraint $\ell \geq t_n$ ensures that the events $\ell_n(1) \geq \ell$ and $\ell_n(1) = \ell$, as in (4.4) and (4.5), respectively, can be determined by step t_n of the Kingman n-coalescent. In what follows, we let t_n grow sufficiently slow so that this constraint is satisfied for any choice of ℓ that is of interest.

Proof. Let us start by proving (4.4). We define $\mathcal{E}_n := \{h_{n,1}(1) \leq h, \ell_n(1) \geq \ell, \mathrm{d}_n(1) \geq d\}$. If we condition on the event $\{\mathcal{S}_{n,1}(1) = J\}$ for some set $J \subseteq \Omega_1$, then we can express the occurrence and probability of the event $\mathcal{E}_n$ in terms of J:

- (i) Conditionally on $\{\mathcal{S}_{n,1}(1) = J\}$, $\mathcal{E}_n$ can only occur if $|[\ell, n] \cap J| \geq d + 1$ by the first and last line of (4.1):
  - (a) By the first line of (4.1), the degree of vertex i is at least d when a streak $h_{1,j_{1,1}} = \ldots = h_{1,j_{1,d}} = 0$ occurs, where we recall that $\mathcal{S}_n(1) = \{j_{1,1}, \ldots, j_{1,S_n(1)}\}$ (and, similarly, $\mathcal{S}_{n,1}(1) = \{j_{1,1}, \ldots, j_{1,S_{n,1}(1)}\}$). This can only happen when vertex 1 is selected at at least d steps, so $S_n(1) \geq d$, and the coin flips associated with the first d of these steps need to be heads.
  - (b) After this streak, vertex 1 needs to be selected at least once more, but not later than step ℓ. Moreover, the associated coin flip at this step has to be tails to ensure that the label of vertex 1 in the random recursive tree is at least ℓ, by the last line of (4.1). So, combined with (a), J needs to contain at least $d + 1$ elements that are at least ℓ, i.e. $|[\ell, n] \cap J| \geq d + 1$. Given this, we then require the first d associated coin flips to favour vertex 1 and the remaining $|[\ell, n] \cap J| - d$ coin flips to not favour vertex 1 at least once, i.e. $X_{n,\ell,1} \geq 1$, to obtain a degree at least d and a label at least ℓ.
- (ii) The required streak of d coin flips favouring vertex 1 occurs with probability 2^{-d}, and is independent from everything else which occurs afterwards (in particular, what occurs in steps $(i)_{(b)}$ and (iii)). Moreover, as the coin flips are independent of the selection set, the degree of 1 is determined by the length of the first streak of coin flips that favour 1. So, $\mathrm{d}_n(1) \stackrel{d}{=} \min\{G, S_n(1)\}$.
- (iii) After the first streak of d coin flips that favour vertex 1, the number of remaining coin flips which do not favour vertex 1, associated to the selection set J, should be at most h. That is, $X_{n,\ell,1} + X_{n,\ell,2} \leq h$.

Combining all of the above, we can then write,

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_n \,|\, \mathcal{S}_{n,1}(1) = J) \\
&\quad = \mathbb{1}_{\{|[\ell,n]\cap J|\geq d+1\}}\mathbb{P}(\mathcal{E}_n \,|\, \mathcal{S}_{n,1}(1) = J) && (i)\\
&\quad = 2^{-d}\mathbb{1}_{\{|[\ell,n]\cap J|\geq d+1\}}\mathbb{P}(h_n(1) \leq h, \ell_n(1) \geq \ell \,|\, \mathcal{S}_{n,1}(1) = J, \mathrm{d}_n(1) \geq d) && (ii)\\
&\quad = 2^{-d}\mathbb{1}_{\{|[\ell,n]\cap J|\geq d+1\}}\mathbb{P}(X_{n,\ell,1} + X_{n,\ell,2} \leq h, X_{n,\ell,1} \geq 1)\,, && (i)_{(b)} + (iii)
\end{aligned}$$

where we remark that we can omit the conditioning due to the fact that the coin flips are independent of everything else.

We now prove (4.5). Let us set $\widetilde{\mathcal{E}}_n := \{h_n(1) \leq h, \ell_n(1) = \ell, \mathrm{d}_n(1) \leq d\}$. Again, we express the occurrence and the probability of the event $\widetilde{\mathcal{E}}_n$ in terms of J:

(i) $\ell_n(1) = \ell$ and $\mathrm{d}_n(1) \leq d$ can only occur together if the following two things occur:

- (a) Vertex 1 is selected at most d times in steps n through $\ell+1$, and all associated coin flips favour vertex 1. The latter occurs with probability $2^{-|J\cap[\ell+1,n]|}$.
- (b) Vertex 1 is selected at step ℓ and is not favoured by the associated coin flip. The latter occurs with probability $1/2$.

Indeed, if (a) does not occur then either the degree or the label of vertex 1 (in the random recursive tree) is too large. If (b) does not occur, then the label of vertex 1 (in the random recursive tree) is not equal to ℓ.

(ii) In steps $\ell-1$ through 2, the number of coin flips which do not favour vertex 1, associated to the selection set J, is at most $h-1$ (since the height of 1 equals one after step ℓ). That is, $X_{n,\ell,2} \leq h-1$.

Combining this, we can write

$$\begin{aligned}
\mathbb{P}(\widetilde{\mathcal{E}}_n \,|\, \mathcal{S}_{n,1}(1) = J) =& \mathbb{1}_{\{|[\ell+1,n]\cap J|\leq d\}}\mathbb{1}_{\{\ell\in J\}}\mathbb{P}(\widetilde{\mathcal{E}}_n \,|\, \mathcal{S}_{n,1}(1) = J) && (i)_{(a)} + (i)_{(b)}\\
=& \mathbb{1}_{\{|[\ell+1,n]\cap J|\leq d\}}\mathbb{1}_{\{\ell\in J\}}2^{-(|[\ell+1,n]\cap J|+1)} && (i)_{(a)} + (i)_{(b)}\\
&\times \mathbb{P}(h_n(1) \leq h-1 \,|\, \ell_n(1) = \ell, \mathcal{S}_{n,1}(1) = J)\\
=& \mathbb{1}_{\{|[\ell+1,n]\cap J|\leq d\}}\mathbb{1}_{\{\ell\in J\}}2^{-(|[\ell+1,n]\cap J|+1)}\mathbb{P}(X_{n,\ell,2} \leq h-1)\,. && (ii)
\end{aligned}$$

We remark that in the last step, as in the proof of (4.4), we can omit the conditional event $\{\ell_n(1) = \ell, \mathcal{S}_n(1) = J\}$, as the coin flips are independent of everything else. Moreover, in the second step we can omit the event $\{\mathrm{d}_n(1) \leq d\}$, as the occurrence of $\{h_n(1) \leq h-1\}$, conditionally on $\{\ell_n(1) = \ell\}$ is independent of $\{\mathrm{d}_n(1) \leq d\}$. This concludes the proof. □

We now extend this result to multiple vertices, which we can do with relative ease as long as the truncated selection sets of the vertices $1, \dots, k$ are disjoint. For ease of writing, we let $\overline{S}_{n,1} := (\mathcal{S}_{n,1}(i))_{i\in[k]}$ and $\bar{J} := (J_i)_{i\in[k]}$ (where $J_i \subseteq \Omega_1$ for each $i \in [k]$).

Lemma 4.4. *Fix $k \in \mathbb{N}$ and consider a truncation sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \leq n$ for all $n \in \mathbb{N}$. Let $h_i, d_i \in \mathbb{N}_0, i \in [k], (J_i)_{i\in[k]} \in \Omega_1^k$ such that the $(J_i)_{i\in[k]}$ are pairwise disjoint. Then,*

$$\mathbb{P}(h_{n,1}(i) \leq h_i, \mathrm{d}_n(i) \geq d_i, i \in [k] \,|\, \overline{\mathcal{S}}_{n,1} = \bar{J}) = \prod_{i=1}^{k}\mathbb{P}(h_{n,1}(i) \leq h_i, \mathrm{d}_n(i) \geq d_i \,|\, \mathcal{S}_{n,1}(i) = J_i).$$

If, additionally, we let $\ell_i \in \Omega_1$ for all $i \in [k]$,

$$\begin{aligned}
&\mathbb{P}(h_{n,1}(i) \leq h_i, \ell_n(i) \geq \ell_i, \mathrm{d}_n(i) \geq d_i, i \in [k] \,|\, \overline{\mathcal{S}}_{n,1} = \bar{J})\\
&\quad = \prod_{i=1}^{k}\mathbb{P}(h_{n,1}(i) \leq h_i, \ell_n(i) \geq \ell_i, \mathrm{d}_n(i) \geq d_i \,|\, \mathcal{S}_{n,1}(i) = J_i),
\end{aligned} \tag{4.6}$$

and

$$\mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \ge d_i, i \in [k] \,|\, \overline{\mathcal{S}}_{n,1} = \bar{J}) = \prod_{i=1}^{k} \mathbb{P}((h_{n,1}(i) \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \ge d_i, i \in [k] \,|\, \mathcal{S}_{n,1}(i) = J_i).$$

Proof. The first result follows from [12, Lemma 3.1]. We prove (4.6), the proof of the last result follows an analogous approach.

The proof is similar to that of [12, Lemma 3.1]. Let us write $J_i = \{j_{i,1}, \dots, j_{i,|J_i|}\}$ with $j_{i,1} > \dots > j_{i,J_i}$ for each $i \in [k]$. Conditionally on $\{\overline{\mathcal{S}}_{n,1} = \bar{J}\}$, we have that for each $i \in [k]$,

$$h_{n,1}(i) = \sum_{m=1}^{|J_i|} h_{i,j_{i,m}}.$$

Also, the event $\{\mathrm{d}_n(i) \ge d_i\}$ holds if and only if $|J_i| \ge d_i$ and $h_{i,j_{i,m}} = 0$ for all $m \in [d_i]$, and the event $\{\ell_n(i) \ge \ell_i\}$ holds if and only if $\max\{m \in J_i : h_{i,m} = 0\} \ge \ell_i$. As $\ell_i \ge t_n$, it follows that the event $\{h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, \mathrm{d}_n(i) \ge d_i, i \in [k]\}$, conditionally on $\overline{\mathcal{S}}_{n,1} = \bar{J}$, depends solely on $(h_{i,m})_{m \in J_i, i \in [k]}$ and J_i. Since the sets $(J_i)_{i \in [k]}$ are pairwise disjoint, the occurrence of the events $\{h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, \mathrm{d}_n(i) \ge d_i\}$, for each $i \in [k]$, depend on disjoint sets of random variables. Moreover, since the random variables $h_{i,m}$ for different values of m are determined by independent coin flips, we have that the events $\{h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, \mathrm{d}_n(i) \ge d_i\}$, for each $i \in [k]$, depend on disjoint sets of independent random variables, from which (4.6) follows. A similar reasoning proves the final result. □

To end the first part of this section, we recall a result from Addario-Berry and Eslava on the degree of vertices $1, \dots, k$ in the Kingman n-coalescent.

Proposition 4.5 (Proposition 4.2, [1])**.** *Fix $k \in \mathbb{N}, c \in (0,2)$. There exists $\beta = \beta(c,k) > 0$ such that uniformly over integers $(d_i)_{i \in [k]} \in [0, c \log n]^k$,*

$$\mathbb{P}(\mathrm{d}_n(i) \ge d_i, i \in [k]) = 2^{-\sum_{i=1}^{k} d_i}(1 + o(n^{-\beta})).$$

4.2 Truncated selection sets

As we have seen in the first part of this section, we can obtain explicit formulations for the probability of events related to the size of the degree, depth, and label of vertices $1, \dots, k$ in the Kingman n-coalescent, under certain conditions on the truncated selection sets $\overline{\mathcal{S}}_{n,1}$. In this part of the section, we formalize these conditions and show that they are met with high probability. We also introduce some other properties of the truncated selection sets that are useful in the analysis that follows in Sections 5 through 8.

Recall Ω_1 from (4.3) and recall that we write $\overline{\mathcal{S}}_{n,1} = (\mathcal{S}_{n,1}(i))_{i \in [k]}, \bar{J} = (J_i)_{i \in [k]}$. For $\delta \in (0,2)$ and $\bar{d} = (d_i)_{i \in [k]} \in \mathbb{Z}^k$, define

$$\begin{aligned} \mathcal{A}_{\bar{d}} &:= \{\bar{J} \in \Omega_1^k : \mathbb{P}\big(\overline{\mathcal{S}}_{n,1} = \bar{J}, \mathrm{d}_n(i) \ge d_i, i \in [k]\big) > 0\}, \\ \mathcal{B}_{n,\delta} &:= \{\bar{J} \in \Omega_1^k : (J_1, \dots, J_k) \text{ are pairwise disjoint and } \big||J_i| - 2\log n\big| \le \delta \log n, i \in [k]\}. \end{aligned} \tag{4.7}$$

$\mathcal{A}_{\bar{d}}$ consists of all possible outcomes of the truncated selection sets that enable the event $\{\mathrm{d}_n(i) \ge d_i, i \in [k]\}$, and $\mathcal{B}_{n,\delta}$ consists of all truncated selection sets which enable the decoupling of the depth, label and degree of the vertices $1, \dots, k$, as follows from Lemma 4.4.

We now present some results related to the sets $\mathcal{A}_{\bar{d}}$ and $\mathcal{B}_{n,\delta}$, which are based on several results from [12]. Though we defined the truncated selection sets and the truncated depth in terms of a general truncation sequence t_n, it suffices to consider only

the case $t_n = \lceil (\log n)^2 \rceil$ in the following lemmas (as we will mostly use this choice for t_n in what follows).

Lemma 4.6 (Lemma 3.1, [12]). *Let $\delta \in (0,2)$ and let $t_n = \lceil (\log n)^2 \rceil$. If $\bar{d} = (d_i)_{i\in[k]} \in \mathbb{N}_0^k$ satisfies $d_i < (2-\delta)\log n$ for all $i \in [k]$, then $\mathcal{B}_{n,\delta} \subseteq \mathcal{A}_{\bar{d}}$.*

We have already discussed that $\tau_k < t_n$ with high probability when the truncation sequence t_n diverges as $n \to \infty$ infinity. The concentration of the size of $\mathcal{S}_{n,1}(i)$ around $2\log n$ for any $i \in [k]$ when $t_n = \lceil (\log n)^2 \rceil$ (or, more generally, when $t_n = o(n)$, which follows from a direct application of Bernstein's inequality, see also [12, (32)] for a more formal statement) yields the following result:

Lemma 4.7 (Lemma 3.2, [12]). *Fix an integer $k \in \mathbb{N}$ and $\delta \in (0,2)$ and let $t_n = \lceil (\log n)^2 \rceil$. Then,*

$$\mathbb{P}\big(\overline{\mathcal{S}}_{n,1} \in \mathcal{B}_{n,\delta}\big) = 1 - o(1).$$

We also know that the elements of $\overline{\mathcal{S}}_{n,1}$ are asymptotically independent for any $k \in \mathbb{N}$, uniformly over the set $\mathcal{B}_{n,\delta}$. Let $\overline{\mathcal{R}}_{n,1} := (\mathcal{R}_{n,1}(1), \dots, \mathcal{R}_{n,1}(k))$ be k independent copies of $\mathcal{S}_{n,1}(1)$. Then, we have the following result:

Lemma 4.8 (Lemma 3.2, [12]). *Fix an integer $k \in \mathbb{N}$ and $\delta \in (0,2)$ and let $t_n = \lceil (\log n)^2 \rceil$. Uniformly over $\bar{J} \in \mathcal{B}_{n,\delta}$,*

$$\mathbb{P}\big(\overline{\mathcal{S}}_{n,1} = \bar{J}\big) = (1+o(1))\mathbb{P}\big(\overline{\mathcal{R}}_{n,1} = \bar{J}\big)\,.$$

The following lemma provides bounds for the decay of the tail distribution of τ_k, conditionally on certain events.

Lemma 4.9. *Fix $k \in \mathbb{N}$ and recall τ_k from (4.2). We have that $(\tau_k)_{n\in\mathbb{N}}$ is a tight sequence of random variables. Furthermore, fix $c \in (0,2)$ and let $(d_i)_{i\in[k]} \in \mathbb{N}_0^k$ such that $d_i \le c\log n$ for all $i \in [k]$. Then,*

$$\mathbb{P}\Big(\tau_k < \lceil (\log n)^2 \rceil \,\Big|\, \mathrm{d}_n(i) \ge d_i, i \in [k]\Big) \ge 1 - \mathcal{O}\Big(\frac{1}{\log n}\Big). \tag{4.8}$$

Furthermore, let $(\ell_i)_{i\in[k]} \in [n]^k$ be distinct such that ℓ_i diverges as $n \to \infty$ for all $i \in [k]$. Then,

$$\mathbb{P}\bigg(\tau_k < \min_{i\in[k]} \log \ell_i \,\bigg|\, \ell_n(i) = \ell_i, i \in [k]\bigg) \ge 1 - \mathcal{O}\Big(\frac{1}{\min_{i\in[k]} \log \ell_i}\Big). \tag{4.9}$$

Proof. We first prove the tightness of $(\tau_k)_{n\in\mathbb{N}}$. Fix $\epsilon > 0$ and set $K_\epsilon := \lceil 2 + k^2/\epsilon \rceil$. We recall that in Definition 3.2, $\{a_j, b_j\}$ denotes the two trees selected at step j in the Kingman n-coalescent, for $2 \le j \le n$. Also, the trees are ordered by their smallest-labelled vertex, so that $\tau_k < K_\epsilon$ is implied by $\{a_j, b_j\} \not\subseteq [k]$ for all $K_\epsilon \le j \le n$. Since the selection of these roots is independent at each step, we obtain

$$\mathbb{P}(\tau_k < K_\epsilon) \ge \mathbb{P}\bigg(\bigcap_{j=K_\epsilon}^{n} \big\{\{a_j, b_j\} \not\subseteq [k]\big\}\bigg) = \prod_{j=K_\epsilon}^{n} \mathbb{P}(\{a_j, b_j\} \not\subseteq [k]) = \prod_{j=K_\epsilon}^{n} \Big(1 - \frac{k(k-1)}{j(j-1)}\Big).$$

We then bound the product from below to obtain the lower bound

$$\mathbb{P}(\tau_k < K_\epsilon) \ge 1 - \sum_{j=K_\epsilon}^{n} \frac{k^2}{(j-1)^2} \ge 1 - k^2 \int_{K_\epsilon - 2}^{\infty} x^{-2}\,\mathrm{d}x = 1 - \frac{k^2}{K_\epsilon - 2} \ge 1 - \epsilon. \tag{4.10}$$

As a result, $\mathbb{P}(\tau_k \ge K_\epsilon) \le \epsilon$ for all $n \in \mathbb{N}$, from which the tightness follows.

We then prove (4.8) and set $s_n := \lceil (\log n)^2 \rceil$ for ease of writing. Using Bayes' theorem, the bound in (4.10) and that $\mathbb{P}(\mathrm{d}_n(i) \geq d_i, i \in [k]) = 2^{-\sum_{i=1}^k d_i}(1+o(1))$ by Proposition 4.5, we obtain

$$\begin{aligned}\mathbb{P}(\tau_k < s_n \,|\, \mathrm{d}_n(i) \geq d_i, i \in [k]) &= \frac{\mathbb{P}(\tau_k < s_n)}{\mathbb{P}(\mathrm{d}_n(i) \geq d_i, i \in [k])} \mathbb{P}(\mathrm{d}_n(i) \geq d_i, i \in [k] \,|\, \tau_k < s_n) \\ &= \big(1 - \mathcal{O}\big(s_n^{-1}\big)\big) 2^{\sum_{i=1}^k d_i} \mathbb{P}(\mathrm{d}_n(i) \geq d_i, i \in [k] \,|\, \tau_k < s_n)\,.\end{aligned}$$

As in the proof of Lemma 4.4, the event $\{\mathrm{d}_n(i) \geq d_i, i \in [k]\}$ occurs when both $|\mathcal{S}_n(i) \cap [s_n, n]| \geq d_i$ holds and when the first d_i associated coin flips favour vertex i, for all $i \in [k]$. Conditionally on $\{\tau_k < s_n\}$, we know that all these coin flips occurs at different steps for *all* vertices $1, \dots, k$, so that they are independent. Moreover, they are independent of the selection sets, so that we obtain the lower bound

$$\begin{aligned}\mathbb{P}(\tau_k < s_n \,|\, \mathrm{d}_n(i) \geq d_i, i \in [k]) &\geq \big(1 - \mathcal{O}\big(s_n^{-1}\big)\big) \mathbb{P}(|\mathcal{S}_n(i) \cap [s_n, n]| \geq d_i, i \in [k] \,|\, \tau_k < s_n) \\ &= \big(1 - \mathcal{O}\big(s_n^{-1}\big)\big) \mathbb{P}\bigg(\Big| \cup_{i=1}^k \mathcal{S}_n(i) \cap [s_n, n] \Big| \geq \sum_{i=1}^k d_i \,\Big|\, \tau_k < s_n \bigg). \end{aligned} \tag{4.11}$$

Again, the last step uses the conditional event, on which we have that all $\mathcal{S}_n(i) \cap [s_n, n]$ are disjoint, so that $|\mathcal{S}_n(i) \cap [s_n, n]| \geq d_i$ for all $i \in [k]$ is equivalent to the cardinality of the union of all these sets being greater than the sum of the d_i. We also know, conditionally on $\{\tau_k < s_n\}$, that for every $j \in [s_n, n]$, at most one $s_{i,j}$ can equal one among all $i \in [k]$. So, for every $j \in [s_n, n]$,

$$\mathbb{P}\big(\cup_{i=1}^k \{s_{i,j} = 1\} \,|\, \tau_k < s_n\big) = k\mathbb{P}(s_{1,j} = 1 \,|\, \tau_k < s_n) = 2k \frac{j-k}{j(j-1) - k(k-1)} = \frac{2k}{j+k-1}.$$

Hence, if we let $(\widetilde{s}_j)_{j=s_n}^n$ be independent indicator random variables such that $\mathbb{P}(\widetilde{s}_j = 1) = 2k/(j+k-1)$, we can write, conditionally on $\{\tau_k < s_n\}$.

$$\Big| \big(\cup_{i=1}^k \mathcal{S}_n(i) \big) \cap [s_n, n] \Big| \stackrel{d}{=} \sum_{j=s_n}^n \widetilde{s}_j.$$

Since $\log s_n = o(\log n)$, it is readily checked that

$$\mathbb{E}\left[\sum_{j=s_n}^n \widetilde{s}_j \right] = 2k(1+o(1)) \log n, \qquad \mathrm{Var}\Big(\sum_{j=s_n}^n \widetilde{s}_j \Big) = 2k(1+o(1)) \log n.$$

Again using that $d_i \leq c \log n$ for each $i \in [k]$ and all n sufficiently large, where $c < 2$, we obtain for some $\widetilde{c} \in (0, 2-c)$ by using Chebychev's inequality,

$$\begin{aligned}&\mathbb{P}\bigg(\Big| \big(\cup_{i=1}^k \mathcal{S}_n(i) \big) \cap [s_n, n] \Big| \geq \sum_{i=1}^k d_i \,\bigg|\, \tau_k < s_n \bigg) \\ &\quad \geq \mathbb{P}\bigg(\sum_{j=s_n}^n \widetilde{s}_j \geq ck \log n \bigg) \\ &\quad \geq 1 - \mathbb{P}\bigg(\Big(\sum_{j=s_n}^n \widetilde{s}_j - \mathbb{E}\Big[\sum_{j=s_n}^n \widetilde{s}_j \Big] \Big)^2 \geq (\widetilde{c} k \log n)^2 \bigg) \\ &\quad \geq 1 - \frac{1}{(\widetilde{c} k \log n)^2} \mathrm{Var}\Big(\sum_{j=s_n}^n \widetilde{s}_j \Big) = 1 - \mathcal{O}\Big(\frac{1}{\log n} \Big),\end{aligned}$$

which, combined with (4.11), completes the proof of (4.8).

We now prove (4.9) and we set $t_n = \min_{i\in[k]} \log \ell_i$ and note that t_n diverges with n. As in the proof of (4.8),

$$\begin{aligned}
&\mathbb{P}\Big(\tau_k < t_n \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big) \\
&\quad \geq \mathbb{P}\Big(\bigcap_{j=t_n}^{n} \big\{\{a_j, b_j\} \not\subseteq [k]\big\} \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big) \\
&\quad = \prod_{j=t_n}^{n} \mathbb{P}(\{a_j, b_j\} \not\subseteq [k] \,|\, \ell_n(i) = \ell_i, i \in [k]) \\
&\quad \geq \prod_{\substack{j=t_n \\ j\neq \ell_i, i\in[k]}}^{n} \mathbb{P}(\{a_j, b_j\} \not\subseteq [k]) \prod_{j=1}^{k} \mathbb{P}\big(\{a_{\ell_j}, b_{\ell_j}\} \not\subseteq [k] \,|\, \ell_n(i) = \ell_i, i \in [k]\big)\,.
\end{aligned}$$

Here, omitting the conditional event for $j \neq \ell_i$ for any $i \in [k]$ yields a lower bound. Indeed, for any two distinct $i, i' \in [k]$, if $j > \max\{\ell_i, \ell_{i'}\}$ then $\{a_j, b_j\} = \{i, i'\}$ cannot occur conditionally on $\ell_n(i) = \ell_i$. Furthermore, we isolate the steps $\ell_1, \dots, \ell_k$, since the conditional event prescribes that vertex i is selected at step ℓ_i. For any $j \in [k]$,

$$\begin{aligned}
\mathbb{P}\big(\{a_{\ell_j}, b_{\ell_j}\} \not\subseteq [k] \,|\, \ell_n(i) = \ell_i, i \in [k]\big) &= 1 - \mathbb{P}\big(\{a_{\ell_j}, b_{\ell_j}\} \subseteq [k] \,|\, \{a_{\ell_j} = j\} \cup \{b_{\ell_j} = j\}\big) \\
&= 1 - \frac{k-1}{\ell_j - 1}.
\end{aligned}$$

As a result, we obtain

$$\begin{aligned}
\mathbb{P}\Big(\tau_k < t_n \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big) &\geq \prod_{\substack{j=t_n \\ j\neq \ell_i, i\in[k]}}^{m} \Big(1 - \frac{k(k-1)}{j(j-1)}\Big) \prod_{j=1}^{k} \Big(1 - \frac{k-1}{\ell_j - 1}\Big) \\
&\geq \prod_{\substack{j=t_n \\ j\neq \ell_i, i\in[k]}}^{m} \Big(1 - \frac{k(k-1)}{(j-1)^2}\Big)\Big(1 - \frac{k-1}{t_n}\Big)^k \\
&= \prod_{j=t_n}^{n} \Big(1 - \frac{k(k-1)}{(j-1)^2}\Big) \prod_{j=1}^{k} \Big(1 - \frac{k(k-1)}{(\ell_j - 1)^2}\Big)^{-1} \Big(1 - \frac{k-1}{t_n}\Big)^k \\
&\geq \prod_{j=t_n}^{n} \Big(1 - \frac{k(k-1)}{(j-1)^2}\Big)\Big(1 - \frac{k(k-1)}{(n-1)^2}\Big)^{-k} \Big(1 - \frac{k-1}{t_n}\Big)^k \\
&\geq 1 - \mathcal{O}\big(t_n^{-1}\big),
\end{aligned}$$

where the last step follows from (4.10), and which concludes the proof. □

Beyond the sets $\mathcal{A}_{\bar{d}}$ and $\mathcal{B}_{n,\delta}$ and the random variable τ_k, we also want to control of the probability of the events $\{\ell_n(i) = \ell_i, i \in [k]\}$ and $\{\mathrm{d}_n(i) \geq d_i, i \in [k]\}$ conditionally on the truncated selection sets $\overline{S}_{n,1}$. To this end, we define, for $\bar{\ell} := (\ell_i)_{\in[k]} \in \mathbb{N}^k$,

$$\begin{aligned}
\widetilde{\mathcal{A}}_{\bar{\ell}} &:= \{\bar{J} \in \Omega_1^k : \mathbb{P}\big(\ell_n(i) = \ell_i, i \in [k], \overline{S}_{n,1} = \bar{J}\big) > 0\}, \\
\widetilde{\mathcal{B}}_n &:= \{\bar{J} \in \Omega_1^k : (J_1, \dots, J_k) \text{ are pairwise disjoint}\}.
\end{aligned}$$

We then have the following lemma, which is partially covered by [12, Lemma 3.1].

Lemma 4.10. *Fix $k \in \mathbb{N}$ and let $(\ell_i)_{i\in[k]} \in [n]^k$ such that $\ell_i \neq \ell_j$ when $i \neq j$. Then,*

$$\mathbb{P}(\ell_n(i) = \ell_i, i \in [k]) = \frac{1}{(n)_k}. \tag{4.12}$$

Also, when the truncation sequence t_n diverges with n,

$$\mathbb{P}\Big(\ell_n(i) = \ell_i, i \in [k], \overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c\Big) = o(n^{-k}). \tag{4.13}$$

Finally, let $(d_i)_{i\in[k]}, \mathbb{N}_0^k$ and let $t_n = \lceil (\log n)^2 \rceil$. If $\bar{J} \in \mathcal{A}_{\bar{d}}$,

$$\mathbb{P}\big(\mathrm{d}_n(i) \geq d_i, i \in [k] \,|\, \overline{\mathcal{S}}_{n,1} = \bar{J}\big) = 2^{-\sum_{i=1}^k d_i}. \tag{4.14}$$

Proof. The first result in (4.12) follows from Corollary 3.4, as each vertex obtains a uniform label from $[n]$ after the relabelling of the final tree F_1 in the Kingman n-coalescent and all ℓ_i are distinct.

To prove (4.13), we write

$$\begin{aligned}\mathbb{P}\Big(\ell_n(i) = \ell_i, i \in [k], \overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c\Big) &= \mathbb{P}\Big(\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big)\, \mathbb{P}(\ell_n(i) = \ell_i, i \in [k]) \\ &= \mathbb{P}\Big(\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big)\, \frac{1}{(n)_k},\end{aligned}$$

where the last step follows from (4.12). It thus remains to argue to that probability on the right-hand side is $o(1)$. For $\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c$ to hold, the truncated selection sets should overlap at some step $t_n \leq j \leq n$, i.e. $\tau_k \geq t_n$ should hold. Conditionally on the event $\{\ell_n(i) = \ell_i, i \in [k]\}$, however, the truncated selection sets in $\overline{\mathcal{S}}_{n,1}$ cannot overlap at certain steps. Namely, for $j > \max_{i\in[k]} \ell_i$, $j \in \mathcal{S}_{n,1}(i)$ can hold for at most one $i \in [k]$. Indeed, if the converse would be the case, i.e. $j \in \mathcal{S}_{n,1}(i)$ and $j \in \mathcal{S}_{n,1}(i')$ for some distinct $i, i' \in [k]$, then one of the vertices i, i', let us assume this is vertex i, would lose the associated coin flip at step j and hence its label in the random recursive tree would be $j > \ell_i$. This clearly contradicts the conditional event. As a result, on the conditional event $\{\ell_n(i) = \ell_i, i \in [k]\}$, the event $\{\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c\}$ implies that $\{t_n \leq \tau_k \leq \max_{i\in[k]} \ell_i\}$ holds. Hence, by Lemma 4.9 and since t_n diverges with n,

$$\mathbb{P}\Big(\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big) \leq \mathbb{P}\Big(t_n \leq \tau_k \leq \max_{i\in[k]} \ell_i\Big) \leq \mathbb{P}(\tau_k \geq t_n) = o(1),$$

as desired.

The final result in (4.14) is proved in [12, Lemma 3.2]. □

In Lemma 4.4 we saw that, as long as the truncated selection sets $(\mathcal{S}_{n,1}(i))_{i\in[k]}$ are pairwise disjoint, then the events $\{h_{n,1}(i) \leq h_i, \ell_n(i) \geq \ell_i, \mathrm{d}_n(i) \geq d_i\}$, $i \in [k]$, are independent, conditionally on $\overline{\mathcal{S}}_{n,1}$. Furthermore, when the truncation sequence t_n diverges as $n \to \infty$, we already observed that the event $\{\tau_k < t_n\}$ holds with high probability by Lemma 4.9, which implies that the $(\mathcal{S}_{n,1}(i))_{i\in[k]}$ are disjoint.

On the other hand, we use the truncated depths $(h_{n,1}(i))_{i\in[k]}$ merely for technical reasons, and are really interested in the depths $(h_n(i))_{i\in[k]}$. As a result, choosing a truncation sequence $(t_n)_{n\in\mathbb{N}}$ that diverges with n 'too quickly', may lead to different behaviour of $h_{n,1}(1)$ compared to $h_n(1)$. In other words, if t_n grows 'too quickly', then $h_{n,2}(1) = h_n(1) - h_{n,1}(1)$ might become 'too large'. In the following lemma we make this informal statement more precise and provide constraints on t_n to avoid such discrepancies between $h_n(1)$ and $h_{n,1}(1)$.

Lemma 4.11 (Partially from Lemma 2.7, [12])**.** *Fix $k \in \mathbb{N}$ and $c \in (0, 2)$. If $d_i \leq c \log n$ for all $i \in [k]$ and $t_n = \lceil (\log n)^2 \rceil$, then for any $j \in [k]$ and any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\Big(h_{n,2}(j) \geq \epsilon\sqrt{\log n} \,\Big|\, \mathrm{d}_n(i) \geq d_i, i \in [k]\Big) = 0. \tag{4.15}$$

Furthermore, let $(\ell_i)_{i\in[k]} \in [n]^k$ be k distinct integers that diverge as $n \to \infty$. If $\log t_n = o(\min_{i\in[k]} \sqrt{\log \ell_i})$, then for any $j \in [k]$ and any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(h_{n,2}(j) \geq \epsilon\sqrt{\log \ell_j} \,\Big|\, \ell_n(i) = \ell_i, i \in [k]\Big) = 0. \tag{4.16}$$

Remark 4.12. Assume the truncation sequence $(t_n)_{n\in\mathbb{N}}$ satisfies the assumptions of Lemma 4.11. As a direct consequence of Lemma 4.11, the limiting distributions of

$$\frac{h_n(j)-(\log n-d_j/2)}{\sqrt{\log n-d_j/4}} \quad \text{and} \quad \frac{h_{n,1}(j)-(\log n-d_j/2)}{\sqrt{\log n-d_j/4}}$$

conditionally on the event $\mathrm{d}_n(i)\geq d_i$ for al $i\in[k]$, with $d_i\leq c\log n$ for all $i\in[k]$, are identical (assuming they both exist), for any $j\in[k]$. This follows from Slutsky's theorem [35, Lemma 2.8] and since $\sqrt{\log n-d_j/4}=\Theta(\sqrt{\log n})$. Similarly, conditionally on $\ell_n(i)=\ell_i, i\in[k]$ (where the ℓ_i diverge as $n\to\infty$), the limiting distributions of

$$\frac{h_n(j)-\log \ell_j}{\sqrt{\log \ell_j}} \quad \text{and} \quad \frac{h_{n,1}(j)-\log \ell_j}{\sqrt{\log \ell_j}}$$

are identical (assuming they exist), for any $j\in[k]$.

Proof. The result in (4.15) follows from [12, Lemma 2.7].

To prove (4.16), we consider $j=1$ only by the exchangeability of the vertices. We first note that $t_n\leq \min_{i\in[k]}\ell_i$ by the assumption on t_n and since the ℓ_i diverge with n. As a result, the event $\{\ell_n(i)=\ell_i, i\in[k]\}$ is solely dependent on the truncated selection sets $\overline{\mathcal{S}}_{n,1}$ and the associated coin flips of the truncated selection sets, whereas $h_{n,2}(1)$ is determined by the set $\mathcal{S}_n(1)\cap[2,t_n-1]$ and its associated coin flips. It thus follows that $h_{n,2}(1)$ is independent of the event $\{\ell_n(i)=\ell_i, i\in[k]\}$. The result then follows from the Markov inequality and by the assumption on t_n, as

$$\mathbb{P}(h_{n,2}(1)\geq \epsilon\sqrt{\log \ell_1})\leq \frac{\mathbb{E}\left[h_{n,2}(1)\right]}{\epsilon\sqrt{\log \ell_1}}=\frac{1}{\epsilon\sqrt{\log \ell_1}}\sum_{j=2}^{t_n-1}\frac{1}{j}=\mathcal{O}\Big(\frac{\log t_n}{\sqrt{\log \ell_1}}\Big)=o(1),$$

by the assumptions on t_n, which concludes the proof. □

5 Joint properties of high-degree vertices

In this section we use the preliminary results proved in Section 4 to study the joint behaviour of the depth and label of high-degree vertices.

We set

$$h:=(\log n-d/2)+y\sqrt{\log n-d/4},\quad \ell:=n\exp(-d/2+x\sqrt{d/4}),\quad t_n:=\lceil(\log n)^2\rceil, \tag{5.1}$$

with $x,y\in\mathbb{R}$. We then have the following result.

Proposition 5.1. *Fix $a\in[0,2)$, let h,ℓ and t_n be as in* (5.1)*, with $d\in\mathbb{N}_0$, and let M and N be two independent standard normal random variables. When $\limsup_{n\to\infty} d<\infty$,*

$$\lim_{n\to\infty}\mathbb{P}(h_{n,1}(1)\leq h\,|\,\mathrm{d}_n(1)\geq d)=\Phi(y). \tag{5.2}$$

When, instead, d diverges as $n\to\infty$ such that $\lim_{n\to\infty} d/\log n=a$,

$$\lim_{n\to\infty}\mathbb{P}(h_{n,1}(1)\leq h,\ell_n(1)\geq \ell\,|\,\mathrm{d}_n(1)\geq d)=\mathbb{P}\left(M\sqrt{\frac{a}{4-a}}+N\sqrt{1-\frac{a}{4-a}}\leq y, M>x\right). \tag{5.3}$$

Remark 5.2. (i) As $M\sqrt{a/(4-a)}+N\sqrt{1-a/(4-a)}\sim\mathcal{N}(0,1)$, the result in (5.3), when omitting the event $\ell_n(1)\geq\ell$ (or, equivalently, letting $x\to-\infty$ and setting $a=0$), yields

$$\lim_{n\to\infty}\mathbb{P}(h_{n,1}(1)\leq h\,|\,\mathrm{d}_n(1)\geq d)=\Phi(y),$$

and hence complements the result in (5.2) in the case that d diverges with n such that $d \leq c \log n$ for some $c \in (0,2)$. Together, they are a generalisation of [12, Lemma 2.5], where only a parametrised version with $d = \lfloor a \log n \rfloor + b$ and $a \in [0,2), b \in \mathbb{Z}$ is considered.

(ii) Combined with Lemma 4.11 and Remark 4.12, we obtain that the results in Proposition 5.1 hold when we substitute $h_n(1)$ for $h_{n,1}(1)$ as well.

Proof. We first prove (5.3) and briefly discuss how to prove (5.2) using [12, Lemma 2.5] at the end. In the setting of (5.3), we recall that we assume that d diverges as $n \to \infty$ and h, ℓ and t_n are as in (5.1).

Take $c \in (a,2)$. By Lemma 4.1,

$$\begin{aligned}
&\mathbb{P}(h_{n,1} \leq h, \ell_n(1) \geq \ell \,|\, \mathrm{d}_n(1) \geq d) \\
&\quad = \frac{\mathbb{P}(h_{n,1} \leq h, \ell_n(1) \geq \ell, \mathrm{d}_n(1) \geq d)}{\mathbb{P}(\mathrm{d}_n(1) \geq d)} \\
&\quad = \frac{1}{\mathbb{P}(\mathrm{d}_n(1) \geq d)\, 2^d} \mathbb{E}\left[\mathbb{1}_{\{|[\ell,n] \cap \mathcal{S}_{n,1}(1)| \geq d+1\}} \mathbb{P}(X_{n,\ell,1} + X_{n,\ell,2} \leq h, X_{n,\ell,1} \geq 1 \,|\, \mathcal{S}_{n,1}(1))\right],
\end{aligned}$$

where we recall that, conditionally on $\mathcal{S}_{n,1}(1)$, $X_{n,\ell,1} \sim \mathrm{Bin}(|[\ell,n] \cap \mathcal{S}_{n,1}(1)| - d, 1/2)$ and $X_{n,\ell,2} \sim \mathrm{Bin}(|[t_n, \ell-1] \cap \mathcal{S}_{n,1}(1)|, 1/2)$ (where we set $X_{n,\ell,1} = 0, X_{n,\ell,2} = 0$ when $|[\ell,n] \cap \mathcal{S}_{n,1}(1)| - d \leq 0$ and $|[t_n, \ell-1] \cap \mathcal{S}_{n,1}(1)| = 0$, respectively). We observe that, since d diverges with n and $d \leq c \log n$ (with $c \in (a,2)$), $\ell = n \exp(-d/2 + x\sqrt{d/4}) > \lceil (\log n)^2 \rceil = t_n$ for all n large. As a result, $X_{n,\ell,2}$ is non-zero with positive probability.

Since $\mathbb{P}(\mathrm{d}_n(1) \geq d)\, 2^d = 1 + o(1)$ by Proposition 4.5, we obtain

$$\begin{aligned}
&\mathbb{P}(h_{n,1} \leq h, \ell_n(1) \geq \ell \,|\, \mathrm{d}_n(1) \geq d) \\
&\quad = (1 + o(1)) \mathbb{E}\left[\mathbb{1}_{\{|[\ell,n] \cap \mathcal{S}_{n,1}(1)| \geq d+1\}} \mathbb{P}(X_{n,\ell,1} + X_{n,\ell,2} \leq h, X_{n,\ell,1} \geq 1 \,|\, \mathcal{S}_{n,1}(1))\right].
\end{aligned}$$

To prove the expected value has the desired limit, we start by rewriting the binomial random variables $X_{n,\ell,1}$ and $X_{n,\ell,2}$. Let $(I_j^n)_{j \in [n], n \in \mathbb{N}}$, $(\widetilde{I}_j^n)_{j \in [n], n \in \mathbb{N}}$ be two i.i.d. sequences of independent Bernoulli$(1/2)$ random variables. Finally, let $Q_n := |[\ell,n] \cap \mathcal{S}_{n,1}(1)|$, $\widetilde{Q}_n := |[t_n, \ell-1] \cap \mathcal{S}_{n,1}(1)| = S_{n,1}(1) - Q_n$, independent of the $I_j^n, \widetilde{I}_j^n$. Then,

$$X_{n,\ell,1} := \sum_{j=1}^{Q_n - d} I_j^{Q_n - d}, \qquad X_{n,\ell,2} := \sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n}.$$

Here, we set $X_{n,\ell,1} = 0, X_{n,\ell,2} = 0$ if $Q_n - d \leq 0, \widetilde{Q}_n = 0$, respectively. Notice that Q_n and $\widetilde{Q}_n$ are independent, that they can be determined from $\mathcal{S}_{n,1}(1)$ and that the values of the $I_j^n, \widetilde{I}_j^n$ are independent of $\mathcal{S}_{n,1}(1)$, so that conditioning on $\mathcal{S}_{n,1}(1)$ is equivalent to conditioning on $Q_n, \widetilde{Q}_n$. We can then write the expected value in the statement of the proposition as

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{Q_n \geq d+1\}} \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + \sum_{j=1}^{Q_n - d} I_j^{Q_n - d} \leq h, \sum_{j=1}^{Q_n - d} I_i^{Q_n - d} \geq 1 \,\middle|\, Q_n, \widetilde{Q}_n\right)\right] \\
&\quad = \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + \sum_{j=1}^{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}} I_j^{Q_n - d} \leq h, \sum_{j=1}^{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}} I_j^{Q_n - d} \geq 1\right).
\end{aligned}$$

The second line follows from the fact that, by changing the upper limits of the second and third sum in the probability on the first line from $Q_n - d$ to $(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}$, we can remove the indicator in the expected value. Indeed, if $Q_n \leq d$, then $\mathbb{1}_{\{Q_n - d \geq 1\}} = 0$, and hence the second event in the probability cannot occur almost surely, so that the

probability is zero. As a result, the indicator in the expected value is redundant. We thus obtain

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + \sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} \leq h\right) - \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} \leq h, \sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} = 0\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + \sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} \leq h\right) \\
&\quad - \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} \leq h\right)\mathbb{P}\left(\sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} = 0\right),
\end{aligned} \tag{5.4}$$

where the second step follows from the independence of the two sums in the second probability on the first line. The event

$$\left\{\sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} = 0\right\}$$

occurs either when $Q_n \leq d$ or when, given $Q_n \geq d+1$, $I_1^{Q_n-d} = \ldots = I_{Q_n-d}^{Q_n-d} = 0$. Hence,

$$\mathbb{P}\left(\sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} = 0\right) = \mathbb{P}(Q_n \leq d) + \mathbb{E}\left[\mathbb{1}_{\{Q_n\geq d+1\}} 2^{-(Q_n-d)}\right].$$

Combining this with (5.4) yields

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + \sum_{j=1}^{(Q_n-d)\mathbb{1}_{\{Q_n-d\geq 1\}}} I_j^{Q_n-d} \leq h\right) - \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} \leq h\right)\mathbb{P}(Q_n \leq d) \\
&\quad + \mathcal{O}\Big(\mathbb{E}\left[\mathbb{1}_{\{Q_n\geq d+1\}} 2^{-(Q_n-d)}\right]\Big).
\end{aligned} \tag{5.5}$$

What remains is to show that the first two terms yield the desired limit and that the last term is negligible compared to the first two. Let us start with the former and tackle the product of two probabilities on the first line. It follows from Lindeberg's conditions [11, Theorem 3.4.5] that

$$\frac{Q_n - \mathbb{E}[Q_n]}{\sqrt{\mathrm{Var}(Q_n)}} \xrightarrow{d} N, \qquad \frac{\widetilde{Q}_n - \mathbb{E}[\widetilde{Q}_n]}{\sqrt{\mathrm{Var}(\widetilde{Q}_n)}} \xrightarrow{d} \widetilde{N}, \tag{5.6}$$

with $N, \widetilde{N} \sim \mathcal{N}(0,1)$ independent standard normal random variables, as we recall that Q_n and $\widetilde{Q}_n$ are sums of independent Bernoulli random variables. It is readily checked that by the choice of ℓ in (5.1) and since d diverges with n,

$$\begin{aligned}
\mathbb{E}[Q_n] &= \sum_{j=\ell}^{n} \frac{2}{j} = 2\log(n/\ell) + \mathcal{O}(1) = d - x\sqrt{d}(1+o(1)), \\
\mathrm{Var}(Q_n) &= \sum_{j=\ell}^{n} \frac{2}{j}\Big(1 - \frac{2}{j}\Big) = d - x\sqrt{d}(1+o(1)),
\end{aligned} \tag{5.7}$$

and, by the choice of ℓ, d and t_n,

$$\begin{aligned}\mathbb{E}[\widetilde{Q}_n] &= \sum_{j=t_n}^{\ell-1} \frac{2}{j} = 2\log n - d + x\sqrt{d} - (1+o(1))\log\log n,\\ \mathrm{Var}(\widetilde{Q}_n) &= \sum_{j=t_n}^{\ell-1} \frac{2}{j}\Big(1-\frac{2}{j}\Big) = 2\log n - d + x\sqrt{d} - (1+o(1))\log\log n.\end{aligned} \tag{5.8}$$

By (5.6) and (5.7) we thus obtain that

$$\mathbb{P}(Q_n \le d) = \mathbb{P}\left(\frac{Q_n - \mathbb{E}\left[Q_n\right]}{\sqrt{\mathrm{Var}(Q_n)}} \le \frac{d - \mathbb{E}\left[Q_n\right]}{\sqrt{\mathrm{Var}(Q_n)}}\right) = \mathbb{P}\left(\frac{Q_n - \mathbb{E}\left[Q_n\right]}{\sqrt{\mathrm{Var}(Q_n)}} \le \frac{x\sqrt{d}(1+o(1))}{\sqrt{d}(1+o(1))}\right), \tag{5.9}$$

which converges to $\Phi(x)$, where we recall that $\Phi : \mathbb{R} \to (0,1)$ denotes the cumulative density function of a standard normal distribution. By Skorokhod's representation theorem [5, Theorem 6.7] there exists a probability space and a coupling of $(Q_n)_{n\in\mathbb{N}}, (\widetilde{Q}_n)_{n\in\mathbb{N}}$ and $(I_j^n)_{j\in[n],n\in\mathbb{N}}, (\widetilde{I}_j^n)_{j\in[n],n\in\mathbb{N}}$ such that the collections $(I_j^n)_{j\in\mathbb{N}}, (\widetilde{I}_j^n)_{j\in\mathbb{N}}$ are independent of Q_n and $\widetilde{Q}_n$ and the convergence in (5.6) is almost sure rather than in distribution. In particular, $Q_n/d \xrightarrow{a.s.} 1$, $\widetilde{Q}_n/(2\log n - d) \xrightarrow{a.s.} 1$ and $Q_n, \widetilde{Q}_n \xrightarrow{a.s.} \infty$. Moreover, it also follows from this representation that

$$\frac{2\sum_{j=1}^n I_j^n - n}{\sqrt{n}} \xrightarrow{a.s.} N', \qquad \frac{2\sum_{j=1}^n \widetilde{I}_j^n - n}{\sqrt{n}} \xrightarrow{a.s.} N'',$$

as $n \to \infty$ as well, where N', N'' are independent standard normal random variables, also independent of $N, \widetilde{N}$ in (5.6). We then rewrite

$$\begin{aligned}\frac{2\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} - (2\log n - d)}{\sqrt{4\log n - d}} &= \frac{2\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} - \widetilde{Q}_n}{\sqrt{\widetilde{Q}_n}}\sqrt{\frac{\widetilde{Q}_n}{2\log n - d}}\sqrt{\frac{2\log n - d}{4\log n - d}}\\ &\quad + \frac{\widetilde{Q}_n - \mathbb{E}[\widetilde{Q}_n]}{\sqrt{\mathrm{Var}(\widetilde{Q}_n)}}\sqrt{\frac{\mathrm{Var}(\widetilde{Q}_n)}{2\log n - d}}\sqrt{\frac{2\log n - d}{4\log n - d}}\\ &\quad + \frac{\mathbb{E}[\widetilde{Q}_n] - (2\log n - d)}{\sqrt{d}}\sqrt{\frac{d}{4\log n - d}}.\end{aligned} \tag{5.10}$$

Combining this with the Skorokhod representation, the fact that $d/\log n \to a$ and (5.8), yields

$$\frac{2\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} - (2\log n - d)}{\sqrt{4\log n - d}} \xrightarrow{d} N_1\sqrt{\frac{2-a}{4-a}} + N_2\sqrt{\frac{2-a}{4-a}} + x\sqrt{\frac{a}{4-a}},$$

where N_1, N_2 are independent standard normal random variables. Combining this with (5.9) and using that $h = \log n - d/2 + y\sqrt{\log n - d/4}$, we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(Q_n \le d)\,\mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} \le h\right) &= \Phi(x)\mathbb{P}\left(N_1\sqrt{\frac{2-a}{4-a}} + N_2\sqrt{\frac{2-a}{4-a}} + x\sqrt{\frac{a}{4-a}} \le y\right)\\ &= \Phi(x)\mathbb{P}\left(N\sqrt{1-\frac{a}{4-a}} + x\sqrt{\frac{a}{4-a}} \le y\right),\end{aligned} \tag{5.11}$$

where N is again a standard normal random variable. This deals with the second term of (5.5). For the first term, we observe that

$$\mathbb{P}\left(\frac{(Q_n - d)\mathbb{1}_{\{Q_n - d \ge 1\}}}{\sqrt{d}} = 0\right) = \mathbb{P}(Q_n \le d) \to \Phi(x),$$

as $n \to \infty$ by (5.9), and similarly for $z \geq 0$,

$$\mathbb{P}\left(\frac{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}}{\sqrt{d}} > z \right) = \mathbb{P}\left(\frac{Q_n - \mathbb{E}\left[Q_n\right]}{\sqrt{\mathrm{Var}(Q_n)}} > \frac{d - \mathbb{E}\left[Q_n\right] + z\sqrt{d}}{\sqrt{\mathrm{Var}(Q_n)}} \right) \to 1 - \Phi(x+z),$$

as $n \to \infty$. Hence, for $x \in \mathbb{R}$ fixed, let us define a random variable $M_x := \mathbb{1}_{\{M > x\}}(M - x)$, where M is a standard normal random variable. It then follows that $\mathbb{P}(M_x = 0) = \Phi(x)$, $\mathbb{P}(M_x > z) = \mathbb{P}(M > x + z) = 1 - \Phi(x+z), z > 0$, so that

$$\frac{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}}{\sqrt{d}} \xrightarrow{d} M_x. \tag{5.12}$$

By the independence of the Bernoulli random variables $I_j^n, \widetilde{I}_j^n$, we can relabel them as a sequence of i.i.d. random variables. If we set $O_n := \widetilde{Q}_n + (Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}$, then we can write them as $(\hat{I}_j^{O_n})_{j \in [O_n]}$, with $\hat{I}_j^{O_n} := \widetilde{I}_j^{\widetilde{Q}_n}$ if $1 \leq j \leq \widetilde{Q}_n$ and $\hat{I}_j^{O_n} := I_{j - \widetilde{Q}_n}^{Q_n - d}$ if $\widetilde{Q}_n + 1 \leq j \leq \widetilde{Q}_n + (Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}$. Again following Lindeberg's conditions, we find that

$$\frac{2\sum_{j=1}^n \hat{I}_j^n - n}{\sqrt{n}} \xrightarrow{d} N',$$

where N' is a standard normal random variable. Moreover, $O_n/(2\log n - d) \xrightarrow{\mathbb{P}} 1$ by combining (5.6), (5.8) and (5.12). We can then write

$$\begin{aligned}
&\frac{2\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + 2\sum_{j=1}^{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}} I_j^{Q_n - d} - (2\log n - d)}{\sqrt{4\log n - d}} \\
&= \frac{2\sum_{j=1}^{O_n} \hat{I}_j^{O_n} - O_n}{\sqrt{O_n}} \sqrt{\frac{O_n}{2\log n - d}} \sqrt{\frac{2\log n - d}{4\log n - d}} + \frac{\widetilde{Q}_n - \mathbb{E}[\widetilde{Q}_n]}{\sqrt{\mathrm{Var}(\widetilde{Q}_n)}} \sqrt{\frac{\mathrm{Var}(\widetilde{Q}_n)}{2\log n - d}} \sqrt{\frac{2\log n - d}{4\log n - d}} \\
&\quad + \frac{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}}{\sqrt{d}} \sqrt{\frac{d}{4\log n - d}} + \frac{\mathbb{E}[\widetilde{Q}_n] - (2\log n - d)}{\sqrt{d}} \sqrt{\frac{d}{4\log n - d}}.
\end{aligned}$$

If we let N, N', N'' be i.i.d. standard normal random variables, independent of M_x, and use similar steps as in (5.12) and (5.10) (in particular using the Skorokhod representation for the random variables $(\hat{I}_j^n)_{j \in [n]}, O_n, (Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}$ and that $d/\log n \to a$), this converges in distribution to

$$\begin{aligned}
&N' \sqrt{\frac{2-a}{4-a}} + N'' \sqrt{\frac{2-a}{4-a}} + M_x \sqrt{\frac{a}{4-a}} + x \sqrt{\frac{a}{4-a}} \\
&\stackrel{d}{=} N \sqrt{1 - \frac{a}{4-a}} + M_x \sqrt{\frac{a}{4-a}} + x \sqrt{\frac{a}{4-a}},
\end{aligned}$$

Combining this with (5.11) in (5.5) yields

$$\begin{aligned}
&\lim_{n \to \infty} \left[\mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} + \sum_{j=1}^{(Q_n - d)\mathbb{1}_{\{Q_n - d \geq 1\}}} I_j^{Q_n - d} \leq h \right) - \mathbb{P}(Q_n \leq d)\, \mathbb{P}\left(\sum_{j=1}^{\widetilde{Q}_n} \widetilde{I}_j^{\widetilde{Q}_n} \leq h \right) \right] \\
&= \mathbb{P}\left(M_x \sqrt{\frac{a}{4-a}} + x \sqrt{\frac{a}{4-a}} + N \sqrt{1 - \frac{a}{4-a}} \leq y \right) \\
&\quad - \Phi(x) \mathbb{P}\left(N \sqrt{1 - \frac{a}{4-a}} + x \sqrt{\frac{a}{4-a}} \leq y \right).
\end{aligned} \tag{5.13}$$

By intersecting the event in the first probability on the right-hand side with the events $\{M_x = 0\}, \{M_x > 0\}$, and using that M_x is independent of N, we arrive at

$$\mathbb{P}\left(M_x \sqrt{\frac{a}{4-a}} + x\sqrt{\frac{a}{4-a}} + N\sqrt{1-\frac{a}{4-a}} \le y, M_x > 0 \right).$$

By the definition of M_x, it follows that the event $\{M_x > 0\}$ is equivalent to $\{M > x\}$, where we recall that M is a standard normal random variable. Moreover, on the event $\{M_x > 0\} = \{M > x\}$, $M_x + x = \mathbb{1}_{\{M>x\}}(M - x) + x = M$. Thus, we obtain

$$\mathbb{P}\left(M\sqrt{\frac{a}{4-a}} + N\sqrt{1-\frac{a}{4-a}} \le y, M > x \right), \tag{5.14}$$

as desired. Finally, we show that

$$\lim_{n\to\infty} \mathbb{E}\left[\mathbb{1}_{\{Q_n \ge d+1\}} 2^{-(Q_n - d)} \right] = 0. \tag{5.15}$$

By splitting the expected value into the cases where Q_n is at most $d + 1 + \lfloor d^{1/2-\eta} \rfloor$ and at least $d + 1 + \lceil d^{1/2-\eta} \rceil$, respectively, for some $\eta \in (0, 1/2)$, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{\{Q_n \ge d+1\}} 2^{-(Q_n - d)} \Big] &= \sum_{m=d+1}^{d+1+\lfloor d^{1/2-\eta} \rfloor} \mathbb{P}(Q_n = m)\, 2^{-(m-d)} + \sum_{m \ge d+1+\lceil d^{1/2-\eta} \rceil} \mathbb{P}(Q_n = m)\, 2^{-(m-d)} \\
&\le \sum_{m=d+1}^{d+1+\lfloor d^{1/2-\eta} \rfloor} \mathbb{P}(Q_n = m) + \sum_{m \ge d+1+\lceil d^{1/2-\eta} \rceil} \mathbb{P}(Q_n = m)\, 2^{-d^{1/2-\eta}} \\
&\le \mathbb{P}\Big(d+1 \le Q_n \le d+1+\lfloor d^{1/2-\eta} \rfloor \Big) + 2^{-d^{1/2-\eta}}.
\end{aligned}$$

Since $d^{1/2-\eta} = o\big(\sqrt{\mathrm{Var}(Q_n)}\big)$ (see (5.7)), it follows from (5.6) that the probability in the last line converges to zero. This proves (5.15), and combining this with the limit (5.14) of the left-hand side of (5.13) in (5.5) yields the desired result and concludes the proof of (5.3).

We now discuss the the proof of (5.2). We recall that now $L := \limsup_{n\to\infty} d < \infty$. Also, conditionally on $\mathcal{S}_{n,1}(1)$, let $X_n = X_n(d) \sim \mathrm{Bin}(|\mathcal{S}_{n,1}(1)| - d, 1/2)$ (where we set $X_n = 0$ when $|\mathcal{S}_{n,1}(1)| - d \le 0$) and let us define $h' := \log n + y\sqrt{\log n}$. Note that, since $L < \infty$, $(h - h')/\sqrt{\log n} = o(1)$, so that using h' instead of h yields the same result. Again using Lemma 4.1 and Proposition 4.5, we obtain

$$\begin{aligned}
\mathbb{P}(h_{n,1}(1) > h' \,|\, \mathrm{d}_n(1) \ge d) &= \frac{\mathbb{P}(h_{n,1}(1) > h', \mathrm{d}_n(1) \ge d)}{\mathbb{P}(\mathrm{d}_n(1) \ge d)} \\
&= (1 + o(1))\mathbb{E}\big[\mathbb{1}_{\{|\mathcal{S}_{n,1}(1)| \ge d\}} \mathbb{P}(X_n > h' \,|\, S_{n,1}(1)) \big].
\end{aligned}$$

Notice that, for any $\mathcal{S}_{n,1}(1) \subseteq \Omega_1$, both the indicator as well as the probability are decreasing functions of d. As a result, we can bound the expected value from above by setting $d = 0$ in the indicator and using $X_n(0)$ in the probability. The upper bound has the desired limit by [12, Lemma 2.5] with $a = b = 0$. Similarly, we can bound the expected value from below by setting $d = L$ in the indicator and using $X_n(L)$ in the probability. The result then follows from [12, Lemma 2.5] with $a = 0, b = L$, which yields a matching lower bound. □

To finish this section, we use the results related to the truncated selection sets developed in Section 4 to extend Proposition 5.1 to the case of multiple vertices.

Proposition 5.3. *Fix $k \in \mathbb{N}, (a_i)_{i\in[k]} \in [0,2)^k$. Let $(d_i)_{i\in[k]}$ be k integer-valued sequences such that, for all $i \in [k]$, $\lim_{n\to\infty} d_i/\log n = a_i$. Let $\ell_i := n\exp(-d_i/2 + x_i\sqrt{d_i/4})$ and $h_i := (\log n - d_i/2) + y_i\sqrt{\log n - d_i/4}$ with $(x_i)_{i\in[k]}, (y_i)_{i\in[k]} \in \mathbb{R}^k$, and set $t_n = \lceil(\log n)^2\rceil$. Then,*

$$\lim_{n\to\infty} \mathbb{P}(h_{n,1}(i) \le h_i, i \in [k] \,|\, \mathrm{d}_n(i) \ge d_i, i \in [k]) = \prod_{i=1}^{k} \Phi(y_i). \tag{5.16}$$

If, additionally, d_i diverges as $n \to \infty$ for all $i \in [k]$, let M and N be independent standard normal random variables. Then,

$$\begin{aligned}\lim_{n\to\infty} &\mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, i \in [k] \,|\, \mathrm{d}_n(i) \ge d_i, i \in [k]) \\ &= \prod_{i=1}^{k} \mathbb{P}\left(M\sqrt{\frac{a_i}{4-a_i}} + N\sqrt{1-\frac{a_i}{4-a_i}} \le y_i, M > x_i\right).\end{aligned} \tag{5.17}$$

Remark 5.4. As is the case in Remark 5.2, it follows from Lemma 4.11 and Remark 4.12 that the result in Proposition 5.3 holds when substituting $h_n(i)$ for $h_{n,1}(i)$ as well.

Proof. We provide a proof for (5.17), the proof of (5.16) uses the same steps.

It suffices to prove that

$$\begin{aligned}&\mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, \mathrm{d}_n(i) \ge d_i, i \in [k]) \\ &\quad = (1+o(1))2^{-\sum_{i=1}^k d_i} \prod_{i=1}^{k} \mathbb{P}\left(M\sqrt{\frac{a_i}{4-a_i}} + N\sqrt{1-\frac{a_i}{4-a_i}} \le y_i, M > x_i\right),\end{aligned}$$

since then, by Proposition 4.5,

$$\begin{aligned}\lim_{n\to\infty} &\mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, i \in [k] \,|\, \mathrm{d}_n(i) \ge d_i, i \in [k]) \\ &= \lim_{n\to\infty} \frac{(1+o(1))2^{-\sum_{i=1}^k d_i} \prod_{i=1}^k \mathbb{P}\left(M\sqrt{\frac{a_i}{4-a_i}} + N\sqrt{1-\frac{a_i}{4-a_i}} \le y_i, M > x_i\right)}{\mathbb{P}(\mathrm{d}_n(i) \ge d_i, i \in [k])} \\ &= \prod_{i=1}^{k} \mathbb{P}\left(M\sqrt{\frac{a_i}{4-a_i}} + N\sqrt{1-\frac{a_i}{4-a_i}} \le y_i, M > x_i\right).\end{aligned}$$

Let us define

$$\begin{aligned}f_n(\bar{J}) &:= \mathbb{P}\big(h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, \mathrm{d}_n(i) \ge d_i, i \in [k] \,|\, \overline{\mathcal{S}}_{n,1} = \bar{J}\big), \\ g_n(\bar{J}) &:= \prod_{i=1}^{k} \mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, \mathrm{d}_n(i) \ge d_i \,|\, \mathcal{S}_{n,1}(i) = J_i).\end{aligned}$$

Then, take $c \in (\max_{i\in[k]} a_i, 2)$ and set $\delta := 2 - c$ so that $\mathcal{B}_{n,\delta} \subseteq \mathcal{A}_{\bar{d}}$ by Lemma 4.6. We write,

$$\begin{aligned}&\mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) \ge \ell_i, d_n(i) \ge d_i, i \in [k]) \\ &\quad = \mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\right] \\ &\quad = \mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1}\in\mathcal{B}_{n,\delta}\}}\right] + \mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1}\in\mathcal{A}_{\bar{d}}\setminus\mathcal{B}_{n,\delta}\}}\right].\end{aligned} \tag{5.18}$$

For the first term on the right-hand side, we use that the truncated selection sets are pairwise disjoint by the definition of $\mathcal{B}_{n,\delta}$ in (4.7) and that by Lemma 4.4, $f_n(\bar{J}) = g_n(\bar{J})$ for all $\bar{J} \in \mathcal{B}_{n,\delta}$ and n sufficiently large as a result. Together with Lemma 4.8, recalling

that $\overline{\mathcal{R}}_{n,1}$ is a tuple of k independent copies of $\mathcal{S}_{n,1}(1)$, this yields

$$\begin{aligned}\mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1}\in\mathcal{B}_{n,\delta}\}}\right] &= \sum_{\bar{J}\in\mathcal{B}_{n,\delta}} f_n(\bar{J})\mathbb{P}\big(\overline{\mathcal{S}}_{n,1}=\bar{J}\big)\\ &= \sum_{\bar{J}\in\mathcal{B}_{n,\delta}} g_n(\bar{J})\mathbb{P}\big(\overline{\mathcal{R}}_{n,1}=\bar{J}\big)\,(1+o(1)) \qquad (5.19)\\ &= \mathbb{E}\left[g_n(\overline{\mathcal{R}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{R}}_{n,1}\in\mathcal{B}_{n,\delta}\}}\right](1+o(1)).\end{aligned}$$

Moreover, since $f_n(\bar{J}), g_n(\bar{J}) \le 2^{-\sum_{i=1}^k d_i}$ when $\bar{J}\in\mathcal{A}_{\bar{d}}$ by (4.14) in Lemma 4.10, and using Lemmas 4.7 and 4.8,

$$\begin{aligned}&\left|\mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1}\in\mathcal{A}_{\bar{d}}\setminus\mathcal{B}_{n,\delta}\}}\right]-\mathbb{E}\left[g_n(\overline{\mathcal{R}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{R}}_{n,1}\in\mathcal{A}_{\bar{d}}\setminus\mathcal{B}_{n,\delta}\}}\right]\right|\\ &\le 2^{-\sum_{i=1}^k d_i}\big(\mathbb{P}\big(\overline{\mathcal{S}}_{n,1}\in\mathcal{A}_{\bar{d}}\setminus\mathcal{B}_{n,\delta}\big)+\mathbb{P}\big(\overline{\mathcal{R}}_{n,1}\in\mathcal{A}_{\bar{d}}\setminus\mathcal{B}_{n,\delta}\big)\big) \qquad (5.20)\\ &= o\Big(2^{-\sum_{i=1}^k d_i}\Big).\end{aligned}$$

Thus, combining (5.18), (5.19) and (5.20), we arrive at

$$\mathbb{P}(h_{n,1}(i)\le h_i,\ell_n(i)\ge\ell_i,d_n(i)\ge d_i,i\in[k]) = \mathbb{E}\left[g_n(\overline{\mathcal{R}}_{n,1})\right](1+o(1))+o\Big(2^{-\sum_{i=1}^k d_i}\Big). \qquad (5.21)$$

As the elements of $\overline{\mathcal{R}}_{n,1}$ are i.i.d., we obtain

$$\begin{aligned}\mathbb{E}[g_n(\overline{\mathcal{R}}_{n,1})] &= \prod_{i=1}^k \mathbb{P}(h_{n,1}(i)\le h_i,\ell_n(i)\ge\ell_i,\mathrm{d}_n(i)\ge d_i)\\ &= \prod_{i=1}^k \mathbb{P}(h_{n,1}(i)\le h_i,\ell_n(i)\ge\ell_i\,|\,\mathrm{d}_n(i)\ge d_i)\,\mathbb{P}(\mathrm{d}_n(i)\ge d_i)\,.\end{aligned}$$

By combining Proposition 4.5, Proposition 5.1 and (5.21), we thus have

$$\begin{aligned}&\mathbb{P}(h_{n,1}(i)\le h_i,\ell_n(i)\ge\ell_i,d_n(i)\ge d_i,i\in[k])\\ &= (1+o(1))2^{-\sum_{i=1}^k d_i}\prod_{i=1}^k\mathbb{P}\left(M\sqrt{\frac{a_i}{4-a_i}}+N\sqrt{1-\frac{a_i}{4-a_i}}\le y_i, M>x_i\right),\end{aligned}$$

as desired, which concludes the proof. □

6 Joint properties of vertices with a given label

This section is devoted to studying the joint behaviour of the degree and depth of vertices with a given label. We use the preliminary results proved in Section 4 to obtain the required results. The section is structured in the same way as Section 5.

We let $\ell\in[n]$ be increasing in n such that ℓ diverges as $n\to\infty$, and set

$$h := \log\ell + x\sqrt{\log\ell},\quad \begin{cases} d := \log(n/\ell)+y\sqrt{\log(n/\ell)} & \text{if } \ell=o(n),\\ d\in\mathbb{N}_0 \text{ fixed} & \text{otherwise,}\end{cases}\quad t_n := \lceil\log\ell\rceil, \qquad (6.1)$$

with $x,y\in\mathbb{R}$. Moreover, we define, for the same $y\in\mathbb{R}$ used in the definition of d, $\rho\in(0,1)$ and with $P(\rho)\sim\mathrm{Poi}(\log(1/\rho))$,

$$\mathrm{Pr} = \mathrm{Pr}(y,\rho,\ell) := \begin{cases}\Phi(y) & \text{if } \ell=o(n),\\ \mathbb{P}(P(\rho)\le d) & \text{if } \ell=(1+o(1))\rho n,\\ 1 & \text{if } \ell=n-o(n).\end{cases} \qquad (6.2)$$

We then have the following result.

Proposition 6.1. *Let d, h, ℓ and t_n be as in* (6.1) *with $x, y \in \mathbb{R}$, and recall* Pr *from* (6.2)*, with $y \in \mathbb{R}, \rho \in (0,1)$. Then,*

$$\lim_{n\to\infty} \mathbb{P}(h_{n,1}(1) \le h, \mathrm{d}_n(1) \le d \,|\, \ell_n(1) = \ell) = \Phi(x)\mathrm{Pr}.$$

Remark 6.2. As is the case in Remark 5.2, it follows from Lemma 4.11 and Remark 4.12 that the result in Proposition 6.1 holds when substituting $h_n(i)$ for $h_{n,1}(i)$ as well.

Proof. We start by using Lemmas 4.1 and 4.10 to obtain

$$\begin{aligned}
&\mathbb{P}(h_{n,1}(1) \le h, \mathrm{d}_n(1) \le d \,|\, \ell_n(1) = \ell) \\
&\quad = \frac{\mathbb{P}(h_{n,1}(1) \le h, \mathrm{d}_n(1) \le d, \ell_n(1) = \ell)}{\mathbb{P}(\ell_n(1) = \ell)} \\
&\quad = n\mathbb{E}\Big[2^{-([\ell+1,n]\cap\mathcal{S}_{n,1}(1)|+1)}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\mathbb{1}_{\{\ell\in\mathcal{S}_{n,1}\}}\mathbb{P}\big(X_{n,\ell,2}(1) \le h-1\big|\mathcal{S}_{n,1}(1)\big)\Big].
\end{aligned}$$

We observe that we can divide the terms in the expected value into three parts which are pairwise independent. Namely, the exponent and the first indicator, the second indicator, and finally the conditional probability, respectively. Indeed, the exponent and first indicator only depend on $[\ell+1,n]\cap\mathcal{S}_{n,1}(1)$, the second indicator only on the event $\{\ell \in \mathcal{S}_{n,1}(1)\}$, and the conditional probability depends only on $[t_n, \ell-1]\cap\mathcal{S}_{n,1}(1)$. Since $\ell > \lceil \log \ell \rceil = t_n$ for all n sufficiently large, these three parts depend on disjoint sets of independent random variables and are hence independent. As a result, we obtain

$$\begin{aligned}
&\mathbb{E}\big[2^{-([\ell+1,n]\cap\mathcal{S}_{n,1}(1)|+1)}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\mathbb{1}_{\{\ell\in\mathcal{S}_{n,1}\}}\mathbb{P}(X_{n,\ell,2}(1) \le h-1 \,|\, \mathcal{S}_{n,1}(1))\big] \\
&= \frac{1}{2}\mathbb{P}(\ell \in \mathcal{S}_{n,1}(1))\,\mathbb{E}\left[2^{-|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\right]\mathbb{P}(X_{n,\ell,2}(1) \le h-1)\,.
\end{aligned} \tag{6.3}$$

The first probability on the right-hand side equals $2/\ell$. The expected value on the right-hand side can be rewritten as follows. First, by summing over all possible truncated selection sets $\mathcal{S}_{n,1}(1)$,

$$\begin{aligned}
&\mathbb{E}\Big[2^{-|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\Big] \\
&\quad = \sum_{m=0}^{d}\sum_{\substack{S\subseteq[\ell+1,n]\\|S|=m}} 2^{-m}\prod_{j\in S}\frac{2}{j}\prod_{j\notin S}\Big(1-\frac{2}{j}\Big) \\
&\quad = \sum_{m=0}^{d}\sum_{\substack{S\subseteq\{\ell+1,\dots,n\}\\|S|=m}}\prod_{j\in S}\frac{1}{j}\prod_{j\notin S}\Big(1-\frac{1}{j}\Big)\prod_{j\notin S}\frac{j-2}{j-1} \\
&\quad = \frac{\ell-1}{n-1}\sum_{m=0}^{d}\sum_{\substack{S\subseteq\{\ell+1,\dots,n\}\\|S|=m}}\prod_{j\in S}\frac{j-1}{j-2}\prod_{j\in S}\frac{1}{j}\prod_{j\notin S}\Big(1-\frac{1}{j}\Big).
\end{aligned}$$

As ℓ diverges with n, $(\ell-1)/(n-1) = (1+o(1))\ell/n$. Defining $\widetilde{\mathcal{S}}_{n,1}(1)$ as a random subset of $\{\ell+1,\dots,n\}$ which includes each integer $j \in \{\ell+1,\dots,n\}$ independently with probability $1/j$, the double sum and triple product can be interpreted as

$$\mathbb{E}\left[\mathbb{1}_{\{|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\}}\prod_{j=\ell+1}^{n}\Big(1+\mathbb{1}_{\{j\in\widetilde{\mathcal{S}}_{n,1}(1)\}}\frac{1}{j-2}\Big)\right].$$

Combining both observations we obtain

$$\begin{aligned}&\mathbb{E}\Big[2^{-|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\Big]\\&\quad=(1+o(1))\frac{\ell}{n}\mathbb{E}\left[\mathbb{1}_{\{|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\}}\prod_{j=\ell+1}^{n}\Big(1+\mathbb{1}_{\{j\in\widetilde{\mathcal{S}}_{n,1}(1)\}}\frac{1}{j-2}\Big)\right].\end{aligned}\tag{6.4}$$

By bounding the product from below by one and using (6.3), we obtain the lower bound

$$\begin{aligned}&n\mathbb{E}\big[2^{-([\ell+1,n]\cap\mathcal{S}_{n,1}(1)|+1)}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\mathbb{1}_{\{\ell\in\mathcal{S}_{n,1}\}}\mathbb{P}(X_{n,\ell,2}(1)\le h-1\,|\,\mathcal{S}_{n,1}(1))\,\big]\\&\quad\ge(1+o(1))\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)\,\mathbb{P}(X_{n,\ell,2}(1)\le h-1)\\&\quad=\Phi(x)\mathrm{Pr}+o(1),\end{aligned}\tag{6.5}$$

where the last step follows if we assume that the two probabilities in the first step are asymptotically equal to Pr and $\Phi(x)$, respectively. For an upper bound, we first expand the product in the expected value of (6.4) to obtain

$$\begin{aligned}&\mathbb{E}\Big[2^{-|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\Big]\\&=(1+o(1))\frac{\ell}{n}\Bigg(\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)\\&\quad+\sum_{m=1}^{n-\ell}\sum_{\ell+1\le j_1<\ldots<j_m\le n}\Big(\prod_{t=1}^{m}\frac{1}{j_t-2}\Big)\mathbb{E}\left[\mathbb{1}_{\{|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\}}\prod_{t=1}^{m}\mathbb{1}_{\{j_t\in\widetilde{\mathcal{S}}_{n,1}(1)\}}\right]\Bigg).\end{aligned}\tag{6.6}$$

We then use the Cauchy-Schwarz inequality to bound

$$\begin{aligned}&\sum_{m=1}^{n-\ell}\sum_{\ell+1\le j_1<\ldots<j_m\le n}\Big(\prod_{t=1}^{m}\frac{1}{j_t-2}\Big)\mathbb{E}\left[\mathbb{1}_{\{|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\}}\prod_{t=1}^{m}\mathbb{1}_{\{j_t\in\widetilde{\mathcal{S}}_{n,1}(1)\}}\right]\\&\quad\le\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)^{1/2}\sum_{m=1}^{n-\ell}\sum_{\ell+1\le j_1<\ldots<j_m\le n}\prod_{t=1}^{m}\left(\frac{1}{j_t-2}\mathbb{P}\Big(j_t\in\widetilde{\mathcal{S}}_{n,1}(1)\Big)^{1/2}\right).\end{aligned}$$

As $\mathbb{P}(j_t\in\widetilde{\mathcal{S}}_{n,1}(1))=1/j_t\le 1/(j_t-2)$, we arrive at the upper bound

$$\begin{aligned}&\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)^{1/2}\sum_{m=1}^{n-\ell}\sum_{\ell-1\le j_1<\ldots<j_m\le n-2}\prod_{t=1}^{m}\frac{1}{j_t^{3/2}}\\&\quad\le\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)^{1/2}\sum_{m=1}^{n-\ell}\big(2(\ell-2)^{-1/2}\big)^m\\&\quad\le\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)^{1/2}\frac{2}{(\ell-2)^{1/2}-2}.\end{aligned}$$

Combining this with (6.6) and (6.3) and since ℓ diverges with n, we thus obtain the upper bound

$$\begin{aligned}&n\mathbb{E}\big[2^{-([\ell+1,n]\cap\mathcal{S}_{n,1}(1)|+1)}\mathbb{1}_{\{|[\ell+1,n]\cap\mathcal{S}_{n,1}(1)|\le d\}}\mathbb{1}_{\{\ell\in\mathcal{S}_{n,1}\}}\mathbb{P}(X_{n,\ell,2}(1)\le h-1\,|\,\mathcal{S}_{n,1}(1))\,\big]\\&\le(1+o(1))\Big(\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)+\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)|\le d\Big)^{1/2}\frac{2}{(\ell-2)^{1/2}+2}\Big)\mathbb{P}(X_{n,\ell,2}(1)\le h-1)\\&=\Phi(x)\mathrm{Pr}+o(1),\end{aligned}$$

when we (again) assume that the first and last probability on the second line are asymptotically equal to Pr and $\Phi(x)$, respectively. As this upper bound matches the lower bound in (6.5), we arrive at the desired result.

It remains to prove that

$$\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)| \le d\Big) = \Pr + o(1), \qquad \mathbb{P}(X_{n,\ell,2}(1) \le h-1) = \Phi(x) + o(1). \tag{6.7}$$

For the first result, let us start by considering $\ell = o(n)$, so that $d := \log(n/\ell) + y\sqrt{\log(n/\ell)}$ for $y \in \mathbb{R}$ fixed. It then follows from Lindeberg's conditions [11, Theorem 3.4.5] that

$$\begin{aligned}\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)| \le d\Big) &= \mathbb{P}\left(\frac{|\widetilde{\mathcal{S}}_{n,1}(1)| - \mathbb{E}[|\widetilde{\mathcal{S}}_{n,1}(1)|]}{\sqrt{\mathrm{Var}(|\widetilde{\mathcal{S}}_{n,1}(1)|)}} \le \frac{\log(n/\ell) + y\sqrt{\log(n/\ell)} - \mathbb{E}[|\widetilde{\mathcal{S}}_{n,1}(1)|]}{\sqrt{\mathrm{Var}(|\widetilde{\mathcal{S}}_{n,1}(1)|)}}\right)\\ &= \Phi(y) + o(1),\end{aligned}$$

since

$$\begin{aligned}\mathbb{E}[|\widetilde{\mathcal{S}}_{n,1}(1)|] &= \sum_{j=\ell+1}^{n} \frac{1}{j} = \log(n/\ell) + \mathcal{O}(1),\\ \mathrm{Var}(|\widetilde{\mathcal{S}}_{n,1}(1)|) &= \sum_{j=\ell+1}^{n} \frac{1}{j}\Big(1 - \frac{1}{j}\Big) = \log(n/\ell) + \mathcal{O}(1),\end{aligned}$$

when $\ell = o(n)$. When $\ell = (1+o(1))\rho n$ for some $\rho \in (0,1)$, we recall that $d \in \mathbb{N}_0$ is fixed, and instead use that for any $t \in \mathbb{R}$,

$$\mathbb{E}\left[\mathrm{e}^{t|\widetilde{\mathcal{S}}_{n,1}(1)|}\right] = \prod_{j=\ell+1}^{n}\Big(1 - \frac{1}{j} + \mathrm{e}^t\frac{1}{j}\Big) = \prod_{j=\ell+1}^{n}\Big(1 + \big(\mathrm{e}^t - 1\big)\frac{1}{j}\Big).$$

Using that $x - x^2 \le \log(1+x) \le x$ for all $x > 0$ and that

$$\sum_{j=\ell+1}^{n} \frac{1}{j} = (1+o(1))\int_{\rho}^{1} x^{-1}\,\mathrm{d}x = (1+o(1))\log(1/\rho),$$

yields

$$\mathbb{E}\left[\mathrm{e}^{t|\widetilde{\mathcal{S}}_{n,1}(1)|}\right] = \mathrm{e}^{(\mathrm{e}^t-1)\log(1/\rho)} + o(1).$$

Since, for any $t \in \mathbb{R}$, the moment generating function (MGF) of $|\widetilde{\mathcal{S}}_{n,1}(1)|$ converges to the MGF of $P(\rho)$,

$$\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)| \le d\Big) = \mathbb{P}(P(\rho) \le d) + o(1).$$

Finally, when $\ell = n - o(n)$, using Markov's inequality yields

$$\mathbb{P}\Big(|\widetilde{\mathcal{S}}_{n,1}(1)| = 0\Big) \ge 1 - \mathbb{E}[|\widetilde{\mathcal{S}}_{n,1}(1)|] = 1 - \sum_{j=\ell+1}^{n}\frac{1}{j} = 1 - (1+o(1))\log(n/\ell) = 1 - o(1),$$

as desired.

For the latter result in (6.7) we set $\widetilde{Q}_n := |[t_n, \ell-1] \cap \mathcal{S}_{n,1}(1)|$, let $(I_j^n)_{j\in[n]}$ denote independent Bernoulli random variables with success probability $1/2$, also independent of $\widetilde{Q}_n$, and write

$$\frac{2X_{n,\ell,2}(1) - 2\log\ell}{2\sqrt{\log\ell}} = \frac{2\sum_{j=1}^{\widetilde{Q}_n} I_j^{\widetilde{Q}_n} - \widetilde{Q}_n}{\sqrt{\widetilde{Q}_n}}\sqrt{\frac{\widetilde{Q}_n}{4\log\ell}} + \frac{\widetilde{Q}_n - \mathbb{E}[\widetilde{Q}_n]}{\sqrt{\mathrm{Var}(\widetilde{Q}_n)}}\sqrt{\frac{\mathrm{Var}(\widetilde{Q}_n)}{4\log\ell}} + \frac{\mathbb{E}[\widetilde{Q}_n] - 2\log\ell}{2\sqrt{\log\ell}}.$$

We then use a similar approach as (5.10). In particular, we use the Skorokhod embedding which provides us with a coupling of the random variables $\widetilde{Q}_n$ and $(I_i^n)_{i\in[n]}$ such that

$$\frac{2\sum_{j=1}^{n} I_j^n - n}{\sqrt{n}} \xrightarrow{a.s.} N_1, \qquad \frac{\widetilde{Q}_n - \mathbb{E}[\widetilde{Q}_n]}{\sqrt{\mathrm{Var}(\widetilde{Q}_n)}} \xrightarrow{a.s.} N_2,$$

where N_1, N_2 are two independent standard normal random variables. Moreover, a straightforward computation of $\mathbb{E}[\widetilde{Q}_n]$ and $\mathrm{Var}(\widetilde{Q}_n)$ shows that $\widetilde{Q}_n/(2\log\ell) \xrightarrow{a.s.} 1$ (and hence $\widetilde{Q}_n \xrightarrow{a.s.} \infty$), that $\mathrm{Var}(\widetilde{Q}_n)/(2\log\ell) \to 1$ and that $\mathbb{E}[\widetilde{Q}_n] - 2\log\ell = o(\sqrt{\log\ell})$ as $n \to \infty$. As a result, it follows that

$$\frac{2X_{n,\ell,2}(1) - 2\log\ell}{2\sqrt{\log\ell}} \xrightarrow{d} \frac{1}{\sqrt{2}}N_1 + \frac{1}{\sqrt{2}}N_2 \overset{d}{=} N,$$

where N is a standard normal random variable. As a result, we obtain (recalling that $h := \log\ell + x\sqrt{\log\ell}$),

$$\mathbb{P}(X_{n,\ell,2}(1) \le h-1) = \mathbb{P}\left(\frac{2X_{n,\ell,2}(1) - 2\log\ell}{2\sqrt{\log\ell}} \le \frac{h-1-\log\ell}{\sqrt{\log\ell}}\right) = \Phi(x) + o(1),$$

as required, which concludes the proof. □

To finish this section, we use the results related to the truncated selection sets developed in Section 4 to extend Proposition 6.1 to the case of multiple vertices. The choice of t_n is imperative, and so we define, for some $(\ell_i)_{i\in[k]} \in [n]^k$,

$$t_n := \min_{i\in[k]} \lceil \log \ell_i \rceil. \tag{6.8}$$

We can then formulate the following result.

Proposition 6.3. *Fix $k \in \mathbb{N}$, let $(\ell_i)_{i\in[k]}$ be k distinct integer-valued sequences such that ℓ_i increases with n and ℓ_i diverges as $n \to \infty$ for all $i \in [k]$. Let, for $i \in [k]$, $h_i := \log\ell_i + x_i\sqrt{\log\ell_i}$ and $d_i := \log(n/\ell_i) + y_i\sqrt{\log(n/\ell_i)}$ if $\ell_i = o(n)$ and $d_i \in \mathbb{N}_0$ fixed otherwise, where $(x_i)_{i\in[k]}, (y_i)_{i\in[k]} \in \mathbb{R}^k$ and let t_n as in* (6.8)*. Furthermore, recall the definition of* Pr *in* (6.2)*. Then,*

$$\lim_{n\to\infty} \mathbb{P}(h_{n,1} \le h_i, \mathrm{d}_n(i) \le d_i, i \in [k] \,|\, \ell_n(i) = \ell_i, i \in [k]) = \prod_{i=1}^{k} \Phi(x_i)\mathrm{Pr}(y_i, \rho_i, \ell_i).$$

Remark 6.4. As is the case in Remark 5.2, it follows from Lemma 4.11 and Remark 4.12 that the result in Proposition 6.3 holds when substituting $h_n(i)$ for $h_{n,1}(i)$ as well.

Proof. The proof follows a similar approach as the proof of Proposition 5.3. We first write

$$\begin{aligned}
&\mathbb{P}(h_{n,1} \le h_i, \mathrm{d}_n(i) \le d_i, i \in [k] \,|\, \ell_n(i) = \ell_i, i \in [k]) \\
&\quad = \frac{\mathbb{P}(h_{n,1} \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \le d_i, i \in [k])}{\mathbb{P}(\ell_n(i) = \ell_i, i \in [k])} \\
&\quad = (n)_k \mathbb{P}(h_{n,1} \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \le d_i, i \in [k]),
\end{aligned} \tag{6.9}$$

where the last step follows from Lemma 4.10. We then define

$$\begin{aligned}
f_n(\bar{J}) &:= \mathbb{P}\big(h_{n,1}(i) \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \le d_i, i \in [k] \,|\, \overline{\mathcal{S}}_{n,1} = \bar{J}\big), \\
g_n(\bar{J}) &:= \prod_{i=1}^{k} \mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \le d_i, \,|\, \mathcal{S}_{n,1}(i) = J_i).
\end{aligned}$$

With similar steps as in (5.18) through (5.21), we then have

$$\begin{aligned}
&\mathbb{P}(h_{n,1}(i) \le h_i, \ell_n(i) = \ell_i, \mathrm{d}_n(i) \le d_i, i \in [k]) \\
&= \mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n\}}\right] + \mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c\}}\right] \\
&= \mathbb{E}\left[g_n(\overline{\mathcal{R}}_{n,1})\right](1 + o(1)) + \left(\mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1} \in \widetilde{\mathcal{B}}_n^c\}}\right] - \mathbb{E}\left[g_n(\overline{\mathcal{R}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{R}}_{n,1} \in \widetilde{\mathcal{B}}_n^c\}}\right]\right).
\end{aligned} \tag{6.10}$$

It follows from (4.13) in Lemma 4.10 that

$$\mathbb{E}\left[f_n(\overline{\mathcal{S}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{S}}_{n,1}\in\widetilde{\mathcal{B}}_n^c\}}\right] = \mathbb{P}\Big(\ell_n(i)=\ell_i, i\in[k], \overline{\mathcal{S}}_{n,1}\in\widetilde{\mathcal{B}}_n^c\Big) = o(n^{-k}).$$

A similar argument as in the proof of (4.13) can be used to show that

$$\mathbb{E}\left[g_n(\overline{\mathcal{R}}_{n,1})\mathbb{1}_{\{\overline{\mathcal{R}}_{n,1}\in\widetilde{B}^c\}}\right] = o(n^{-k}),$$

as well. As the elements of $\overline{\mathcal{R}}_{n,1}$ are i.i.d., we have

$$\begin{aligned}\mathbb{E}[g_n(\overline{\mathcal{R}}_{n,1})] &= \prod_{i=1}^{k}\mathbb{P}(h_{n,1}(i)\le h_i, \ell_n(i)=\ell_i, \mathrm{d}_n(i)\le d_i)\\ &= \prod_{i=1}^{k}\mathbb{P}(h_{n,1}(i)\le h_i, \mathrm{d}_n(i)\le d_i \,|\, \ell_n(i)=\ell_i)\,\mathbb{P}(\ell_n(i)=\ell_i)\,.\end{aligned}$$

The product on the right-hand side equals

$$(1+o(1))n^{-k}\prod_{i=1}^{k}\Phi(y_i)\Pr(y_i,\rho_i,\ell_i)$$

by Proposition 6.1 and Lemma 4.10. Using this in (6.10) yields

$$\mathbb{P}(h_{n,1}(i)\le h_i, \ell_n(i)=\ell_i, \mathrm{d}_n(i)\le d_i, i\in[k]) = \frac{1+o(1)}{n^k}\prod_{i=1}^{k}\Phi(y_i)\Pr(y_i,\rho_i,\ell_i).$$

Combining this with (6.9) then yields the desired result. □

7 Proof of Theorem 2.2

This section is devoted to proving Theorem 2.2. We first provide some additional theory on top of what is introduced in Section 3 prior to stating the proof.

7.1 Convergence of marked point processes via finite dimensional distributions

We recall that, as discussed in Section 3, Theorem 2.2 can be understood as the weak convergence of the marked point process $\mathcal{MP}^{(n_t)}$ to $\mathcal{MP}^{\epsilon}$, as defined in (3.2) and (3.1), respectively. The approach to prove this is via the convergence of its finite dimensional distributions (FDDs) along suitable subsequences. The FDDs of a random measure $\mathcal{P}$ on $\mathcal{X}$ are defined as the joint distributions, for all finite families of bounded Borel sets $(B_1,\dots,B_k)\in\mathcal{B}(\mathcal{X})^k$, of the random variables $(\mathcal{P}(B_1),\dots,\mathcal{P}(B_k))$, see [6, Definition 9.2.II]. Moreover, by [6, Proposition 9.2.III], the distribution of a random measure $\mathcal{P}$ on $\mathcal{X}$ is completely determined by the FDDs for all finite families $(B_1,\dots,B_k)$ of *disjoint* sets from a semiring $\mathcal{A}$ that generates $\mathcal{B}(\mathcal{X})$. In our case, we consider the marked point process $\mathcal{MP}^{(n)}$ on $\mathcal{X}:=\mathbb{Z}^*\times\mathbb{R}^2$, see (3.1). Here, we let

$$\mathcal{A} := \{\{s\}\times(a,b]\times(c,d] : s\in\mathbb{Z}, a,b,c,d\in\mathbb{R}\}\cup\{[s,\infty]\times(a,b]\times(c,d] : s\in\mathbb{Z}, a,b,c,d\in\mathbb{R}\} \tag{7.1}$$

be the semiring that generates $\mathcal{B}(\mathbb{Z}^*\times\mathbb{R}^2)$.

Recall the counting measures $X_s^{(n)}(B), X_{\ge s}^{(n)}(B), \widetilde{X}_s^{(n)}(B), \widetilde{X}_{\ge s}^{(n)}(B)$ defined in (3.7) (in terms of the Kingman n-coalescent) and $X_s(B), X_{\ge s}(B)$ defined in (3.3). We observe that $\widetilde{X}_s^{(n)}(B)=\mathcal{MP}^{(n)}(\{s\}\times B), \widetilde{X}_{\ge s}^{(n)}(B)=\mathcal{MP}^{(n)}([s,\infty]\times B), X_s(B)=\mathcal{MP}^{\epsilon}(\{s\}\times B)$ and

$X_{\geq s}(B) = \mathcal{MP}^{\epsilon}([s,\infty] \times B)$. As a result, the convergence of the FDDs of $\mathcal{MP}^{(n_t)}$ to the FDDs of $\mathcal{MP}^{\epsilon}$ can be obtained via the convergence of any finite collection of these counting measures.

For any $K \in \mathbb{N}$, take any (fixed) increasing integer sequence $(s_m)_{m\in[K]}$ and define $0 \leq K' := \min\{m : s_{m+1} = s_K\}$. Also fix any sequence $(B_m)_{m\in[K]}$ with $B_m \in \mathcal{B}(\mathbb{R}^2)$ such that $B_m \cap B_\ell = \varnothing$ when $s_m = s_\ell$ and $m \neq \ell$. The conditions on the sets B_m ensure that the elements $\{s_1\} \times B_1, \ldots, \{s'_K\} \times B_{K'}, \{s_{K'+1}, \ldots\} \times B_{K'+1}, \ldots, \{s_K, \ldots\} \times B_K$ of $\mathcal{A}$ are disjoint. We are thus required to prove the joint distributional convergence of the random variables

$$(\widetilde{X}^{(n)}_{s_1}(B_1), \ldots, \widetilde{X}^{(n)}_{s_{K'}}(B_{K'}), \widetilde{X}^{(n)}_{\geq s_{K'+1}}(B_{K'+1}), \ldots, \widetilde{X}^{(n)}_{\geq s_K}(B_K)),$$

to prove Theorem 2.2. We use the method of moments combined with Proposition 3.1 to achieve this:

Proof of Theorem 2.2 subject to Proposition 3.1. As discussed, it suffices to prove the weak convergence of $\mathcal{MP}^{(n_t)}$ to $\mathcal{MP}^{\epsilon}$ along subsequences $(n_t)_{t\in\mathbb{N}}$ such that $\epsilon_{n_t} \to \epsilon$ (where $\epsilon \in [0,1]$) as $t \to \infty$. In turn, this is implied by the convergence of the FDDs, i.e., by the joint convergence of the counting measures $\widetilde{X}^{(n)}_s(B), \widetilde{X}^{(n)}_{\geq s}(B)$ of finite collections of disjoint subsets of $\mathcal{A}$ (see (7.1)).

We recall that the points P_i in the definition of the variables $X_s(B), X_{\geq s}(B)$ in (3.3) are the points of the Poisson point process $\mathcal{P}$ with intensity measure $\lambda(\mathrm{d}x) := 2^{-x}\log 2\,\mathrm{d}x$ in decreasing order. As a result, as the random variables $(M_i, N_i)_{i\in\mathbb{N}}$ are i.i.d. and also independent of $\mathcal{P}$, $X_s(B) \sim \mathrm{Poi}(\lambda_s(B)), X_{\geq s}(B) \sim \mathrm{Poi}(2\lambda_s(B))$, where

$$\lambda_s(B) = 2^{-(s+1)+\epsilon}\mathbb{P}\left(M_1\sqrt{1-\frac{\mu}{\sigma^2}} + N_1\sqrt{\frac{\mu}{\sigma^2}} \in B\right).$$

We also recall that $(n_\ell)_{\ell\in\mathbb{N}}$ is a subsequence such that $\epsilon_{n_\ell} \to \epsilon$ as $\ell \to \infty$. We now take $c \in (1/\log 2, 2)$ and for any $K \in \mathbb{N}$ consider any fixed non-decreasing integer sequence $(s_m)_{m\in[K]}$. It follows from the choice of c and the fact that the s_m are fixed with respect to n that $s_1 + \log_2 n = \omega(1)$ and that $s_K + \log_2 n < c\log n$ for all $n \geq 2$. Moreover, let $K' := \min\{m : s_{m+1} = s_K\}$ and let $(B_m)_{m\in[K]}$ be a sequence of sets in $\mathcal{B}(\mathbb{R}^2)$ such that $B_m \cap B_\ell = \varnothing$ when $s_m = s_\ell$ and $m \neq \ell$. We can then, for any $(c_m)_{m\in[K]} \in \mathbb{N}_0^K$, obtain from Proposition 3.1 and since $s_1, \ldots, s_K = o(\sqrt{\log n})$, that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\Big[\prod_{m=1}^{K'}\Big(\widetilde{X}^{(n_\ell)}_{s_m}(B_m)\Big)_{c_m} \prod_{m=K'+1}^{K}\Big(\widetilde{X}^{(n_\ell)}_{\geq s_m}(B_m)\Big)_{c_m}\Big] \\
= \prod_{m=1}^{K'}\lambda_{s_m}(B_m)^{c_m} \prod_{m=K'+1}^{K}(2\lambda_{s_m}(B_m))^{c_m} \\
= \mathbb{E}\Big[\prod_{m=1}^{K'}\Big(X_{s_m}(B_m)\Big)_{c_m} \prod_{m=K'+1}^{K}\Big(X_{\geq s_m}(B_m)\Big)_{c_m}\Big],
\end{aligned}$$

where the last step follows from the independence property of (marked) Poisson point processes and the choice of the sequences $(s_m, B_m)_{m\in[K]}$. The method of moments [21, Section 6.1] then concludes the proof. □

It remains to prove Proposition 3.1.

Proof of Proposition 3.1. The proof essentially follows a similar approach as the proof of [23, Proposition 5.4]. However, as certain estimations and definitions differ, we include it here for completeness.

Recall that $\mu = 1 - 1/(2\log 2)$, $\sigma^2 = 1 - 1/(4\log 2)$, and that we have fixed $K \in \mathbb{N}, (a_m)_{m\in[K]} \in [0,2)^K$. Moreover, we have a non-decreasing integer sequence $(s_m)_{m\in[K]}$ such that $s_1 + \log_2 n = \omega(1)$ and

$$\lim_{n\to\infty} \frac{s_m + \log_2 n}{\log n} = a_m,$$

for all $m \in [K]$, and a sequence $(B_m)_{m\in[K]}$ such that $B_m \in \mathcal{B}(\mathbb{R}^2)$ for all $m \in [K]$ and $B_m \cap B_\ell = \varnothing$ when $s_m = s_\ell$ and $m \neq \ell$. We also define $K' := \min\{m : s_{m+1} = s_K\}$. Finally, we recall that M and N are two independent standard normal random variables. Then, take an arbitrary sequence $(c_m)_{m\in[K]} \in \mathbb{N}_0^K$ and set $L := \sum_{m=1}^{K} c_m$ and $L' := \sum_{m=1}^{K'} c_m$.

We define $\bar{d} = (d_i)_{i\in[L]} \in \mathbb{Z}^L, (a'_i)_{i\in[L]}$, and $\bar{A} = (A_i)_{i\in[L]} \subset \mathcal{B}(\mathbb{R}^2)^L$ as follows: For each $i \in [L]$, find the unique $m \in [K]$ such that $\sum_{\ell=1}^{m-1} c_\ell < i \leq \sum_{\ell=1}^{m} c_\ell$ and define $d_i := \lfloor \log_2 n \rfloor + s_m, a'_i := a_m, A_i := B_m$. We note that this construction implies that the first c_1 many d_i, a'_i and A_i equal $\lfloor \log_2 n \rfloor + s_1, a_1$ and B_1, respectively, that the next c_2 many d_i, a'_i and A_i equal $\lfloor \log_2 n \rfloor + s_2, a_2$ and B_2, respectively, etcetera. Furthermore, $\lim_{n\to\infty} d_i/\log n = a'_i$ for all $i \in [L]$. We then define the events

$$\begin{aligned}
\mathcal{HL}_{\bar{A},\bar{d}} &:= \Big\{ \Big(\frac{h_n(i) - (\log n - d_i/2)}{\sqrt{\log n - d_i/4}}, \frac{\log \ell_n(i) - (\log n - d_i/2)}{\sqrt{d_i/4}} \Big) \in A_i, i \in [L] \Big\}, \\
\mathcal{D}_{\bar{d}}(L', L) &:= \{ \mathrm{d}_n(i') = d'_i, i' \in [L'], \mathrm{d}_n(i) \geq d_i, L' < i \leq L \}, \\
\mathcal{E}_{\bar{d}}(S) &:= \{ \mathrm{d}_n(i) \geq d_i + \mathbb{1}_{\{i\in S\}}, i \in [L] \}.
\end{aligned}$$

We know from [1, Lemma 5.1] that by the inclusion-exclusion principle,

$$\mathbb{P}(\mathcal{D}_{\bar{d}}(L', L)) = \sum_{j=0}^{L'} \sum_{\substack{S\subseteq[L']: \\ |S|=j}} (-1)^j \mathbb{P}(\mathcal{E}_{\bar{d}}(S)),$$

so that intersecting the event $\mathcal{HL}_{\bar{A},\bar{d}}$ in the probabilities on both sides yields

$$\mathbb{P}\big(\mathcal{D}_{\bar{d}}(L', M) \cap \mathcal{HL}_{\bar{A},\bar{d}}\big) = \sum_{j=0}^{L'} \sum_{\substack{S\subseteq[L']: \\ |S|=j}} (-1)^j \mathbb{P}\big(\mathcal{E}_{\bar{d}}(S) \cap \mathcal{HL}_{\bar{A},\bar{d}}\big). \tag{7.2}$$

Let us then define

$$\begin{aligned}
\widetilde{\mathcal{HL}}_{\bar{A},\bar{d}}(S) := \Big\{ \Big(& \frac{h_n(i) - (\log n - (d_i + \mathbb{1}_{\{i\in S\}})/2)}{\sqrt{\log n - (d_i + \mathbb{1}_{\{i\in S\}})/4}}, \\
& \frac{\log \ell_n(i) - (\log n - (d_i + \mathbb{1}_{\{i\in S\}})/2)}{\sqrt{(d_i + \mathbb{1}_{\{i\in S\}})/4}} \Big) \in A_i, i \in [L] \Big\}.
\end{aligned}$$

We use Proposition 5.3 (combined with Remark 5.4) with $a'_i = \lim_{n\to\infty}(d_i + \mathbb{1}_{\{i\in S\}})/\log n$ for all $i \in [L]$ and Proposition 4.5 to then obtain

$$\begin{aligned}
&\mathbb{P}\Big(\mathcal{E}_{\bar{d}}(S) \cap \widetilde{\mathcal{HL}}_{\bar{A},\bar{d}}(S)\Big) \\
&\quad = \mathbb{P}\Big(\widetilde{\mathcal{HL}}_{\bar{A},\bar{d}}(S) \,\Big|\, \mathcal{E}_{\bar{d}}(S)\Big) \mathbb{P}(\mathcal{E}_{\bar{d}}(S)) \\
&\quad = (1 + o(1)) 2^{-\sum_{i=1}^{L} (d_i + \mathbb{1}_{\{i\in S\}})} \prod_{i=1}^{L} \mathbb{P}\left(\left(M\sqrt{\frac{a'_i}{4 - a'_i}} + N\sqrt{1 - \frac{a'_i}{4 - a'_i}}, M \right) \in A_i \right).
\end{aligned}$$

Since

$$\Big(\frac{h_n(i)-(\log n-(d_i+\mathbb{1}_{\{i\in S\}})/2)}{\sqrt{\log n-(d_i+\mathbb{1}_{\{i\in S\}})/4}},\frac{\log \ell_n(i)-(\log n-(d_i+\mathbb{1}_{\{i\in S\}})/2)}{\sqrt{(d_i+\mathbb{1}_{\{i\in S\}})/4}}\Big)$$
$$=(1+o(1))\Big(\frac{h_n(i)-(\log n-d_i/2)}{\sqrt{\log n-d_i/4}},\frac{\log \ell_n(i)-(\log n-d_i/2)}{\sqrt{d_i/4}}\Big),$$

we also obtain from Slutsky's theorem [35, Lemma 2.8] that

$$\mathbb{P}(\mathcal{E}_{\bar{d}}(S)\cap\mathcal{HL}_{\bar{A},\bar{d}})$$
$$=(1+o(1))2^{-\sum_{i=1}^{L}(d_i+\mathbb{1}_{\{i\in S\}})}\prod_{i=1}^{L}\mathbb{P}\left(\left(M\sqrt{\frac{a_i'}{4-a_i'}}+N\sqrt{1-\frac{a_i'}{4-a_i'}},M\right)\in A_i\right).$$

The right-hand side of (7.2) then equals

$$\prod_{i=1}^{L}\left[\mathbb{P}\left(\left(M\sqrt{\frac{a_i'}{4-a_i'}}+N\sqrt{1-\frac{a_i'}{4-a_i'}},M\right)\in A_i\right)\right]\sum_{j=0}^{L}\sum_{\substack{S\subseteq[L']:\\|S|=j}}\frac{(1+o(1))(-1)^j}{2^{\sum_{i=1}^{L}(d_i+\mathbb{1}_{\{i\in S\}})}},\tag{7.3}$$

where the product is independent of S and j and can therefore be taken out of the double sum. The double sum equals

$$(1+o(1))\sum_{j=0}^{L}\sum_{\substack{S\subseteq[L']:\\|S|=j}}(-1)^j2^{-j-\sum_{i=1}^{L}d_i}=(1+o(1))2^{-L'-\sum_{i=1}^{L}d_i}.\tag{7.4}$$

Now, recall the definition of the variables $X_s^{(n)}(B), X_{\geq s}^{(n)}(B)$ as in (3.7). Combining (7.2), (7.3) and (7.4) together with the exchangeability of the degree, depth, and label of vertices $1,\dots,K$, we arrive at

$$\begin{aligned}&\mathbb{E}\left[\prod_{m=1}^{K'}\left(X_{s_m}^{(n)}(B_m)\right)_{c_m}\prod_{m=K'+1}^{K}\left(X_{\geq s_m}^{(n)}(B_m)\right)_{c_m}\right]\\&=(n)_L\mathbb{P}\big(\mathcal{D}_{\bar{d}}(L',L)\cap\mathcal{HL}_{\bar{A},\bar{d}}\big)\\&=(1+o(1))2^{L\log_2 n-L'-\sum_{i=1}^{L}d_i}\prod_{i=1}^{L}\mathbb{P}\left(\left(M\sqrt{\frac{a_i'}{4-a_i'}}+N\sqrt{1-\frac{a_i'}{4-a_i'}},M\right)\in A_i\right),\end{aligned}\tag{7.5}$$

since $(n)_L:=n(n-1)\cdots(n-(L-1))=(1+o(1))n^L$. We now recall that there are exactly c_m many d_i,a_i', and A_i that equal $\lfloor\log_2 n\rfloor+s_m,a_m$, and B_m, respectively, for each $m\in[K]$ and that $s_{K'+1}=\ldots=s_K$, so that

$$\prod_{i=1}^{L}\mathbb{P}\left(\left(M\sqrt{\frac{a_i'}{4-a_i'}}+N\sqrt{1-\frac{a_i'}{4-a_i'}},M\right)\in A_i\right)$$
$$=\prod_{m=1}^{K}\mathbb{P}\left(\left(M\sqrt{\frac{a_m}{4-a_m}}+N\sqrt{1-\frac{a_m}{4-a_m}},M\right)\in B_m\right)^{c_m},$$
$$L\log_2 n-L'-\sum_{i=1}^{L}d_i=-\sum_{m=1}^{K'}(s_m+1-\epsilon_n)c_m-\sum_{m=K'+1}^{K}(s_K-\epsilon_n)c_m.$$

Combined with (7.5), this finally yields

$$\mathbb{E}\Bigg[\prod_{m=1}^{K'}\Big(X_{s_m}^{(n)}(B_m)\Big)_{c_m}\prod_{m=K'+1}^{K}\Big(X_{\geq s_m}^{(n)}(B_m)\Big)_{c_m}\Bigg]$$
$$= (1+o(1))\prod_{m=1}^{K'}\left(\mathbb{P}\left(\left(M\sqrt{\frac{a_m}{4-a_m}}+N\sqrt{1-\frac{a_m}{4-a_m}},M\right)\in B_m\right)2^{-(s_m+1)+\epsilon_n}\right)^{c_m}$$
$$\times\prod_{m=K'+1}^{K}\left(\mathbb{P}\left(\left(M\sqrt{\frac{a_m}{4-a_m}}+N\sqrt{1-\frac{a_m}{4-a_m}},M\right)\in B_m\right)2^{-s_K+\epsilon_n}\right)^{c_m}.$$

To prove the second result in Proposition 3.1, we use that for $s_1,\dots,s_K=o(\sqrt{\log n})$,

$$\Big(\frac{h_n(i)-(\log n-d_i/2)}{\sqrt{\log n-d_i/4}},\frac{\log\ell_n(i)-(\log n-d_i/2)}{\sqrt{d_i/4}}\Big)$$
$$=(1+o(1))\Big(\frac{h_n(i)-\mu\log n}{\sqrt{\sigma^2\log n}},\frac{\log\ell_n(i)-\mu\log n}{\sqrt{(1-\sigma^2)\log n}}\Big),$$

and that in this case, $a_m=\lim_{n\to\infty}(s_m+\log_2 n)/\log n=1/\log 2$ for all $m\in[K]$. As a result, noting that

$$\frac{1/\log 2}{4-1/\log 2}=1-\mu/\sigma^2,$$

a similar approach as the above proof for the random variables $\widetilde{X}_s^{(n)}(B),\widetilde{X}_{\geq s}^{(n)}(B)$ yields the desired result. □

8 Proof of Theorems 3.5 and 3.7

In this section we provide the final steps that build on Propositions 5.1 and 6.1 to prove Theorems 3.5 and 3.7. In particular, we show how to include the graph distance between vertices $1,\dots,k$ in the Kingman n-coalescent. As mentioned at the end of Section 3, combining Theorems 3.5 and 3.7 with Corollary 3.4 then immediately implies Theorems 2.4 and 2.6, respectively.

Intuitively, the graph distance between vertices can be related to their (truncated) depth. By the definition of τ_k, the largest common ancestor of any two distinct vertices $i,j\in[k]$ in the random recursive tree has label at most τ_k and hence the sum of the depths and truncated depths of vertices i and j form an upper and lower bound for the graph distance between these vertices in the Kingman n-coalescent, respectively. Since the depth and the truncated depth are asymptotically equal under certain constraints on the truncation sequence t_n (see Lemma 4.11 and Remark 4.12), and since $(\tau_k)_{n\in\mathbb{N}}$ forms a tight sequence of random variables by Lemma 4.9, these bounds on the graph distance are sufficiently sharp. Using the largest common ancestor to provide a lower bound on the graph distance has been used by Munsonius and Rüschendorf for b-are recursive trees [29] and by Ryvkina for random split trees [33], previously.

We formalise the above intuition in the remainder of the section, in which we prove Theorems 3.5 and 3.7.

Proof of Theorem 3.5. We prove (3.6). The proof of (3.5) uses an analogous approach with (5.16) in Proposition 5.3, and hence the proof is omitted.

We set $t_n=\lceil(\log n)^2\rceil$, $\mathcal{D}_k:=\{\mathrm{d}_n(i)\geq d_i,i\in[k]\}$ and, for ease of writing, let $f_i:=\log n-d_i/2$ and recall that d_i diverges with n such that $a_i:=\lim_{n\to\infty}d_i/\log n$ exists for all $i\in[k]$. From Proposition 5.3, we obtain that the tuple

$$\Big(\frac{h_{n,1}(i)-f_i}{\sqrt{\log n-d_i/4}},\frac{\log\ell_n(i)-f_i}{\sqrt{d_i/4}}\Big)_{i\in[k]},$$

conditionally on the event $\mathcal{D}_k$, converges in distribution to

$$\Big(M_i\sqrt{\frac{a_i}{4-a_i}}+N_i\sqrt{1-\frac{a_i}{4-a_i}},M_i\Big)_{i\in[k]},$$

where we recall that $a_i := \lim_{n\to\infty} d_i/\log n$. Moreover, by the choice of t_n at the start of the proof, combining the above result with Lemma 4.11 and Remark 4.12 yields the same result when substituting $h_n(i)$ for $h_{n,1}(i)$. What remains is to include the graph distance between the vertices $1,\dots,k$ to prove Theorem 3.5. We use the trivial upper bound $\mathrm{dist}_n(i,j)\le h_n(i)+h_n(j)$ $i,j\in[n]$ to obtain

$$\begin{aligned}&\Big(\Big(\frac{h_n(i)-f_i}{\sqrt{\log n-d_i/4}},\frac{\log\ell_n(i)-f_i}{\sqrt{d_i/4}}\Big)_{i\in[k]},\Big(\frac{\mathrm{dist}_n(i,j)-(f_i+f_j)}{\sqrt{2\log n-(d_i+d_j)/4}}\Big)_{1\le i<j\le k}\Big)\\&\quad\le\Big(\Big(\frac{h_n(i)-f_i}{\sqrt{\log n-d_i/4}},\frac{\log\ell_n(i)-f_i}{\sqrt{d_i/4}}\Big)_{i\in[k]},\Big(\frac{h_n(i)+h_n(j)-(f_i+f_j)}{\sqrt{2\log n-(d_i+d_j)/4}}\Big)_{1\le i<j\le k}\Big),\end{aligned}\tag{8.1}$$

where the inequality holds element-wise and almost surely. We now observe that

$$\begin{aligned}\frac{h_n(i)+h_n(j)-(f_i+f_j)}{\sqrt{2\log n-(d_i+d_j)/4}}&=\frac{h_n(i)-(f_i)}{\sqrt{\log n-d_i/4}}\sqrt{\frac{\log n-d_i/4}{2\log n-(d_i+d_j)/4}}\\&\quad+\frac{h_n(j)-f_j}{\sqrt{\log n-d_j/4}}\sqrt{\frac{\log n-d_j/4}{2\log n-(d_i+d_j)/4}}.\end{aligned}\tag{8.2}$$

Since $d_i/\log n\to a_i$, it follows that the two deterministic square root terms on the right-hand side converge to $\sqrt{(4-a_i)/(8-(a_i+a_j))}$ and $\sqrt{(4-a_j)/(8-(a_i+a_j))}$, respectively. Furthermore, by the joint convergence of the depth and label of vertices $1,\dots,k$, conditionally on $\mathcal{D}_k$, it thus follows from the continuous mapping theorem [5] and Slutsky's theorem [35, Lemma 2.8], that

$$\begin{aligned}&\Big(\Big(\frac{h_n(i)-f_i}{\sqrt{\log n-d_i/4}},\frac{\log\ell_n(i)-f_i}{\sqrt{d_i/4}}\Big)_{i\in[k]},\Big(\frac{h_n(i)+h_n(j)-(f_i+f_j)}{\sqrt{2\log n-(d_i+d_j)/4}}\Big)_{1\le i<j\le k}\Big)\\&\xrightarrow{d}\Big(\Big(M_i\sqrt{\frac{a_i}{4-a_i}}+N_i\sqrt{1-\frac{a_i}{4-a_i}},M_i\Big)_{i\in[k]},\\&\qquad\Big(\frac{M_i\sqrt{a_i}+N_i\sqrt{4-2a_i}+M_j\sqrt{a_j}+N_j\sqrt{4-2a_j}}{\sqrt{8-(a_i+a_j)}}\Big)_{1\le i<j\le k}\Big).\end{aligned}$$

Combined with (8.1), and letting, for $(x_i,y_i)_{i\in[k]}\in(\mathbb{R}^2)^k$, and $(z_{i,j})_{1\le i<j\le k}\in\mathbb{R}^{k(k-1)/2}$ fixed,

$$\begin{aligned}h_i&:=(\log n-d_i/2)+x_i\sqrt{\log n-d_i/4},\quad\widetilde{\ell}_i:=(\log n-d_i/2)+y_i\sqrt{d_i/4},\\L_{i,j}&:=(2\log n-(d_i+d_j)/2)+z_{i,j}\sqrt{2\log n-(d_i+d_j)/4},\end{aligned}\tag{8.3}$$

this yields,

$$\begin{aligned}&\liminf_{n\to\infty}\mathbb{P}(h_n(i)\le h_i,\log\ell_n(i)\le\widetilde{\ell}_i,i\in[k],\mathrm{dist}_n(i,j)\le L_{i,j},1\le i<j\le k\,|\,\mathcal{D}_k)\\&\quad\ge\mathbb{P}\Big(M_i\sqrt{\frac{a_i}{4-a_i}}+N_i\sqrt{1-\frac{a_i}{4-a_i}}\le x_i,M_i\le y_i,i\in[k],\\&\qquad\frac{M_i\sqrt{a_i}+N_i\sqrt{4-2a_i}+M_j\sqrt{a_j}+N_j\sqrt{4-2a_j}}{\sqrt{8-(a_i+a_j)}}\le z_{i,j},1\le i<j\le k\Big).\end{aligned}\tag{8.4}$$

It remains to obtain a matching lower bound. We make use of the following observation: In the Kingman n-coalescent process, assume two vertices i_1,i_2 are in distinct trees

at step j of the coalescent. Then, the sum of their depths at step j is bounded from above by the graph distance between i_1 and i_2 in the final tree of the coalescent. That is, $h_{F_j}(i_1) + h_{F_j}(i_2) \le \mathrm{dist}_{F_1}(i_1, i_2)$ on the event that i_1, i_2 are in two distinct trees in the forest F_j. See Figure 1 for an example, where the graph distance between vertices 1 and 3 in F_1 is larger than the sum of the depths of 1 and 3 in F_2.

This observation allows us to use the truncated depths $h_{n,1}(i)$ to bound the graph distances between the vertices $1, \dots, k$. Indeed, $h_{n,1}(i) = h_{F_{t_n}}(i)$ denotes the depth of vertex i in the tree at the truncation time t_n. Recall that the event $\{\tau_k < t_n\}$ denotes that the vertices $1, \dots, k$ are in distinct trees at step t_n, which holds with high probability by Lemma 4.9. For $h_i, \widetilde{\ell}_i, L_{i,j}$ as in (8.3), we thus have

$$\begin{aligned}
&\mathbb{P}(h_n(i) \le h_i, \log \ell_n(i) \le \widetilde{\ell}_i, i \in [k], \mathrm{dist}_n(i,j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{D}_k) \\
&\le \mathbb{P}(h_{n,1}(i) \le h_i, \log \ell_n(i) \le \widetilde{\ell}_i, i \in [k], h_{n,1}(i) + h_{n,1}(j) \le L_{i,j}, 1 \le i < j \le k, \tau_k < t_n \,|\, \mathcal{D}_k) \\
&\qquad + \mathbb{P}(\tau_k \ge t_n \,|\, \mathcal{D}_k) \\
&\le \mathbb{P}(h_{n,1}(i) \le h_i, \log \ell_n(i) \le \widetilde{\ell}_i, i \in [k], h_{n,1}(i) + h_{n,1}(j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{D}_k) \\
&\qquad + \mathbb{P}(\tau_k \ge t_n \,|\, \mathcal{D}_k).
\end{aligned}$$

The last term tends to zero with n by Lemma 4.9. With the same approach as in (8.2) and (8.4), we thus obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}(h_n(i) \le h_i, \log \ell_n(i) \le \widetilde{\ell}_i, i \in [k], \mathrm{dist}_n(i,j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{D}_k) \\
&\quad \le \mathbb{P}\Bigg(M_i \sqrt{\frac{a_i}{4-a_i}} + N_i \sqrt{1 - \frac{a_i}{4-a_i}} \le x_i, M_i \le y_i, i \in [k], \\
&\qquad \frac{M_i\sqrt{a_i} + N_i\sqrt{4-2a_i} + M_j\sqrt{a_j} + N_j\sqrt{4-2a_j}}{\sqrt{8-(a_i+a_j)}} \le z_{i,j}, 1 \le i < j \le k \Bigg).
\end{aligned}$$

Combined with the matching lower bound which follows from (8.4), this concludes the proof. □

In a similar spirit, we prove Theorem 3.7. Again, combined with Corollary 3.4, this implies Theorem 2.6.

Proof of Theorem 3.7. The proof follows a similar approach to the proof of Theorem 3.5. Recall the random variables $(d_n^*(i))_{i\in[k]}$ and $(Z_i)_{i\in[k]}$ from (3.9) and set $t_n = \min_{i\in[k]} \log \ell_i$. Proposition 6.1 provides that the tuple

$$\Big(d_n^*(i), \frac{h_{n,1}(i) - \log \ell_i}{\sqrt{\log \ell_i}} \Big)_{i\in[k]}, \tag{8.5}$$

conditionally on the event $\mathcal{L}_k := \{\ell_n(i) = \ell_i, i \in [k]\}$, converges in distribution to $(Z_i, N_i)_{i\in[k]}$, where the N_i are i.i.d. standard normal random variables, also independent of the Z_i. By our choice of t_n, Lemma 4.11 and Remark 4.12 yield that the result holds when $h_{n,1}(i)$ is substituted by $h_n(i)$ as well. As in (8.1), we can use the trivial upper bound $\mathrm{dist}_n(i,j) \le h_n(i) + h_n(i), i, j \in [n]$. We can thus write, similar to (8.2),

$$\begin{aligned}
\frac{\mathrm{dist}_n(i,j) - (\log \ell_i + \log \ell_j)}{\sqrt{\log \ell_i + \log \ell_j}} \le{}& \frac{h_n(i) - \log \ell_i}{\sqrt{\ell_i}} \sqrt{\frac{\log \ell_i}{\log \ell_i + \log \ell_j}} \\
&+ \frac{h_n(j) - \log \ell_j}{\sqrt{\log \ell_j}} \sqrt{\frac{\log \ell_j}{\log \ell_i + \log \ell_j}}.
\end{aligned} \tag{8.6}$$

Define, for $(y_i)_{i\in[k]} \in \mathbb{R}^k, (z_{i,j})_{1\le i<j\le k} \in \mathbb{R}^{k(k-1)/2}$ fixed,

$$h_i := \log \ell_i + y_i\sqrt{\log \ell_i}, \qquad L_{i,j} := (\log \ell_i + \log \ell_j) + z_{i,j}\sqrt{\log \ell_i + \log \ell_j}, \quad 1 \le i < j \le k.$$

Recall the limits $c_{i,j}, c_{j,i}$ of the two square-root terms on the right-hand side of (8.6) from (3.8). We thus obtain, for $(x_i)_{i\in[k]} \in \mathbb{R}^k$ fixed, by (8.6) and (8.5) (and the remark on the $h_n(i)$ below (8.5)) together with the continuous mapping theorem [5],

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}(d_n^*(i) \le x_i, h_n(i) \le h_i, i \in [k], \mathrm{dist}_n(i,j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{L}_k) \\ \ge \mathbb{P}(Z_i \le x_i, N_i \le y_i, i \in [k], c_{i,j}N_i + c_{j,i}N_j \le z_{i,j}, 1 \le i < j \le k)\,.\end{aligned} \tag{8.7}$$

We now use the same observation made after (8.4). That is, on the event $\{\tau_k < t_n\}$, $\mathrm{dist}_n(i,j) \ge h_{n,1}(i) + h_{n,1}(j)$ holds for any two distinct vertices $i, j \in [k]$. We hence have

$$\begin{aligned}&\mathbb{P}(d_n^*(i) \le x_i, h_n(i) \le h_i, i \in [k], \mathrm{dist}_n(i,j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{L}_k) \\ &\le \mathbb{P}(d_n^*(i) \le x_i, h_n(i) \le h_i, i \in [k], h_{n,1}(i) + h_{n,1}(j) \le L_{i,j}, 1 \le i < j \le k, \tau_k < t_n \,|\, \mathcal{L}_k) \\ &\quad + \mathbb{P}(\tau_k \ge t_n \,|\, \mathcal{L}_k) \\ &\le \mathbb{P}(d_n^*(i) \le x_i, h_n(i) \le h_i, i \in [k], h_{n,1}(i) + h_{n,1}(j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{L}_k) \\ &\quad + \mathbb{P}(\tau_k \ge t_n \,|\, \mathcal{L}_k)\,.\end{aligned}$$

The last term on the right-hand side tends to zero by Lemma 4.10. Using the right-hand side of (8.6) to rewrite the event $\{h_{n,1}(i) + h_{n,1}(j) \le L_{i,j}\}$, we thus obtain

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(d_n^*(i) \le x_i, h_n(i) \le h_i, i \in [k], \mathrm{dist}_n(i,j) \le L_{i,j}, 1 \le i < j \le k \,|\, \mathcal{L}_k) \\ \le \mathbb{P}(Z_i \le x_i, N_i \le y_i, i \in [k], c_{i,j}N_i + c_{j,i}N_j \le z_{i,j}, 1 \le i < j \le k)\,,\end{aligned}$$

which matches the lower bound in (8.7) and concludes the proof. □

References

[1] L. Addario-Berry and L. Eslava. High degrees in random recursive trees. *Random Structures & Algorithms*, 52(4):560–575, 2018. MR3809688

[2] K. B. Athreya and S. Karlin. Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *The Annals of Mathematical Statistics*, 39(6):1801–1817, 1968. MR0232455

[3] S. Banerjee and S. Bhamidi. Persistence of hubs in growing random networks. *Probability Theory and Related Fields*, pages 1–63, 2021. MR4288334

[4] S. Bhamidi. Universal techniques to analyze preferential attachment trees: Global and local analysis. *Preprint available at http://www.unc.edu/~bhamidi*, 2007.

[5] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics, second edition, 1999. MR1700749

[6] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. II*. Probability and its Applications (New York). Springer, New York, second edition, 2008. General theory and structure. MR2371524

[7] L. Devroye. Applications of the theory of records in the study of random trees. *Acta Informatica*, 26(1):123–130, 1988. MR0969872

[8] L. Devroye and J. Lu. The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures & Algorithms*, 7(1):1–14, 1995. MR1346281

[9] R. P. Dobrow. On the distribution of distances in recursive trees. *Journal of Applied Probability*, 33(3):749–757, 1996. MR1401472

[10] M. Drmota. *Random trees: an interplay between combinatorics and probability*. Springer Science & Business Media, 2009. MR2484382

[11] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. MR3930614

[12] L. Eslava. Depth of vertices with high degree in random recursive trees. *arXiv preprint arXiv:1611.07466*, 2016. MR4436029

[13] L. Eslava. A non-increasing tree growth process for recursive trees and applications. *Combinatorics, Probability and Computing*, 30(1):79–104, 2021. MR4205660

[14] L. Eslava, B. Lodewijks, and M. Ortgiese. Fine asymptotics for the maximum degree in weighted recursive trees with bounded random weights. *Preprint, arXiv:2109.15270*, 2021.

[15] Q. Feng, J. Liu, and C. Su. A note on the distance in random recursive trees. *Statistics & probability letters*, 76(16):1748–1755, 2006. MR2274136

[16] J. L. Gastwirth and P. Bhattacharya. Two probability models of pyramid or chain letter schemes demonstrating that their promotional claims are unreliable. *Operations Research*, 32(3):527–536, 1984. MR0755999

[17] W. Goh and E. Schmutz. Limit distribution for the maximum degree of a random recursive tree. *Journal of computational and applied mathematics*, 142(1):61–82, 2002. MR1910519

[18] T. Iyer. Degree distributions in recursive trees with fitnesses. *Preprint, arXiv:2005.02197*, 2020.

[19] S. Janson. Functional limit theorems for multitype branching processes and generalized pólya urns. *Stochastic Processes and their Applications*, 110(2):177–245, 2004. MR2040966

[20] S. Janson. Asymptotic degree distribution in random recursive trees. *Random Structures & Algorithms*, 26(1-2):69–83, 2005. MR2116576

[21] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*. Wiley-Interscience Series, New York, 2000. MR1782847

[22] M. Kuba and A. Panholzer. On the degree distribution of the nodes in increasing trees. *Journal of Combinatorial Theory, Series A*, 114(4):597–618, 2007. MR2319165

[23] B. Lodewijks. Location of maximum degree vertices in weighted recursive graphs with bounded random weights. *arXiv preprint arXiv:2110.00522*, 2021.

[24] H. M. Mahmoud. Limiting distributions for path lengths in recursive trees. *Probability in the Engineering and Informational Sciences*, 5(1):53–59, 1991. MR1183165

[25] H. M. Mahmoud and G. S. Lueker. *Evolution of random search trees*, volume 200. Wiley New York, 1992. MR1140708

[26] H. M. Mahmoud and R. T. Smythe. Asymptotic joint normality of outdegrees of nodes in random recursive trees. *Random Structures & Algorithms*, 3(3):255–266, 1992. MR1164839

[27] A. Meir and J. Moon. Recursive trees with no nodes of out-degree one. *Congressus Numerantium*, 66:49–62, 1988. MR0992887

[28] J. W. Moon. The distance between nodes in recursive trees. *London Mathematical Society Lecture Notes*, 13:125–132, 1974. MR0357186

[29] G. O. Munsonius and L. Rüschendorf. Limit theorems for depths and distances in weighted random b-ary recursive trees. *Journal of applied probability*, 48(4):1060–1080, 2011. MR2896668

[30] H. S. Na and A. Rapoport. Distribution of nodes of a tree by degree. *Mathematical Biosciences*, 6:313–329, 1970. MR0278985

[31] D. Najock and C. Heyde. On the number of terminal vertices in certain random trees with an application to stemma construction in philology. *Journal of Applied Probability*, 19(3):675–680, 1982. MR0664852

[32] B. Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures & Algorithms*, 5(2):337–347, 1994. MR1262983

[33] J. Ryvkina. *Ein universeller zentraler Grenzwertsatz für den Abstand zweier Kugeln in zufälligen Splitbäumen*. PhD thesis, Frankfurt am Main, Johann Wolfgang Goethe-Univ., Diplomarbeit, 2008, 2008.

[34] J. Szymanski. On the maximum degree and the height of a random recursive tree. In *Random graphs*, volume 87, pages 313–324, 1990. MR1094139

[35] A. W. van der Vaart. *Asymptotic statistics,* volume 3. Cambridge university press, 2000. MR1652247

Acknowledgments. Bas Lodewijks would like to thank Laura Eslava for some useful discussions related to the Kingman n-coalescent and for providing the source code of the figures in this paper.

He would also like to thank the anonymous referees for providing helpful suggestions which led to an improved presentation of the results and proofs.

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: `https://vtex.lt/services/ejms-peer-review/`
[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: `https://imstat.org/shop/donation/`