Multivariate normal limit laws for the numbers of fringe subtrees in m-ary search trees and preferential attachment trees

Cecilia Holmgren* Svante Janson†

Department of Mathematics
Uppsala University
PO Box 480
751 06 Uppsala, Sweden
{cecilia.holmgren,svante}@math.uu.se

Matas Šileikis

Department of Applied Mathematics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Praha, Czech Republic
matas.sileikis@gmail.com

Submitted: Aug 12, 2016; Accepted: Jun 22, 2017; Published: Jun 30, 2017
Mathematics Subject Classifications: 60C05; 05C05; 05C80; 60F05; 68P05; 68P10

Abstract

We study fringe subtrees of random m-ary search trees and of preferential attachment trees, by putting them in the context of generalised Pólya urns. In particular we show that for the random m-ary search trees with $m \leqslant 26$ and for the linear preferential attachment trees, the number of fringe subtrees that are isomorphic to an arbitrary fixed tree T converges to a normal distribution; more generally, we also prove multivariate normal distribution results for random vectors of such numbers for different fringe subtrees. Furthermore, we show that the number of protected nodes in random m-ary search trees for $m \leqslant 26$ has asymptotically a normal distribution.

Keywords: Random trees; Fringe trees; Normal limit laws; Pólya urns; m-ary search trees; Preferential attachment trees; Protected nodes

1 Introduction

The main focus of this paper is to consider fringe subtrees of random m-ary search trees and of general preferential attachment trees (including the random recursive tree); these random trees are defined in Section 2. Recall that a *fringe subtree* is a subtree consisting of some node and all its descendants, see Aldous [1] for a general theory, and note that

*Partly supported by the Swedish Research Council.
†Partly supported by the Knut and Alice Wallenberg Foundation.

fringe subtrees typically are "small" compared to the whole tree. (All subtrees considered in the present paper are of this type, and we will use 'subtree' and 'fringe subtree' as synonyms.)

We will use (generalised) Pólya urns to analyze vectors of the numbers of fringe subtrees of different types in random m-ary search trees and general (linear) preferential attachment trees, and in the former class we will also analyze the number of protected nodes (that is, nodes with distance to a nearest leaf at least two). As a result, we prove multivariate normal asymptotic distributions for these random variables, for m-ary search trees when $m \leqslant 26$ and for preferential attachment trees with linear weights.

Pólya urns have earlier been used to study the total number of nodes in random m-ary search trees, see [33, 24, 34]. In that case one only needs to consider an urn with $m-1$ different types, describing the nodes holding i keys, where $i \in \{0, 1, \dots, m-2\}$. For this case it is well-known that asymptotic normality does not hold for m-ary search trees with $m > 26$, see [6]. Recently, in [22] more advanced Pólya urns (with $\binom{2m}{m-1}$ types) were used to describe protected nodes in random m-ary search trees. Only the cases $m = 2, 3$ were treated in detail in [22], and the cases $m = 4, 5, 6$ were further treated in [20]. In [22] a simpler urn (similar to the urn describing the total number of nodes) was also used to describe the total number of leaves in random m-ary search trees.

In this work we further extend the approach used in [22] for analyzing arbitrary fringe subtrees of a fixed size in random m-ary search trees as well as in preferential attachment trees. For the random m-ary search trees we furthermore extend the methods used in [22] and in [20] to analyze the number of protected nodes in m-ary search trees for $m \leqslant 26$.

Remark 1.1. The Pólya urns yield asymptotic results for the numbers of fringe subtrees in m-ary search trees for $m \geqslant 27$ too, using [24, Theorem 3.24] or [40], but the normalization is different and the (subsequence) limits are presumably not normal. In fact, our proofs show that for any m, the second largest (in real part) eigenvalue in any of our Pólya urns is the same as the one in the simple Pólya urn mentioned above for counting the total number of nodes, see Theorem 6.2. As is well-known (see e.g. Theorem 4.1 and [24]), the asymptotic behaviour depends crucially on the second largest eigenvalue, so we expect the same type of asymptotic behavior as for the number of nodes, which for $m \geqslant 27$ is not normal [6], see also [4]. We will not consider this case further in the present paper.

1.1 Composition of the paper

The m-ary search trees and general preferential attachment trees are defined in Section 2. Our main results are presented in Section 3; the results in Section 3.1 and in Section 3.2 concern the case of random m-ary search trees and the results in Section 3.3 concern the case of linear preferential attachment trees.

The results in the case of the random m-ary search trees are extensions of results that previously have been shown for the special case of the random binary search tree with the use of other methods, see e.g., [8, 9, 15, 21]. Furthermore, the results for the m-ary search trees in Section 3.2 (where we consider applications to protected nodes in such

trees) are extensions of the results that were proved for the random binary search trees using other methods in [38, 21], and extensions of both the results and the methods that were used for $m = 2, 3$ in [22] and for $m = 4, 5, 6$ in [20]. The results for the preferential attachment trees in Section 3.3 are extensions of results that previously have been shown (using different methods) for the random recursive trees, see e.g., [21] and [18].

In particular we show that for the random m-ary search trees with $m \leqslant 26$ and for the linear preferential attachment trees, the number of fringe subtrees that are isomorphic to an arbitrary fixed tree T has an asymptotic normal distribution; more generally, we also prove multivariate normal distribution results for random vectors of such numbers for different trees.

In Section 4 we describe the theory of generalised Pólya urns developed in [24] that we use in our proofs.

In Section 5 we describe the specific Pólya urns that we use for analyzing fringe subtrees in random m-ary search trees, and in Section 6 we use them to prove the main results for m-ary search trees in Section 3.1. Similarly, in Section 7 we describe the specific Pólya urns that we use for analyzing fringe subtrees in preferential attachment trees, and in Section 8 we use them to prove the main results for preferential attachment trees in Section 3.3. In Section 9 we describe the specific Pólya urns that we use for analyzing protected nodes in random m-ary search trees, and in Section 10 we use them to prove the result on protected nodes in m-ary search trees in Section 3.2.

In Section 11 we present some examples with explicit calculations.

Finally, in Section 12 we use related but simpler Pólya urns to analyze the out-degrees of the nodes in the random trees.

2 The random trees

2.1 m-ary search trees

We recall the definition of m-ary search trees, see e.g. [32] or [11]. An m-ary search tree, for an integer $m \geqslant 2$, is constructed recursively from a sequence of n *keys* (real numbers); we assume that the keys are distinct. Each node may contain up to $m-1$ keys. We start with a tree containing just an empty root. The first $m-1$ keys are put in the root, and are placed in increasing order from left to right; they divide the set of real numbers into m intervals $J_1, \dots, J_m$. When the root is full (after the first $m-1$ keys are added), it gets m children that are initially empty, and each further key is passed to one of the children depending on which interval it belongs to; a key in J_i is passed to the ith child. (The binary search tree, i.e., the case $m = 2$, is the simplest case where keys are passed to the left or right child depending on whether they are larger or smaller than the key in the root.) The procedure repeats recursively in the subtrees until all keys are added to the tree.

We are primarily interested in the random case when the keys form a uniformly random permutation of $\{1, \dots, n\}$, and we let $\mathcal{T}_n$ denote the random m-ary search tree constructed from such keys. Only the order of the keys matter, so alternatively, we may assume that

the keys are n i.i.d. uniform random numbers in $[0,1]$. Moreover, considering an infinite sequence of i.i.d. keys, and defining $\mathcal{T}_n$, for $n = 1, 2, \dots$ as the tree constructed from the n first keys, we obtain a Markov process $(\mathcal{T}_n)_{n=1}^{\infty}$.

Nodes that contain at least one key are called *internal*, while empty nodes are called *external*. We regard the m-ary search tree as consisting only of the internal nodes; the external nodes are places for potential additions, and are useful when discussing the tree (e.g. below), but are not really part of the tree. (However, the positions of the external nodes are significant. For example, when a node in a binary search tree has exactly one internal child, we want to know whether that is a left or a right child.) Thus, a *leaf* is an internal node that has no internal children, but it may have external children. (It will have external children if it is full, but not otherwise.)

From now on, when considering an m-ary search tree, we will ignore the values of the keys, but we will keep track of the number of keys in each node. Hence, a non-random m-ary search tree is a (finite) ordered rooted tree where each node is marked with the number of keys it contains, with this number being in $\{0, \dots, m-1\}$, and such that if we include the external nodes, the nodes with $m-1$ keys have exactly m children while the remaining nodes have no children.

We say that a node (external or internal) with $i \leqslant m-2$ keys has $i+1$ *gaps*, while a full node has no gaps. It is easily seen that an m-ary search tree with n keys has $n+1$ gaps; the gaps correspond to the intervals of real numbers between the keys (and $\pm\infty$).

If we condition on the isomorphism class of $\mathcal{T}_n$ (or even on the underlying permutation), then a new key has the same probability $1/(n+1)$ of being inserted into any of the $n+1$ gaps. Thus the $\mathcal{T}_{n+1}$ is obtained from $\mathcal{T}_n$ by choosing gap uniformly at random and inserting a key there.

Remark 2.1. In applications where the order of the children of a node does not matter, we can simplify things by ignoring the order and regard the m-ary search tree as an unordered tree. (In this case, we can also ignore the external nodes completely.)

If we treat $(\mathcal{T}_n)_{n=1}^{\infty}$ as a sequence of unordered trees without external nodes, then without external nodes $\mathcal{T}_{n+1}$ is obtained from $\mathcal{T}_n$ by choosing a node with probability proportional to $k+1-l$, where k is the number of keys in the node and l is the number of children of the node (of course $l = 0$, if $k < m-1$), and giving this node a new key if $k < m-1$ and a new child if $k = m-1$.

Remark 2.2. Each permutation of $\{1, \dots, n\}$ defines an m-ary search tree; however, different permutations may define the same m-ary search tree. It is possible to obtain a bijection by giving each key in the m-ary search tree a time stamp, which is its number in the sequence of keys used to construct the tree. (For binary trees we thus obtain so-called increasing trees, see e.g. [11].) This gives a labelled version of m-ary search trees. In this context two trees T and T' are isomorphic if there is an isomorphism with the additional property that it maps the ith largest time stamp of T to the ith largest time stamp of T'.

2.2 General preferential attachment trees

Suppose that we are given a sequence of non-negative weights $(w_k)_{k=0}^{\infty}$, with $w_0 > 0$. Grow a random unordered tree Λ_n (with n nodes) recursively, starting with a single node and adding nodes one by one. Each new node is added as a child of some randomly chosen existing node; when a new node is added to Λ_{n-1}, the probability of choosing a node $v \in \Lambda_{n-1}$ as the parent is proportional to $w_{d^+(v)}$, where $d^+(v)$ is the out-degree of v in Λ_{n-1}. (More formally, this is the conditional probability, given Λ_{n-1} and the previous history. The sequence $(\Lambda_n)_{n=1}^{\infty}$ thus constitutes a Markov process.)

We will mainly consider the case of *linear preferential attachment trees*, i.e., when

$$w_k = \chi k + \rho, \tag{1}$$

for some real parameters χ and ρ, with $\rho = w_0 > 0$; this includes the most studied cases of preferential attachment trees. Note that we obtain the same random trees T_n if we multiply all w_k by some positive constant. Hence, only the quotient χ/ρ matters, and it suffices to consider $\chi \in \{1, 0, -1\}$. In the case $\chi = -1$, so $w_k = \rho - k$, w_k is eventually negative. This is not allowed; however, this is harmless if (and only if) $\rho = m$ is an integer; then $w_m = 0$ so no node ever gets more than m children and thus the values w_k for $k > m$ do not matter and can be replaced by 0. (We exclude the trivial case $\chi = -1$, $\rho = 1$, when $w_1 = 0$ so no node ever gets more than one child and the tree Λ_n deterministically is a path with n nodes.)

Example 2.3. The *random recursive tree* is constructed recursively by adding nodes one by one, with each new node attached as a child of a uniformly randomly chosen existing node, see [11, Section 1.3.1]. Hence, this is the case $w_k = 1$ for all k, which is a special case of a linear preferential attachment tree (1) with $\chi = 0$ and $\rho = 1$. (Any $\rho > 0$ yields the same tree when $\chi = 0$.)

Example 2.4. The random *plane oriented recursive tree*, introduced by Szymański [41], is constructed similarly to the random recursive tree, but we now consider the trees as ordered; an existing node with k children thus has $k + 1$ positions in which a new node can be added, and we give all possible positions of the new node the same probability. The probability of choosing a node v as the parent is thus proportional to $d^+(v) + 1$, so the plane oriented recursive tree is the case of a linear preferential attachment tree with $w_k = k + 1$, i.e., $\chi = \rho = 1$.

This model with $w_k = k + 1$ is also the preferential attachment model by Barabási and Albert [2], which has become popular and has been studied by many authors, as has the generalization $w_k = k + \rho$ with arbitrary $\rho > 0$, i.e., (1) with $\chi = 1$.

Example 2.5. The binary search tree is the special case with $w_0 = 2$, $w_1 = 1$ and $w_k = 0$, $k \geqslant 2$ (and, furthermore, each first child randomly assigned to be left or right); as said above, we may regard this as the case $\chi = -1$ and $\rho = 2$ of (1). However, m-ary search trees with $m \geqslant 3$ are not preferential attachment trees.

Remark 2.6. It is often natural to consider preferential attachment trees as unordered; it is also possible to consider them as ordered, either by assigning random orders as in Example 2.4 or by ordering the children of each node in the order that they are added to the tree.

For further descriptions of preferential attachment trees, see e.g., [23, Section 6].

3 Main results

In this section we state the results on fringe subtrees and protected nodes in random m-ary search trees as well as fringe trees in preferential attachment trees.

3.1 Fringe subtrees in random m-ary search trees

Remark 3.1. As said in the introduction, m-ary search trees can be regarded as either ordered or unordered trees; it is further possible to consider the labelled version as in Remark 2.2. (See also [21, Remark 1.2] for the special case of the binary search tree.) The most natural interpretation is perhaps the one as ordered trees, and it immediately implies the corresponding result for unordered trees in, for example, Theorem 3.2. However, in some applications it is preferable to regard the fringe trees as unordered trees, since this gives fewer types to consider in the Pólya urns that we use, see, e.g., Example 5.1 and Section 9. The theorems in this section apply to all these interpretations, via the choice of an appropriate notion of isomorphism.

The following theorem generalises [21, Theorem 1.22], where the special case of the binary search tree was analyzed.

Let $H_m := \sum_{k=1}^{m} 1/k$ be the mth harmonic number. Here and below we write $T = T'$ whenever two trees T and T' are isomorphic.

Theorem 3.2. *Assume that $2 \leqslant m \leqslant 26$. Let $T^1, \dots, T^d$ be a fixed sequence of non-isomorphic non-random m-ary search trees and let $\mathbf{Y}_n = \big(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\big)$, where $X_n^{T^i}$ is the (random) number of fringe subtrees that are isomorphic to T^i in the random m-ary search tree $\mathcal{T}_n$ with n keys. Let k_i be the number of keys of T^i for $i \in \{1, \dots, d\}$. Let*

$$\boldsymbol{\mu}_n := \mathbb{E}\,\mathbf{Y}_n = \Big(\mathbb{E}(X_n^{T^1}), \mathbb{E}(X_n^{T^2}), \dots, \mathbb{E}(X_n^{T^d})\Big).$$

Then

$$n^{-1/2}(\mathbf{Y}_n - \boldsymbol{\mu}_n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{2}$$

where $\Sigma = (\sigma_{ij})_{i,j=1}^{d}$ is some covariance matrix. Furthermore, in (2), the vector $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n := n\hat{\boldsymbol{\mu}}$ with

$$\hat{\boldsymbol{\mu}} := \Big(\frac{\mathbb{P}(\mathcal{T}_{k_1} = T^1)}{(H_m - 1)(k_1 + 1)(k_1 + 2)}, \dots, \frac{\mathbb{P}(\mathcal{T}_{k_d} = T^d)}{(H_m - 1)(k_d + 1)(k_d + 2)}\Big). \tag{3}$$

Moreover, if the trees $T^1, \dots, T^d$ have at least one internal node each, then the covariance matrix Σ is non-singular.

Remark 3.3. The fact that $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n$ means that

$$\mathbb{E}\bigl(X_n^{T^i}\bigr) = \frac{\mathbb{P}(\mathcal{T}_{k_i} = T^i)}{(H_m - 1)(k_i + 1)(k_i + 2)} n + o\bigl(n^{1/2}\bigr). \tag{4}$$

A weaker version of (4) with the error term $o(n)$ follows, for any $m \geqslant 2$, from the branching process analysis of fringe subtrees in [23], see the proof in Section 6. Moreover, the proof also shows that (4) holds, for any $m \geqslant 2$, with the error term $O\bigl(n^{\max(\gamma_m,0)}\bigr)$, where γ_m is the second largest real part of a root of the polynomial ϕ_m in Theorem 6.2; if $m \leqslant 26$ then $\gamma_m < \frac{1}{2}$ (yielding (4)) but if $m \geqslant 27$ then $\gamma_m > \frac{1}{2}$, as shown by [35] and [14].

The vector $\hat{\boldsymbol{\mu}}_n$ can also, using (30) below, be calculated from an eigenvector of the intensity matrix of the Pólya urn defined in Section 5, see Theorem 4.1(i). See also [27].

Also the covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1}^d$ can be calculated explicitly from the intensity matrix of the Pólya urn, see Theorem 4.1(ii)–(iii). We give one example in Section 11. The results in [27] also show

$$\sigma_{ij} = \lim_{n\to\infty} \frac{1}{n} \operatorname{Cov}\bigl(X_n^{T^i}, X_n^{T^j}\bigr). \tag{5}$$

More generally, it follows from the results in [28] that all moments converge in (2), see Remark 6.3. Similarly, as a consequence of [28] and [27], moment convergence and, in particular, asymptotics of variance and covariances as in (5) hold in all theorems in this section.

Remark 3.4. The covariance matrix may be singular if some T^i is the tree consisting of a single external node. For example, if $d = m - 1$ and T^i is the tree consisting of a single node with $i - 1$ keys, and thus i gaps, then, by counting the number of gaps in the tree $\mathcal{T}_n$,

$$\sum_{i=1}^{d} i X_n^{T^i} = n + 1, \tag{6}$$

so this sum is deterministic, and thus the covariance matrix is singular. (In particular, if $m = 2$, then the number of external nodes is deterministic, namely $n + 1$.) Moreover, (6) shows that the number of external nodes X^{T^1} is an affine function of the numbers X^{T^j}, $j \geqslant 2$; thus it is always possible to reduce to the case when every tree T^i has at least one internal node and Σ is non-singular.

The following theorem is an important corollary of Theorem 3.2. It also follows from Fill and Kapur [14, Theorem 5.1]. The special case of the random binary search tree was proved by Devroye [8], and the covariances for $Y_{n,k}$ in that case were given by Dennert and Grübel [7], see also [21, Theorem 1.19 and Proposition 1.10] and the references therein.

Theorem 3.5. *Assume that $2 \leqslant m \leqslant 26$. Let $k \geqslant 0$ be an arbitrary fixed integer and let $Y_{n,k}$ be the (random) number of fringe subtrees with k keys in the random m-ary search tree $\mathcal{T}_n$ with n keys. Then, as $n \to \infty$,*

$$n^{-1/2}\bigl(Y_{n,k} - \mathbb{E}\, Y_{n,k}\bigr) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{7}$$

where σ_k^2 is some constant with $\sigma_k^2 > 0$ except when $k = 0$ and $m = 2$. We also have

$$n^{-1/2}\Big(Y_{n,k} - \frac{n}{(H_m - 1)(k+1)(k+2)}\Big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2). \tag{8}$$

Remark 3.6. The asymptotic mean $\frac{n}{(H_m-1)(k+1)(k+2)}$ in (8) easily follows from (4), see the proof in Section 6. The constant σ_k^2 can again be calculated explicitly from our proof.

We give one example of Theorem 3.5 in Section 11.1, where we let $m = 3$ and $k = 4$.

3.2 Protected nodes in random m-ary search trees

There are many recent studies of so-called protected nodes in various classes of random trees, see e.g. [3, 5, 10, 12, 38, 39, 21, 22, 23]. A node is *protected* (more precisely, two-protected) if it is not a leaf and none of its children is a leaf.

The following result was proved by using Pólya urns in [22, Theorem 1.1] for $m = 3$ and in [20] for $m = 4$, 5 and 6.

Theorem 3.7. *Let Z_n be the number of protected nodes in the random m-ary search tree $\mathcal{T}_n$ with n keys. Then, if $m \leqslant 26$, we have*

$$n^{-1/2}\big(Z_n - \mathbb{E}\, Z_n\big) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\right), \tag{9}$$

where σ^2 is some positive constant. Furthermore, $\mathbb{E}\, Z_n$ can be replaced by μn with

$$\mu := \frac{1}{m(H_m - 1)} \sum_{\ell=0}^{m-1} \frac{m!}{(m-\ell)!} \cdot \frac{(m(m-\ell))!}{(m(m-\ell)+\ell+1)!}. \tag{10}$$

Remark 3.8. The fact that $\mathbb{E}\, Z_n$ can be replaced by μn means that

$$\mathbb{E}(Z_n) = \mu n + o\big(n^{1/2}\big). \tag{11}$$

As in Remark 3.3, a weaker version with $o(n)$ follows for any $m \geqslant 2$ from [23], see the proof in Section 10. Moreover, our proof shows that (11) holds, for any $m \geqslant 2$, with the error term $O\big(n^{\max(\gamma_m, 0)}\big)$, with γ_m as in Remark 3.3.

The constant σ^2 can be calculated explicitly from our proof of Theorem 3.7. For examples of Theorem 3.7 with explicit calculations of the asymptotic variance σ^2, we refer the reader to [22] for $m = 3$ and [20] for $m = 4$.

3.3 Fringe subtrees in preferential attachment trees

The following theorem was proved for the random recursive tree in [21, Theorem 1.22] using Stein's method, and (under a technical condition) by Gopaladesikan, Mahmoud and Ward [18] using the contraction method (see also Feng and Mahmoud [13] for the univariate case).

Here we give a generalisation to the linear preferential attachment trees. The result applies to all three versions of the tree as mentioned in Remark 2.6 with the notion of isomorphism chosen appropriately.

Theorem 3.9. *Let $\Lambda^1, \dots, \Lambda^d$ be a fixed sequence of non-isomorphic unordered (or ordered) trees and let $\mathbf{Z}_n = (X_n^{\Lambda^1}, X_n^{\Lambda^2}, \dots, X_n^{\Lambda^d})$, where $X_n^{\Lambda^i}$ is the number of fringe subtrees that are isomorphic to Λ^i in the linear preferential attachment tree Λ_n. Let k_i be the number of nodes in Λ^i. Let*

$$\boldsymbol{\mu}_n := \mathbb{E}\,\mathbf{Z}_n = \Big(\mathbb{E}(X_n^{\Lambda^1}), \mathbb{E}(X_n^{\Lambda^2}), \dots, \mathbb{E}(X_n^{\Lambda^d})\Big).$$

Then

$$n^{-1/2}(\mathbf{Z}_n - \boldsymbol{\mu}_n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{12}$$

where the vector $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n := n\hat{\boldsymbol{\mu}}$ with

$$\hat{\boldsymbol{\mu}} := \Big(\frac{\mathbb{P}(\Lambda_{k_1} = \Lambda^1)\cdot\kappa}{(k_1+\kappa-1)(k_1+\kappa)}, \dots, \frac{\mathbb{P}(\Lambda_{k_d} = \Lambda^d)\cdot\kappa}{(k_d+\kappa-1)(k_d+\kappa)}\Big), \tag{13}$$

with

$$\kappa := \frac{\rho}{\chi+\rho} = \frac{w_0}{w_1}, \tag{14}$$

and $\Sigma = (\sigma_{ij})_{i,j=1}^d$ is some non-singular covariance matrix.

Note that for the random recursive tree $\kappa = 1$ and for the plane oriented recursive tree $\kappa = \frac{1}{2}$.

Remark 3.10. The proof shows also that

$$\mathbb{E}(X_n^{\Lambda^i}) = \frac{\mathbb{P}(\Lambda_{k_i} = \Lambda^i)\cdot\kappa}{(k_i+\kappa-1)(k_i+\kappa)}\, n + O(1). \tag{15}$$

A weaker version of (15) with the error term $o(n)$ follows from the branching process analysis of fringe subtrees, see [23, (5.29) and Example 6.4, in particular (6.24)].

The vector $\hat{\boldsymbol{\mu}}$, and thus the coefficient of n in (15), can also be calculated from an eigenvector of the intensity matrix in the proof; similarly, the covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1}^d$ can be calculated explicitly from our proof.

The following theorem is an important corollary of Theorem 3.9. The cases of the random recursive tree ($\kappa = 1$) and binary search tree ($\kappa = 2$) were proved in [8, Theorems 4 and 5] and the case of the plane oriented recursive tree ($\kappa = \frac{1}{2}$) was proved in [17, Theorem 1.1].

Theorem 3.11. *Let k be an arbitrary fixed integer. Let $Y_{n,k}$ be the number of subtrees with k nodes in the linear preferential attachment tree Λ_n. Then, as $n \to \infty$,*

$$n^{-1/2}\big(Y_{n,k} - \mathbb{E}\,Y_{n,k}\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{16}$$

where σ_k^2 is some constant with $\sigma_k^2 > 0$. Furthermore, we also have

$$n^{-1/2}\Big(Y_{n,k} - \frac{\kappa}{(k+\kappa-1)(k+\kappa)}\, n\Big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{17}$$

with κ as in (14).

Remark 3.12. It follows from (15), see the proof, that

$$\mathbb{E}\bigl(Y_{n,k}\bigr) = \frac{\kappa}{(k+\kappa-1)(k+\kappa)}\, n + O(1). \tag{18}$$

The constant σ_k^2 can again be calculated explicitly from our proof.

We give one example in Section 11.3, where we let $k = 3$.

4 Generalised Pólya urns

A (generalised) Pólya urn process is defined as follows, see e.g. [24] or [34]. There are balls of q types (or colours) $1, \dots, q$, and for each n a random vector $\mathcal{X}_n = (X_{n,1}, \dots, X_{n,q})$, where $X_{n,i}$ is the number of balls of type i in the urn at time n. The urn starts with a given vector $\mathcal{X}_0$. For each type i, there is an activity (or weight) $a_i \in \mathbb{R}_{\geqslant 0}$, and a random vector $\xi_i = (\xi_{i1}, \dots, \xi_{iq})$. The urn evolves according to a discrete time Markov process. At each time $n \geqslant 1$, assuming there is a ball of positive activity (see assumption A7 below), one ball is drawn at random from the urn, with the probability of any ball proportional to its activity. Thus, the drawn ball has type i with probability $\frac{a_i X_{n-1,i}}{\sum_j a_j X_{n-1,j}}$. If the drawn ball has type i, it is replaced together with $\Delta X_{n,j}^{(i)}$ balls of type j, $j = 1, \dots, q$, where the random vector $\Delta\mathcal{X}_n^{(i)} = (\Delta X_{n,1}^{(i)}, \dots, \Delta X_{n,q}^{(i)})$ has the same distribution as ξ_i and is independent of everything else that has happened so far. We allow $\Delta X_{n,i}^{(i)} = -1$, which means that the drawn ball is *not* replaced.

Usually, the random variables $X_{n,i}$ and ξ_{ij} are integer-valued, with $X_{n,i} \geqslant 0$, in accordance with the interpretation as numbers of balls; we assume this unless we explicitly make an exception. However, see Remark 4.4 for an extension, which will be used in Section 7.

The *intensity matrix* of the Pólya urn is the $q \times q$ matrix

$$A := (a_j\, \mathbb{E}\, \xi_{ji})_{i,j=1}^{q}. \tag{19}$$

The intensity matrix A with its eigenvalues and eigenvectors is central for proving limit theorems. As noted in [24], $\alpha I + A$ has all non-negative entries for a sufficiently large $\alpha > 0$ and thus by the standard Perron-Frobenius theory, see e.g., [30, Appendix 2], A has a real eigenvalue λ_1 such that all eigenvalues $\lambda \neq \lambda_1$ satisfy $\operatorname{Re}\lambda < \lambda_1$.

The basic assumptions in [24] are the following. We say that a type i is *dominating*, if every other type j can be found with positive probability at some time in an urn started with a single ball of type i. The urn (and its matrix A) is *irreducible* if every type is dominating.

(A1) $\xi_{ij} \geqslant 0$ for $j \neq i$ and $\xi_{ii} \geqslant -1$. (I.e., the drawn ball may be removed, but no other ball.)

(A2) $\mathbb{E}(\xi_{ij}^2) < \infty$ for all $i, j \in \{1, \dots, q\}$.

(A3) The largest real eigenvalue λ_1 of A is positive.

(A4) The largest real eigenvalue λ_1 is simple.

(A5) There exists a dominating type i with $X_{0,i} > 0$, i.e., we start with at least one ball of a dominating type.

(A6) λ_1 is an eigenvalue of the submatrix of A given by the dominating types.

We will also use the following simplifying assumption.

(A7) At each time $n \geqslant 1$, there exists a ball of a dominating type.

Before stating the results that we use, we need some notation. By a vector v we mean a column vector, and we write v' for its transpose (a row vector). More generally, we denote the transpose of a matrix A by A'. By an eigenvector of A we mean a right eigenvector; a left eigenvector is the same as the transpose of an eigenvector of the matrix A'. If u and v are vectors then $u'v$ is a scalar while uv' is a $q \times q$ matrix of rank 1. We also use the notation $u \cdot v$ for $u'v$. Let $a = (a_1, \dots, a_q)$ denote the (column) vector of activities, and let u_1' and v_1 denote left and right eigenvectors of A corresponding to the eigenvalue λ_1, i.e., vectors satisfying

$$u_1' A = \lambda_1 u_1', \qquad\qquad Av_1 = \lambda_1 v_1.$$

We assume that v_1 and u_1 are normalised so that

$$a \cdot v_1 = a'v_1 = v_1'a = 1, \qquad\qquad u_1 \cdot v_1 = u_1'v_1 = v_1'u_1 = 1, \tag{20}$$

see [24, equations (2.2)–(2.3)]. We write $v_1 = (v_{11}, \dots, v_{1q})$.

We define

$$P_{\lambda_1} = v_1 u_1',$$

and $P_I = I_q - P_{\lambda_1}$, where I_q is the $q \times q$ identity matrix. (Thus P_{λ_1} is the one-dimensional projection onto the eigenspace corresponding to λ_1 such that P_{λ_1} commutes with the matrix A, see [24, equation (2.5)]; note that P_{λ_1} typically is not orthogonal). We define the matrices

$$B_i := \mathbb{E}(\xi_i \xi_i'), \tag{21}$$

$$B := \sum_{i=1}^{q} v_{1i} a_i B_i. \tag{22}$$

In the case when $\operatorname{Re}\lambda < \lambda_1/2$ for every eigenvalue $\lambda \neq \lambda_1$, we define

$$\Sigma_I := \int_0^\infty P_I e^{sA} B e^{sA'} P_I' e^{-\lambda_1 s} ds, \tag{23}$$

where we recall that $e^{tA} = \sum_{j=0}^{\infty} t^j A^j / j!$. (It follows from [24], see also [27], that the matrix-valued integral Σ_I in (23) is absolutely convergent.)

It is proved in [24] that, under assumptions (A1)–(A7), $\mathcal{X}_n$ is asymptotically normal if $\operatorname{Re}\lambda \leqslant \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$; more precisely, if $\operatorname{Re}\lambda < \lambda_1/2$ for each such λ, then $n^{-1/2}(\mathcal{X}_n - n\mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\Sigma)$ for some $\mu = (\mu_1,\dots,\mu_k)$ and $\Sigma = (\sigma_{i,j})_{i,j=1}^k$. (If $\operatorname{Re}\lambda = \lambda_1/2$ for some eigenvalue λ, then $\mathcal{X}_n$ is still asymptotically normal, however with another normalisation.) The asymptotic covariance matrix Σ may be calculated in different ways; we refer to [24, Theorem 3.22] for a general formula, but we will instead use two simpler formulas that apply under (different) additional assumptions; see further [24, Section 5].

Theorem 4.1 ([24, Theorem 3.22 and Lemmas 5.4 and 5.3(i)])**.** *Assume* (A1)–(A7) *and that we have normalised as in* (20)*. Also assume that* $\operatorname{Re}\lambda < \lambda_1/2$*, for each eigenvalue* $\lambda \neq \lambda_1$*.*

(i) *Then, as* $n \to \infty$*,*

$$n^{-1/2}(\mathcal{X}_n - n\mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\Sigma), \tag{24}$$

with $\mu = \lambda_1 v_1$ *and some covariance matrix* Σ*.*

(ii) *Suppose further that, for some* $c > 0$*,*

$$a \cdot \mathbb{E}(\xi_i) = c, \qquad i = 1,\dots,q. \tag{25}$$

Then the covariance matrix $\Sigma = c\Sigma_I$*, with* Σ_I *as in* (23)*.*

(iii) *Suppose that* (25) *holds and that the matrix* A *is diagonalisable, and let* $\{u_i'\}_{i=1}^q$ *and* $\{v_i\}_{i=1}^q$ *are dual bases of left and right eigenvectors, respectively, i.e.,* $u_i'A = \lambda_i u_i'$*,* $Av_i = \lambda_i v_i$ *and* $u_i \cdot v_j = \delta_{ij}$*. Then, the covariance matrix in* (i) *is given by, with the matrix* B *as in* (22)*,*

$$\Sigma = c\sum_{j,k=2}^{q} \frac{u_j' B u_k}{\lambda_1 - \lambda_j - \lambda_k} v_j v_k'. \tag{26}$$

□

Remark 4.2. It is easily seen that (25) implies that $\lambda_1 = c$ and $u_1 = a$, see e.g. [24, Lemma 5.4].

Remark 4.3. From (24) follows immediately a weak law of large numbers:

$$\mathcal{X}_n/n \xrightarrow{\mathrm{p}} \mu. \tag{27}$$

In fact, the corresponding strong law $\mathcal{X}_n/n \xrightarrow{\text{a.s.}} \mu$ holds as well, see [24, Theorem 3.21]. It follows that corresponding strong law of large numbers holds for all theorems in Section 3.

Furthermore, in all applications in the present paper, all ξ_{ij} are bounded and thus each $X_{n,i} \leqslant Cn$ for some deterministic constant; hence (27) implies by dominated convergence that also the means converge:

$$\mathbb{E}\,\mathcal{X}_n/n \to \mu. \tag{28}$$

(In fact, this holds in general, without assuming that ξ_{ij} are bounded, since it is easy to see that (A2) implies that $X_{n,i}/n$ are uniformly integrable, which together with (27) yields (28), see e.g. [19, Theorem 5.5.4].)

Moreover, in all applications in the present paper, $a \cdot \xi_i = c$ for some c and every i (a stronger version of (25)), and then (28) can be improved, with an explicit rate of convergence, see [27].

Remark 4.4. It has been noted several times that the Pólya urn process is also well-defined for *real-valued* X_{ni} and ξ_{ij}, see e.g. [24, Remark 4.2], [26, Remark 1.11] and [40] (cf. also [29] for the related case of branching processes); the "number of balls" X_{ni} may thus be any non-negative real number. (This can be interpreted as an urn containing a certain amount (mass) of each type, rather than discrete balls.) In this version, Condition (A1) is replaced by the more general:

(A1′) For each i, either

- there is a real number $d_i > 0$ such that $X_{0,i}$ and $\xi_{1i}, \dots, \xi_{qi}$ are multiples of d_i and $\xi_{ii} \geqslant -d_i$

  or

- $\xi_{ii} \geqslant 0$.

Moreover, $\xi_{ij} \geqslant 0$ when $i \neq j$.

(Note that (A1), with all variables integer-valued, is the case $d_i = 1$ for every i.) Theorem 4.1 holds with (A1) replaced by (A1′), see [24, Remark 4.2]. (The extra assumptions used there are easy to verify when (A1′) holds together with (A2)–(A7).)

In the Pólya urns used in this paper, it is immediately verified that (A1), or at least (A1′), holds, and also (A2). Furthermore, it is easily seen (from the definitions using trees) that every type with positive activity is dominating. If we remove rows and columns corresponding to the types with activity 0 from A, then the removed columns are identically 0, so the set of non-zero eigenvalues of A is not changed. The remaining matrix is irreducible, and using the Perron–Frobenius theorem, it is easy to verify all conditions (A3)–(A6), see [24, Lemma 2.1]. Furthermore, in our urns there will always be a ball of positive activity, so essential extinction is impossible and (A7) holds. Hence, Theorem 4.1 applies.

5 Pólya urns to count fringe subtrees in random m-ary search trees

In this section we describe the Pólya urns that we will use in the analysis of fringe subtrees to prove Theorem 3.2 and Theorem 3.5 for m-ary search trees. The definitions apply to all interpretations of the trees (ordered/unordered, labeled/unlabeled), see Remark 3.1.

Let $T^1, \dots, T^d$ be a fixed sequence of (non-random) m-ary search trees and let $\mathbf{Y}_n = (X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d})$, where $X_n^{T^i}$ is the number of fringe subtrees in $\mathcal{T}_n$ that are isomorphic

to T^i. We may assume that at least one tree T^i contains at least $m-2$ keys. (Otherwise we simply add one such tree to the sequence.)

We define a partial order on the set of (isomorphism classes of) non-random m-ary search trees, such that $T \preceq T'$ if T' can be obtained from T by adding keys (including the case $T' = T$). Of course, the order depends on the definition of isomorphism (ordered, unordered, labelled) one considers.

Assume that we have a given m-ary search tree $\mathcal{T}_n$ together with its external nodes. Denote the fringe subtree of $\mathcal{T}_n$ rooted at a node v by $\mathcal{T}_n(v)$. We say that a node v is *living* if $\mathcal{T}_n(v) \preceq T^i$ for some $i \in \{1, \dots, d\}$, i.e., if $\mathcal{T}_n(v)$ is isomorphic to some T^i or can be grown to become one of them by adding more keys. Note that this includes all external nodes and all leaves with at most $m-2$ keys (by the assumption above). Furthermore, we let all descendants of a living node be living. All other nodes are *dead*.

Now erase all edges from dead nodes to their children. This yields a forest of small trees, where each tree either consists of a single dead node or is living (meaning that all nodes are living) and can be grown to become one of the T^i. We regard these small trees as the balls in our generalised Pólya urn. Hence, the types in this Pólya urn are all (isomorphism types of) non-random m-ary search trees T such that $T \preceq T^i$ for some $i \in \{1, \dots, d\}$, plus one dead type. We denote the set of living types by

$$\mathcal{S} := \bigcup_{i=1}^{d} \{T : T \preceq T^i\}, \tag{29}$$

and the set of all types by $\mathcal{S}^* := \mathcal{S} \cup \{*\}$, where $*$ is the dead type. (The set $\mathcal{S}$ is thus a down-set for the partial order $\preceq$. Conversely, any finite down-set occurs as $\mathcal{S}$, provided it contains all trees with a single node and thus $\leqslant m-2$ keys; we may simply let $T^1, \dots, T^d$ be the trees in $\mathcal{S}$.)

When a key is added to the tree $\mathcal{T}_n$, it is added to a leaf with at most $m-2$ keys or to an external node, and thus to one of the living subtrees in the forest just described. If the root of that subtree still is living after the addition, then that subtree becomes a living subtree of a different type; if the root becomes dead, then the subtree is further decomposed into one or several dead nodes and several (at least m) living subtrees. In any case, the transformation does not depend on anything outside the subtree where the key is added. The random evolution of the forest obtained by decomposing $\mathcal{T}_n$ is thus described by a Pólya urn with the types $\mathcal{S}^*$, where each type has activity equal to its number of gaps, and certain transition rules that in general are random, since the way a subtree is decomposed (or perhaps not decomposed) typically depends on where the new key is added.

Note that dead balls have activity 0; hence we can ignore them and consider only the living types (i.e., the types in $\mathcal{S}$) and we will still have a Pólya urn. The number of dead balls can be recovered from the numbers of balls of other types if it is desired, since the total number of keys is non-random and each dead ball contains $m-1$ keys.

Let $X_{n,T}$ be the number of balls of type T in the Pólya urn, for $T \in \mathcal{S}$. The trees T^i that we want to count correspond to different types in the Pólya urn, but they may also

appear as subtrees of larger living trees. Hence, if $n(T, T')$ denotes the number of fringe subtrees in T that are isomorphic to T', then $X_n^{T_i}$ is the linear combination

$$X_n^{T^i} = \sum_{T \in \mathcal{S}} n(T, T^i) X_{n,T}. \tag{30}$$

The strategy to prove Theorem 3.2 should now be obvious. We verify that the Pólya urn satisfies the conditions of Theorem 4.1 (this is done in Section 6); then that theorem yields asymptotic normality of the vectors $(X_{n,T})_{T \in \mathcal{S}}$, and then asymptotic normality of $(X_n^{T^1}, \dots, X_n^{T^d})$ follows from (30).

Example 5.1 (a Pólya urn to count fringe subtrees with $\boldsymbol{k}$ keys)**.** As an important example, let us consider the problem of finding the distribution of the number of fringe subtrees with a given number of keys, as in Theorem 3.5. In this case, the order of children in the tree does not matter so it is easier to regard the trees as unordered.

So, fix $k \geqslant m-2$ and let T^i, $i \in \{1, \dots, d\}$, be the sequence of all m-ary search trees with at most k keys. This is a down-set, so (29) simply yields $\mathcal{S} = \{T^i : 1 \leqslant i \leqslant d\}$. We ignore the dead nodes and consider the urn with only the living types $\mathcal{S}$.

In the decomposition of an m-ary search tree constructed above, a node v is living if and only if the fringe subtree rooted at v has at most k keys. Hence the decomposition consists of all maximal fringe subtrees with at most k keys, plus dead nodes.

The replacement rules in the Pólya urn are easy to describe. A type T with j keys has $j+1$ gaps, and thus has activity $j+1$. Suppose we draw a ball of type T and $T_1, \dots, T_{j+1}$ are the trees that can be obtained by adding a key to one of these gaps in T. (Some of T_i's may be equal.) If $j < k$, then each T_i has at most k keys and is itself a type in the urn, so the drawn ball is replaced by one ball of a type chosen uniformly at random among $T_1, \dots, T_{j+1}$. On the other hand, if $j = k$, then each T_i has $k+1$ keys and thus has a dead root; the root contains $m-1$ keys, so after removing it we are left with m subtrees that together contain $k+1-(m-1) \leqslant k$ keys; hence these subtrees are all living and the decomposition stops there. Consequently, when $j = k$, the drawn ball is replaced by m balls of the types obtained by choosing one of $T_1, \dots, T_{k+1}$ uniformly at random and then removing its root.

To find the number of fringe subtrees with k keys, we sum the numbers $X_{n,T}$ of balls of type T in the urn, for all types T with exactly k keys. Note that similarly, using (30), we may obtain the number of fringe subtrees with ℓ keys, for any $\ell \leqslant k$, from the same urn. This enables us to obtain joint convergence in Theorem 3.5 for several different k, with asymptotic covariances that can be computed from this urn.

Note that for $k = m-2$, the urn described here consists of $m-1$ types, viz. a single node with $i-1$ keys for $i = 1, \dots, m-1$. This urn has earlier been used in [33, 24, 34] to study the number of nodes, and the numbers of nodes with different numbers of keys, in an m-ary search tree.

In Section 11.1 we give an example with $m = 3$ and $k = 4$; in that case there are 6 different (living) types in the Pólya urn.

Remark 5.2. The types described by the Pólya urns above all have activities equal to the total number of gaps in the type. Since the total number of gaps increases by 1 in each step, we have $a \cdot \xi_i = 1$ for every i, deterministically; in particular, (25) holds with $c = 1$. Hence, $\lambda_1 = 1$ and $u = a$ by Remark 4.2.

Remark 5.3. In the Pólya urns above, a type that is a tree T with g gaps ($g - 1$ keys) has activity g, and if a ball of that type is drawn, each gap is chosen with probability $1/g$ for the addition of a new key. Each gap in T thus gives a contribution with weight $g/g = 1$ to the corresponding column in the intensity matrix A in (19).

6 Proofs of Theorem 3.2 and Theorem 3.5

As said in Section 4, it is easy to see (with the help of [24, Lemma 2.1], for example) that the Pólya urns constructed in Section 5 satisfy (A1)–(A7). To apply Theorem 4.1 it remains to show that $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$. We will find the eigenvalues of A by using induction on the size of $\mathcal{S}$, the set of (living) types. For definiteness we consider the version with ordered unlabelled trees; the versions with unordered trees or labelled trees are the same up to minor differences that are left to the reader.

Note that there is exactly one type that has activity j for every $j \in \{1, \dots, m-1\}$. (These correspond to the nodes holding $j - 1$ keys.) These types are the $m - 1$ smallest in the partial order $\preceq$, and they always belong to the set $\mathcal{S}$ constructed in Section 5.

Let $q := |\mathcal{S}|$ be the number of types in $\mathcal{S}$, and choose a numbering $T_1, \dots, T_q$ of these q types that is compatible with the partial order $\preceq$. For $k \leqslant q$, let

$$\mathcal{S}_k := \{T_1, \dots, T_k\}, \tag{31}$$

and note that this is a down-set for $\preceq$. For $k \geqslant m - 1$, we may thus consider the Pólya urn with the k types in $\mathcal{S}_k$ constructed as in Section 5. Note that this corresponds to decomposing $\mathcal{T}_n$ into a forest with all components in $\mathcal{S}_k \cup \{*\}$. Furthermore, let $\mathcal{X}_n^k := (X_{n,1}^k, \dots, X_{n,k}^k)$, where $X_{n,i}^k$ is the number of balls of type T_i in the urn with types $\mathcal{S}_k$ at time n and let A_k be the intensity matrix of this Pólya urn. Thus $A = A_q$.

First let us take a look at the diagonal values ξ_{ii}.

Proposition 6.1. (i) *Let $m \geqslant 3$ and $m - 1 \leqslant k \leqslant q$. Then $(A_k)_{ii} = -a_i$ for every type $i = 1, \dots, k$. Hence the trace satisfies*

$$\operatorname{tr}(A_k) = -\sum_{i=1}^{k} a_i. \tag{32}$$

(ii) *Let $m = 2$ and $1 \leqslant k \leqslant q$. Then $(A_k)_{ii} = -a_i$ for every type $i = 1, \dots, k$, except for the cases when T_i is the longest left path in $\mathcal{S}_k$ or the longest right path in $\mathcal{S}_k$. If $k \geqslant 3$ these two exceptional cases are distinct, and $(A_k)_{ii} = -a_i + 1$ for them; if $k = 1$ or $k = 2$, then the exceptional cases coincide and $(A_k)_{ii} = -a_i + 2$ for the single exceptional case. Consequently, for any $k \geqslant 1$, the trace satisfies*

$$\operatorname{tr}(A_k) = 2 - \sum_{i=1}^{k} a_i. \tag{33}$$

Proof. Observe that if we draw a ball of type i with k_i keys, then the ball is replaced either by a single ball of a type with $k_i + 1$ keys or by several different balls obtained by decomposing a tree with $k_i + 1$ keys that has a dead root. In the latter case, $m - 1$ of the keys are in the dead root, so each living tree in the decomposition has at most $k_i + 1 - (m - 1) = k_i - m + 2$ keys. Hence, if $m \geqslant 3$, then in no case will there be a ball with exactly k_i keys among the added balls, and in particular no ball of type i; consequently, $\xi_{ii} = -1$ and $(A_k)_{ii} = -a_i$, see (19).

When $m = 2$, the same holds except if T_i is such that it is possible to add a new key such that the root dies and the tree decomposes into the dead root, a copy of T_i and an external node. This can happen only if the root has at most one child, and it follows by induction that every node has at most one child, so T is a path, and, furthermore, that T_i must be either a left path or a right path, with the new key added at the end; furthermore, it must be the longest such path in $\mathcal{S}_k$, since otherwise the root would not die. If $k \geqslant 3$, then T_3 is a path with two nodes, and it follows that the two exceptional cases are distinct. If T_i is one of them, then T_i has a_{ii} gaps and only one of them will yield a new copy of T_i if the new key is added there. Hence, $\mathbb{E}\,\xi_{ii} = -(a_{ii} - 1)/a_{ii}$, and (19) yields $(A_k)_{ii} = -(a_{ii} - 1)$. The cases $k = 1$ or 2 are similar, but they are so simple that they are simplest treated separately; the matrices A_k are $\bigl(1\bigr)$ and $\left(\begin{smallmatrix} -1 & 2 \\ 1 & 0 \end{smallmatrix}\right)$, with $a_1 = 1$ and $a_2 = 2$. □

Theorem 6.2. *Let $m \geqslant 2$. The eigenvalues of A are the $m - 1$ roots of the polynomial $\phi_m(\lambda) := \prod_{i=1}^{m-1}(\lambda + i) - m!$ plus the multiset*

$$\{-a_i : i = m, m + 1, \dots, q\}\,. \tag{34}$$

Proof. We prove by induction on k that the theorem holds for A_k (with q replaced by k in (34)), for any k with $m - 1 \leqslant k \leqslant q$. The theorem is the case $k = q$.

First, for the initial case $k = m - 1$, T_i is a single node with $i - 1$ keys, $i = 1, \dots, k$; thus $X_{n,i}^{m-1}$ is the number of nodes with $i - 1$ keys, i.e., the number of nodes with i gaps. (In particular, $X_{n,1}^{m-1}$ is the number of external nodes.) This Pólya urn with $m - 1$ types has earlier been analyzed, see e.g., [24, Example 7.8] and [34, Section 8.1.3]. The $(m-1) \times (m-1)$ matrix A_{m-1} has elements $a_{i,i} = -i$ for $i \in \{1, \dots, m-1\}$, $a_{i,i-1} = i - 1$ for $i \in \{2, \dots, m\}$, $a_{1,m-1} = m \cdot (m - 1)$ and all other elements $a_{i,j} = 0$, i.e.,

$$A_{m-1} = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 & m(m-1) \\ 1 & -2 & 0 & \dots & 0 & 0 \\ 0 & 2 & -3 & \dots & 0 & 0 \\ 0 & 0 & 3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & m-2 & -(m-1) \end{pmatrix}. \tag{35}$$

As is well-known, the matrix A_{m-1} has characteristic polynomial $\phi_m(\lambda)$; this shows the theorem for $k = m - 1$, since the set (34) is empty in this case.

We proceed to the induction step. Let $m - 1 \leqslant k < q$. By using arguments similar to those that were used in the proof of [22, Lemma 5.1] we will show that A_{k+1} inherits

(with multiplicities) the eigenvalues of A_k. We write $a^k = (a_1, \dots, a_k)$ for the activity vector of the Pólya urn with types in $\mathcal{S}_k$.

We have $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{T_{k+1}\}$. The vector $\mathcal{X}_n^{k+1}$ determines also the number of subtrees of each type in the decomposition of $\mathcal{T}_n$ into the types in $\mathcal{S}_k$, and there is an obvious linear map $T : \mathbb{R}^{k+1} \to \mathbb{R}^k$ such that $\mathcal{X}_n^k = T\mathcal{X}_n^{k+1}$. Furthermore, starting the urns with an arbitrary (deterministic) non-zero vector $\mathcal{X}_0^{k+1} \in \mathbb{Z}_{\geqslant 0}^{k+1}$ and $\mathcal{X}_0^k = T\mathcal{X}_0^{k+1}$, the urn dynamics yield

$$\mathbb{E}(\mathcal{X}_1^{k+1} - \mathcal{X}_0^{k+1}) = \frac{A_{k+1}\mathcal{X}_0^{k+1}}{a^{k+1} \cdot \mathcal{X}_0^{k+1}}, \tag{36}$$

$$\mathbb{E}(\mathcal{X}_1^{k} - \mathcal{X}_0^{k}) = \frac{A_{k}\mathcal{X}_0^{k}}{a^{k} \cdot \mathcal{X}_0^{k}}. \tag{37}$$

Consequently, since also $a^{k+1} \cdot \mathcal{X}_0^{k+1} = a^k \cdot \mathcal{X}_0^k$ (this is the total activity, i.e., the total number of gaps),

$$\begin{aligned} TA_{k+1}\mathcal{X}_0^{k+1} &= (a^{k+1} \cdot \mathcal{X}_0^{k+1}) T\,\mathbb{E}(\mathcal{X}_1^{k+1} - \mathcal{X}_0^{k+1}) = (a^k \cdot \mathcal{X}_0^k)\,\mathbb{E}(\mathcal{X}_1^k - \mathcal{X}_0^k) = A_k\mathcal{X}_0^k \\ &= A_k T\mathcal{X}_0^{k+1}, \end{aligned}$$

and thus, since $\mathcal{X}_0^{k+1}$ is arbitrary,

$$TA_{k+1} = A_k T. \tag{38}$$

Let u' be a left generalised eigenvector of rank m corresponding to the eigenvalue λ of the matrix A_k, i.e.,

$$u'(A_k - \lambda I_k)^m = 0.$$

Then, by (38),

$$u'T(A_{k+1} - \lambda I_{k+1})^m = u'(A_k - \lambda I_k)^m T = 0,$$

and thus $u'T = (T'u)'$ is a left generalised eigenvector of A_{k+1} for the eigenvalue λ. Since T is onto (it maps $(x_1, \dots, x_k, 0)$ to $(x_1, \dots, x_k)$), T' is injective and thus T' is an injective map of the generalised eigenspace (for λ) of A_k into the generalised eigenspace of A_{k+1}. This shows that λ is an eigenvalue of A_{k+1} with algebraic multiplicity at least as large as for A_k. Consequently, if A_k has eigenvalues $\lambda_1, \dots, \lambda_k$ (including repetitions, if any), then A_{k+1} has eigenvalues $\lambda_1, \dots, \lambda_k, \lambda_{k+1}$ for some complex number λ_{k+1}.

Then the result follows by the following observation. The trace of a matrix is equal to the sum of the eigenvalues; hence,

$$\operatorname{tr} A_{k+1} = \lambda_1 + \cdots + \lambda_{k+1} = \operatorname{tr} A_k + \lambda_{k+1} \tag{39}$$

and thus by (32) (when $m > 2$) or (33) (when $m = 2$),

$$\lambda_{k+1} = \operatorname{tr}(A_{k+1}) - \operatorname{tr}(A_k) = -a_{k+1}. \tag{40}$$

Thus, by induction, Theorem 6.2 holds for every A_k, with $m-1 \leqslant k \leqslant q$, and in particular for $A = A_q$. □

Theorem 6.2 shows that the eigenvalues of A are the roots of ϕ_m plus some negative numbers $-a_i$; hence the condition $\operatorname{Re}\lambda < \lambda_1/2$ in Theorem 4.1 is satisfied for all eigenvalues of A except λ_1 if the condition is satisfied for the roots of ϕ_m except λ_1. Furthermore, $\lambda_1 = 1$ by Remark 5.2. Let

$$\gamma_m := \max_{\phi_m(\lambda)=0:\lambda\neq\lambda_1} \operatorname{Re}\lambda, \tag{41}$$

i.e., the largest real part of a root of ϕ_m except λ_1. (If $m = 2$, when there is no other root, we interpret this as $\gamma_2 = -\infty$.) Thus our condition in Theorem 4.1 is satisfied if and only if $\gamma_m < \frac{1}{2}$; it is well-known that this holds if $m \leqslant 26$, but not for larger m, see [35] and [14].

In the remainder of this section we assume $m \leqslant 26$. Thus

$$\operatorname{Re}\lambda \leqslant \max(\gamma_m, 0) < \frac{1}{2} = \frac{\lambda_1}{2} \tag{42}$$

for every eigenvalue $\lambda \neq \lambda_1$, and Theorem 4.1 applies to the urn defined above.

Proof of Theorem 3.2. By Theorem 4.1(i), (24) holds, with $\mu = \lambda_1 v_1 = v_1$.

By (30), $\mathbf{Y}_n = \bigl(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\bigr) = R\mathcal{X}_n$ for some (explicit) linear operator R. Hence, (24) implies

$$n^{-1/2}\bigl(\mathbf{Y}_n - nR\mu\bigr) = R\bigl(n^{-1/2}(\mathcal{X}_n - \mu)\bigr) \xrightarrow{\mathrm{d}} \mathcal{N}\bigl(0, R\Sigma R'\bigr). \tag{43}$$

Furthermore, as said above, $\operatorname{Re}\lambda \leqslant \gamma_m$ for every eigenvalue $\lambda \neq \lambda_1$. We note also that if $\lambda \neq \lambda_1$ is an eigenvalue with $\operatorname{Re}\lambda = \max(\gamma_m, 0)$, then λ is not in (34) so λ is a root of ϕ_m; furthermore, all roots of ϕ_m are simple [35], and therefore λ is a simple eigenvalue. Hence, by [27], cf. [35, Theorem 1] for a special case proved by other methods,

$$\mathbb{E}\,\mathcal{X}_n = n\mu + O\bigl(n^{\max(\gamma_m,0)}\bigr), \tag{44}$$

and thus, since $\gamma_m < \frac{1}{2}$ for $m \leqslant 26$ as said above,

$$\mathbb{E}\,\mathcal{X}_n = n\mu + o\bigl(n^{1/2}\bigr). \tag{45}$$

Hence,

$$\boldsymbol{\mu}_n = \mathbb{E}\,\mathbf{Y}_n = R\bigl(\mathbb{E}\,\mathcal{X}_n\bigr) = nR\mu + o\bigl(n^{1/2}\bigr). \tag{46}$$

Consequently, (43) implies (2), with the covariance matrix $R\Sigma R'$, where Σ is as in (24).

Moreover, as said in Remark 3.3, it follows from [23], to be precise by combining [23, (5.30), Theorem 7.10 and Theorem 7.11], that (for any $m \geqslant 2$)

$$\mathbb{E}\,\mathbf{Y}_n = n\hat{\boldsymbol{\mu}} + o(n). \tag{47}$$

By combining (46) and (47) we see that $R\mu = \hat{\boldsymbol{\mu}}$ (since neither depends on n), and thus (46) yields (4).

To see that the covariance matrix $R\Sigma R'$ is non-singular when each T^i has an internal node so $k_i > 0$, suppose that, on the contrary, $u'R\Sigma R'u = 0$ for some vector $u \neq 0$. Then,

by [27, Theorem 3.6], $u'\mathbf{Y}_n = u'R\mathcal{X}_n$ is deterministic for every n. We argue as for the case $k = 2$ in the proof of [21, Lemma 3.6]. We may assume that every $u_i \neq 0$, since we may just ignore each T^i with $u_i = 0$; we may also assume that $1 \leqslant k_1 \leqslant k_2 \leqslant \dots$. Let N be a large integer, with $N > k_d$, and let T_1 be a tree consisting of a single path with $N + k_1$ internal nodes, each of them (except the root) the right-most child of the preceding one. Let T_2 consist of a similar right-most path with N internal nodes, together with $m - 1$ copies of T_1, which have their roots as the $m-1$ first children of T_2. Note that both T_1 and T_2 have $(N + k_1)(m - 1)$ keys, so they are possible realizations of $\mathcal{T}_{(N+k_1)(m-1)}$. Moreover, for any tree T^i, $i \geqslant 2$, T_1 and T_2 have the same number of fringe trees isomorphic to T^i, while T_1 contains $m - 1$ more copies of T^1 than T_2 does. Hence the linear combination $u'\mathbf{Y}_n = \sum_i u_i X_n^{T^i}$ may take at least two different values when $n = (N + k_1)(m - 1)$, which is a contradiction. Consequently, the covariance matrix cannot be singular when all $k_i > 0$. □

Proof of Theorem 3.5. Let $T^1, \dots, T^d$ be all non-random m-ary search trees with k keys. (We may consider either ordered or unordered trees; for numerical calculations, the unordered case is simpler.) Then $Y_{n,k} = \sum_{i=1}^d X_n^{T^i}$. The result (7) thus follows from (2), and $\sigma_k^2 > 0$ when $k > 0$ because the covariance matrix Σ in (2) then is non-singular. Also in the case $k = 0$ and $m > 2$ we have $\sigma_k^2 > 0$, because $Y_{n,0} = n + 1 - \sum_{j=1}^{m-2}(j + 1)Y_{n,j}$ by (6), and the asymptotic variance of the right-hand side is non-zero by an application of Theorem 3.2 to the sequence of all m-ary search trees with at least 1 and at most $m - 2$ keys.

Furthermore, summing (4) over the trees T^i yields

$$\mathbb{E}\big(Y_{n,k}\big) = \frac{n}{(H_m - 1)(k + 1)(k + 2)} + o\big(n^{1/2}\big). \tag{48}$$

and thus (8) follows. □

Remark 6.3. The results on moment convergence in Janson and Pouyanne [28] are stated for a Pólya urn with deterministic replacements, but as is mentioned [28, Remark 1.9] the results hold also for random replacement vectors (and by [28, Remark 1.7], different types may have different activities), which yields convergence of all moments in (2), as said in Remark 3.3. As a consequence, (7) and (8) also hold with convergence of all moments. Alternatively, in the present case we can instead consider an urn where the balls represent gaps in the trees in the construction in Section 5. This yields an urn with deterministic replacements, although now also some subtractions of balls occur. The results in [28] apply to this version too by [28, Remark 1.8], which again yields moment convergence.

7 Pólya urns to count fringe subtrees in preferential attachment trees

We now describe the Pólya urns that we use for proving Theorem 3.9 for linear preferential attachment trees. We may consider either ordered or unordered trees, see Remark 2.6.

7.1 A Pólya urn with infinitely many types for the general case

Let $\Lambda^1, \dots, \Lambda^d$ be a fixed sequence of rooted trees and let $\mathbf{Z}_n = (X_n^{\Lambda^1}, X_n^{\Lambda^2}, \dots, X_n^{\Lambda^d})$, where $X_n^{\Lambda^i}$ is the number of fringe subtrees that are isomorphic to Λ^i in Λ_n.

Assume that we have a given preferential attachment tree Λ_n. As in Section 5, we say that a node v is *living* if the fringe subtree $\Lambda_n(v) \preceq \Lambda^i$ for some $i \in \{1, \dots, d\}$. Furthermore, we let all descendants of a living node be living. All other nodes we declare *dead*.

Now erase all edges from dead nodes to their children. This yields a forest of small trees, where each tree either consists of a single dead node or is living and can be grown to become one of the Λ^i. Again, we regard these small trees as the balls in our generalised Pólya urn. However, unlike the situation in Section 5, we now cannot ignore the dead nodes, since they may get new children; furthermore, the probability of this depends on their degree. Hence we label each dead node by the number of children it has in Λ_n.

Hence, the types in this Pólya urn are all (isomorphism types of) rooted trees Λ such that $\Lambda \preceq \Lambda^i$ for some $i \in \{1, \dots, d\}$ (these are called *ordinary types*), plus one type $*_k$ for each positive integer k, consisting of a single dead node labelled by k (these are called *special types*). In other words, the set of types is $\mathcal{S}' \cup \mathcal{S}''$, where

$$\mathcal{S}' := \bigcup_{i=1}^{d} \{\Lambda : \Lambda \preceq \Lambda^i\} \tag{49}$$

is the set of ordinary types (cf. (29)) and $\mathcal{S}'' := \{*_k : k \geqslant 1\}$ is the set of special types. (As in the corresponding case in Section 5, the set $\mathcal{S}'$ is thus a down-set for the partial order $\preceq$. Conversely, any finite down-set occurs as $\mathcal{S}'$, for example, by choosing $\Lambda^1, \dots, \Lambda^d$ to be the trees in $\mathcal{S}'$.)

Unfortunately, $\mathcal{S}''$ is infinite, so this is a Pólya urn with infinitely many types. Theorem 4.1 does not apply to such urns, and we do not know any extension to infinite-type urns that can be used here. However, in the linear case (1), we can reduce the urn to a finite-type one; this is done in the following subsection.

Nevertheless, it is easy to describe the replacement rules for this urn in general. The activity a_Λ of an ordinary type Λ is the total weight $w_\Lambda := \sum_{v \in \Lambda} w_{d^+(v)}$ of all nodes in Λ, while a special type $*_k$ has activity w_k. If a ball of an ordinary type Λ is drawn, we add a new child to one of its nodes, with probabilities determined by the weights w_k and the out-degrees as in the definition of Λ_n, and if the resulting tree is dead (does not belong to $\mathcal{S}'$), it is decomposed into several trees (including at least one dead node). If a ball of a special type $*_k$ is drawn, it is replaced by one ball of type $*_{k+1}$ and one ball of type $\bullet$, the tree with a single node. (The tree $\bullet$ is always an ordinary type.)

7.2 A Pólya urn with finitely many types for the linear case

Consider from now on only the linear case (1). We then can replace the infinite-type Pólya urn in Section 7.1 by a Pólya urn with finitely many types by using a version of the trick used in [25, 24] to study node degrees in random recursive trees and plane oriented recursive trees.

Recall that we may assume that $\chi \in \{-1, 0, 1\}$. For simplicity, consider first the case when ρ (and thus every w_k) is an integer. Change each ball of type $*_k$ to w_k balls of a new type $*$. Let $*$ have activity 1; then the activities are preserved by the change. Moreover, if a ball of type $*_k$ is drawn, it is, as said above, replaced by one ball of type $*_{k+1}$ and one of type $\bullet$; after the change, this means that the number of balls $*$ is increased by $w_{k+1} - w_k = \chi$. (This is where the linearity of the weights is essential.) If $\chi = -1$, this means that the ball is not replaced. Consequently, the new urn also evolves as a Pólya urn with types $\mathcal{S}^* := \mathcal{S}' \cup \{*\}$. An ordinary type Λ has the activity $\sum_{v\in\Lambda} w_{d^+(v)}$ as above in Section 7.1, and when drawn, it is replaced as above, except that instead of any ball of a type $*_k$ we add w_k balls of type $*$. The special type $*$ has activity 1, and when drawn, it is replaced and we add χ additional balls of type $*$ and one ball of type $\bullet$. (For an example, see Section 11.3.)

Moreover, we can do the same for general ρ. As said in Remark 4.4, the Pólya urn process is well-defined for *real-valued* X_{ni} and ξ_{ij}, interpreting the "number of balls" as the mass in the urn of each type.

With this interpretation, the Pólya urn with types $\mathcal{S}^*$ just described exists and describes the evolution of fringe subtrees for any linear preferential attachment model, also with non-integer ρ. (We can also allow $\chi \notin \{-1, 0, 1\}$, but as said earlier, this is not more general.)

Note that the extension to real-valued urns is needed only when $\chi = 1$; when $\chi = 0$ we may assume that $\rho = 1$ and when $\chi = -1$ we necessarily have $\rho \in \mathbb{Z}_+$, see Section 2.2; hence the integer version is enough in these cases. Note further that then (A1′) holds, with $d_i = 1$ for every ordinary type and $\xi_{**} \geqslant 0$. Hence, Theorem 4.1 holds also for the real-valued urns considered here; see Remark 4.4.

Remark 7.1. In the linear case, the total activity increases by $\chi + \rho = w_1$ each time a ball is drawn; the easiest way to see this is by going back to the preferential attachment tree Λ_n and noting that: the total activity in the urn equals the total weight of the tree; when we add a new node, it contributes extra weight $w_0 = \rho$ and the weight of its parent increases by χ. In other words, we have $a \cdot \xi_i = \chi + \rho$ deterministically, and thus (25) holds with $c = \chi + \rho$. Consequently, $\lambda_1 = \chi + \rho$, see Remark 4.2.

As a consequence, we also see that the total activity of the urn before the nth step, i.e., the total weight of the preferential attachment tree Λ_n with n nodes, is

$$w_{\Lambda_n} = \rho + (n-1)(\chi + \rho) = n(\chi + \rho) - \chi, \tag{50}$$

see e.g., [23, (6.22)]; this is deterministic, and any tree with n nodes has the same total weight.

8 Proofs of Theorem 3.9 and Theorem 3.11

To apply Theorem 4.1 in the proofs of Theorem 3.9 and Theorem 3.11 it remains to show that $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$ of the intensity matrix A for the urn in

Section 7.2. We will again (as in the case of random m-ary search trees in Section 5) find the eigenvalues of A by using induction.

We consider either ordered or unordered trees; the results and proofs below apply to both versions.

Recall that the set of types is $\mathcal{S}^* = \mathcal{S} \cup \{*\}$. Let $q := |\mathcal{S}^*|$ be the number of types, and choose a numbering $T_1, \dots, T_{q-1}$ of the $q-1$ ordinary types that is compatible with the partial order $\preceq$. (Hence, $T_1 = \bullet$.) For $2 \leqslant k \leqslant q$, let

$$\mathcal{S}_k := \{T_1, \dots, T_{k-1}, *\}. \tag{51}$$

Since $\mathcal{S}_k \setminus \{*\} = \{T_1, \dots, T_{k-1}\}$ is a down-set for $\preceq$, we may thus consider the Pólya urn with the k types in $\mathcal{S}_k$, defined as in Section 7.2. Let A_k be the $k \times k$ intensity matrix of this Pólya urn, and let $\mathcal{X}_n^k := (X_{n,1}^k, \dots, X_{n,k}^k)$, where $X_{n,i}^k$ is the number of balls of type T_i in this urn at time n. The activities are $a^k := (a_1, \dots, a_{k-1}, 1)$, where

$$a_i := w_{T_i} = |T_i|(\chi + \rho) - \chi. \tag{52}$$

see (50). (Recall that $*$ always has activity 1.)

Proposition 8.1. *Let $2 \leqslant k \leqslant q$.*

(i) *For every ordinary type T_i, $i = 1, \dots, k-1$, except the type that is a path (rooted at an endpoint) of maximal length,*

$$(A_k)_{ii} = -a_i. \tag{53}$$

(ii) *For the ordinary type T_j that is a path of maximal length among all paths in $\mathcal{S}_k$,*

$$(A_k)_{jj} = \rho - a_j. \tag{54}$$

(iii) *For the special type $*$, we have*

$$(A_k)_{kk} = \chi. \tag{55}$$

Consequently,

$$\operatorname{tr}(A_k) = \chi + \rho - \sum_{j=1}^{k-1} a_j. \tag{56}$$

Proof. This is similar to the proof of Proposition 6.1. Consider first an ordinary type T_i. If we draw T_i, then we add a node to it, which either gives a new living type $T_j \neq T_i$, or a tree with a dead root that is decomposed. In the latter case, the only possibility to get a copy of T_i is that only the root is dead, and that when we remove it, the remaining $|T_i|$ living nodes form a tree isomorphic to T_i, i.e., that we first add a node to T_i and then remove the root, and obtain a copy of T_i. In this exceptional case, the root of T_i has to have degree at most 1, and by induction it follows that every node in T_i has out-degree at most 1, so T_i is a path; furthermore, the new node has to be added at the leaf, and T_i

has to have maximal length (since otherwise the root would not die when the new node is added).

(i): In this case, a drawn ball of type T_i is never replaced by a ball of the same type; thus $\xi_{ii} = -1$ and, by (19), $(A_k)_{ii} = -a_i$, which yields (53).

(ii): If T_j is a maximal path, then when adding a new node to T_j, the probability of adding it at the leaf is w_0/w_{T_j}. In this case, and only then, we obtain a new ball T_j. Consequently,

$$\mathbb{E}\,\xi_{jj} = -1 + w_0/w_{T_j} = -1 + w_0/a_j \tag{57}$$

and, by (19),

$$(A_k)_{jj} = a_j\,\mathbb{E}\,\xi_{jj} = a_j(-1 + w_0/a_j) = -a_j + w_0 = -a_j + \rho, \tag{58}$$

which yields (54).

(iii): The special type $*$ has activity 1, and if we draw a ball of type $*$, then the total change in the urn is χ additional balls of type $*$ and one additional ball of type $\bullet$; hence $(A_k)_{kk} = \mathbb{E}\,\xi_{kk} = \chi$. □

Theorem 8.2. *For the linear preferential attachment trees, the eigenvalues of the intensity matrix A are $\chi + \rho$ and $-a_i$ for $i \in \{1, \dots, q-1\}$, where a_i is given by* (52).

Proof. The proof is similar to the proof of Theorem 6.2. Again we will use induction, and consider the $k \times k$ matrices A_k defined above.

We start the induction with $k = 2$, when the types are $\{\bullet, *\}$. This means that all nodes with children are dead. If we draw a ball of type $\bullet$, we add a node to it and get a tree with two nodes; the root is dead and we obtain a new ball of type $\bullet$ and a dead node of type $*_1$, which is changed to a mass $w_1 = \chi + \rho$ of type $*$. A ball of type $\bullet$ is thus replaced by another ball of type $\bullet$ and a mass $\chi + \rho$ of type $*$, so $\xi_{11} = 1 - 1 = 0$ (in accordance with (54)) and $\xi_{12} = \chi + \rho$. Similarly, $\xi_{21} = 1$ and $\xi_{22} = \chi$. Consequently, by (19), since $a_1 = w_{T_1} = w_0 = \rho$,

$$A_2 = \begin{pmatrix} 0 & 1 \\ \rho(\chi+\rho) & \chi \end{pmatrix}. \tag{59}$$

The matrix A_2 has eigenvalues $\chi + \rho$ and $-\rho = -a_1$, verifying the theorem for A_2.

The induction step is identical to the one in the proof of Theorem 6.2. We see again that the eigenvalues of A_{k+1} are inherited from A_k, i.e., that the eigenvalues of A_{k+1} can be listed (with multiplicities) as $\lambda_1, \dots, \lambda_{k+1}$ where $\lambda_1, \dots, \lambda_k$ are the eigenvalues of A_k. Finally, we use (56) and deduce

$$\lambda_{k+1} = \operatorname{tr}(A_{k+1}) - \operatorname{tr}(A_k) = -a_k, \tag{60}$$

and the theorem follows by induction. □

Remark 8.3. For the linear preferential attachment tree the eigenvalues are easier to describe than for the m-ary search tree. The reason for this is that we can start the

induction already when we have a Pólya urn consisting of only two types. (The brave reader might even start with only one type, regarding all nodes as dead, and $A_1 = (\chi+\rho)$.) In the m-ary search tree, on the other hand, we always have at least $m-1$ different types as explained above, and these first types are the reason why we get more complicated eigenvalues.

Proof of Theorem 3.9. The proof is analogous to the proof of Theorem 3.2. By Theorem 8.2, $\lambda_1 = \chi+\rho$ and all other eigenvalues are negative. Hence, Theorem 4.1(i) applies, so (24) holds, with $\mu = \lambda_1 v_1$. Furthermore, by [27],

$$\mathbb{E}\,\mathcal{X}_n = n\mu + O(1). \tag{61}$$

By an analogue of (30), $\mathbf{Z}_n = \bigl(X_n^{\Lambda^1}, X_n^{\Lambda^2}, \dots, X_n^{\Lambda^d}\bigr) = R\mathcal{X}_n$ for some (explicit) linear operator R. Hence, (24) implies

$$n^{-1/2}\bigl(\mathbf{Z}_n - nR\mu\bigr) = R\bigl(n^{-1/2}(\mathcal{X}_n - \mu)\bigr) \xrightarrow{\mathrm{d}} \mathcal{N}\bigl(0, R\Sigma R'\bigr) \tag{62}$$

and (61) implies

$$\boldsymbol{\mu}_n = \mathbb{E}\,\mathbf{Z}_n = R\bigl(\mathbb{E}\,\mathcal{X}_n\bigr) = nR\mu + O(1), \tag{63}$$

which together yield (12), with the covariance matrix $R\Sigma R'$, where Σ is as in (24).

Moreover, as said in Remark 3.10, it follows from [23, (5.29) and Example 6.4, in particular (6.24)], that

$$\mathbb{E}\,\mathbf{Z}_n = n\hat{\boldsymbol{\mu}} + o(n). \tag{64}$$

By (63) and (64), we have $R\mu = \hat{\boldsymbol{\mu}}$, and thus (63) yields (15).

To see that the covariance matrix $R\Sigma R'$ is non-singular, suppose that, on the contrary, $u'R\Sigma R'u = 0$ for some vector $u \neq 0$. Then, by [27, Theorem 3.6], $u'\mathbf{Z}_n = u'R\mathcal{X}_n$ is deterministic for every n. However, this is impossible by the same construction as in the proof of Theorem 3.2, taking $m = 2$ (so the number of internal nodes equals the number of keys) and then ignoring keys. See also the proofs of [21, Lemmas 3.6 and 3.17], which give this construction in the special case of the random recursive tree, and note that (at least for $\chi \geqslant 0$), the general case is the same since for a given n, the possible trees Λ_n are the same for all $\chi \geqslant 0$ and ρ (although their different probability distributions are different). □

Proof of Theorem 3.11. Let $\Lambda^1, \dots, \Lambda^d$ be all non-random unordered (or ordered) trees with k keys. Then $Y_{n,k} = \sum_{i=1}^d X_n^{\Lambda^i}$. The result (16) thus follows from (12), and $\sigma_k^2 > 0$ because the covariance matrix Σ in (12) is non-singular.

Furthermore, summing (15) over the trees Λ^i yields (18) and thus (17) follows. □

9 A Pólya urn to count protected nodes in m-ary search trees

We start precisely in the same way as in [22]. Given an m-ary search tree $\mathcal{T}_n$ with n keys together with its external nodes, erase all edges that connect two internal non-leaves. This

yields a forest of small trees, where (assuming $n \geqslant m$) each tree has a root that is a non-leaf in $\mathcal{T}_n$. We regard these small trees as the balls in our generalised Pólya urn. The type of a ball (tree) is the type of the tree as an unordered tree, i.e., up to permutations of the children. The type of a tree in the urn is thus described by the numbers $x_i, i = 1, \dots, m$, of children of the root with $i-1$ keys; each of these children is an external node ($i = 1$) or a leaf ($i \geqslant 2$), and it has itself children only when $i = m$ when it has m external children; thus we can unambiguously label the type by a vector $\mathbf{x} = (x_1, \dots, x_m)$. Since the root of any of the small trees has m children (including external ones) in the original tree $\mathcal{T}_n$, we have $\sum_{i=1}^m x_i \leqslant m$, (with the remainder $m - \sum_{i=1}^m x_i$ equal to the number of erased edges to children in the original tree $\mathcal{T}_n$ that are non-leaves). If $n \geqslant m$, there are no balls of type $(m, 0, \dots, 0)$, since the root of a small tree is never a leaf in $\mathcal{T}_n$, and therefore the total number of types is the number of ways to write m as a sum of $m+1$ non-negative integers minus one, i.e., $\binom{2m}{m} - 1$.

The activity of a type $\mathbf{x}$ is the number of gaps it contains. The root has no gaps and each child with $i-1$ keys contributes i gaps. (These gaps belong to the child itself when $i \leqslant m-1$ and to the m external children of it when $i = m$.) Therefore type $\mathbf{x}$ has activity $a_{\mathbf{x}} := \sum_{i=1}^m i x_i$.

Let us denote the unit vectors

$$\mathbf{e}_1 = (1, 0, \dots, 0), \quad \mathbf{e}_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad \mathbf{e}_m = (0, \dots, 0, 1).$$

If we add a new key to a leaf with $i-1 \leqslant m-2$ keys, which belongs to a small tree of type $\mathbf{x}$, in the Pólya urn this corresponds to replacing the ball by a ball of type $\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}$. The same holds if we add a key to an external node that is a child of the root. However, if we add a key to an external node that is a child of a (full) leaf in a tree of type $\mathbf{x}$, then that leaf becomes a non-leaf, so the edge from it to the root is erased and the tree is split into two trees: one of type $(m-1, 1, 0, \dots 0)$ and the other of type $\mathbf{x} - \mathbf{e}_m$. Thus in general there may be up to m different ways a small tree can be transformed, depending on which gap a new key goes into.

10 Proof of Theorem 3.7

The idea of the proof is the same as in the case of fringe subtrees: reducing the urn one type at a time, until we arrive at the simple urn for counting incomplete leaves with matrix A_{m-1}. In addition to the types described in Section 9, add types numbered 1 to m where type i is a single node with $i-1$ keys for $i \leqslant m-1$ and type m is the type $(m, 0, \dots, 0)$, i.e., a node with $m-1$ keys and m external children. Of course, when $n \geqslant m$, there are no small trees of these types, but they are used in the induction.

Recalling that the number of types described above is $\binom{2m}{m} - 1$ and writing $q = m - 1 + \binom{2m}{m}$, let us enumerate all types from 1 to q in such a way that whenever we insert a key into a tree of type i (and obtain one or two trees, as described above), the new tree which inherits the root of the old one must have index larger than i. Note the similarity with the fringe case where the tree grown by inserting a key must have a larger index.

There is more than one such enumeration of types, but for clarity we choose the following one. We let the first $m-1$ types be the single nodes ordered with an increasing number of keys, while the rest of the types are ordered in such a way that a type with more children precedes a type with fewer children; ordering among types with equal number of children is achieved by treating them as numbers in base $m+1$ (with the coordinate x_1 being the most significant digit) arranged in decreasing order. (For types with the same number of children, this is the reverse lexicographic order.) The vector describing a type $i \geqslant m$ is denoted by $\mathbf{x}^{(i)}$. Thus, for example, type m is type $\mathbf{x}^{(m)} = (m, 0, \ldots, 0)$ and is followed by $\mathbf{x}^{(m+1)} = (m-1, 1, 0, \ldots, 0)$, and the last type is the dead type $\mathbf{x}^{(q)} = (0, \ldots, 0)$. Figure 1 shows the different types in a ternary search tree (i.e., $m = 3$), except for the first 3, which are single nodes (and are the same as the first three in Figure 2).

Let us write $\mathcal{X}_n = (X_{n,1}, \ldots, X_{n,q})$, where $X_{n,i}$ is the number of balls of type i in the urn. The limit distribution of the urn is, of course, described by types $m+1$ to q, since for $n \geqslant m$ we have no balls of the first m types.

For each $k = m-1, \ldots, q$ we define a reduced urn by considering the forest of small trees in the original urn and deleting roots of the trees of types $i > k$ (and the edges leading to their children). This replaces each tree of type i by $\mathbf{x}_j^{(i)}$ trees of type j, for each $j = 1, \ldots, m$. Write

$$\mathcal{X}_n^k = (X_{n,1}^k, \ldots, X_{n,k}^k).$$

Clearly, there is a $k \times q$ matrix P_k such that $\mathcal{X}_n^k = P_k \mathcal{X}_n$ and therefore

$$\Delta \mathcal{X}_n^k = \mathcal{X}_{n+1}^k - \mathcal{X}_n^k = P_k \left(\Delta \mathcal{X}_n \right). \tag{65}$$

It need not be obvious that $\mathcal{X}_n^k$ is actually a Pólya urn, in the sense that the distribution of $\Delta \mathcal{X}_n^k$ depends only on the type i of the ball drawn. To convince ourselves we consider three cases.

Case 1: If we draw a ball of type $i \in \{m+1, \ldots, k\}$, then in the original urn this corresponds to drawing a ball of the same type. Therefore, in view of (65), $\Delta \mathcal{X}_n^k$ has the same distribution as $P_k \xi_i$, which depends on i only.

Case 2: If we draw a ball of type $i < m$, then in the original urn (assuming $n \geqslant m$) this corresponds to replacing a tree of some type $r > k$ (for which $x_i^{(r)} > 0$) by a tree of type $\mathbf{x}^{(s)} := \mathbf{x}^{(r)} - \mathbf{e}_i + \mathbf{e}_{i+1}$. The latter tree inherits the root of the former, which, by the ordering of types, implies that $s > r > k$. Therefore in the reduced urn a ball of type i is replaced by a ball of type $i+1$ (with an exception when $k = m-1$ and $i = m-1$, in which case it is replaced by m balls of type 1), regardless of which particular type r was involved.

Case 3: If we draw a ball of type $i = m$, then in the original urn (assuming $n \geqslant m$) we replace a tree of type $\mathbf{x}^{(r)}$ (for which $x_m^{(r)} > 0$) by a tree of type $m+1$ and a tree of type $\mathbf{x}^{(s)} := \mathbf{x}^{(r)} - \mathbf{e}_m$, which inherits the root and therefore satisfies $s > r > k$. In the reduced urn we remove a ball of type m and what we add depends on the value of k: if $k \geqslant m+1$, we add a ball of type $m+1$; otherwise $k = m$ and we add m balls: $m-1$ external nodes and one node with a single key. In either case the outcome does not depend on r.

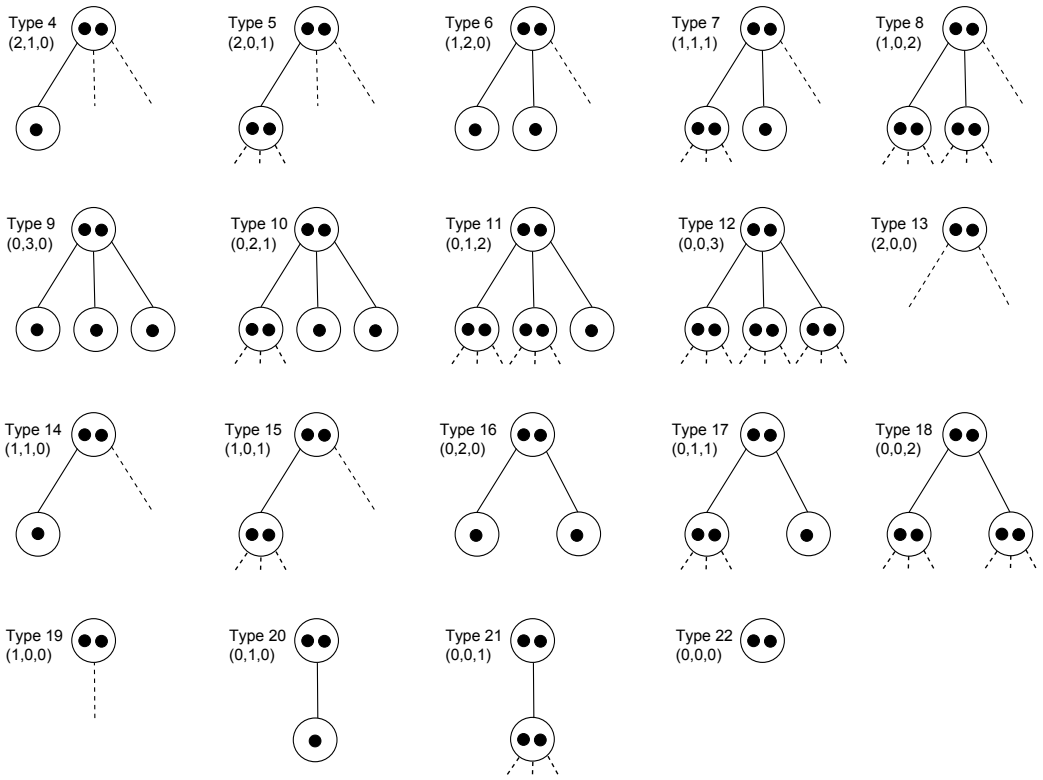

Figure 1: The 19 different types characterizing protected and unprotected nodes in ternary search trees for $n \geqslant 3$. Type 13, type 19 and type 22 are the only ones that include a protected node. External nodes are omitted but indicated by dotted lines.

To sum up, we have shown that $\mathcal{X}_n^k$, $k = m-1, \dots, q$, are Pólya urns with some random replacement vectors $\xi_i^{(k)}$, $i = 1, \dots, k$. Regardless of k, the activity of a type i is the number of gaps in the corresponding tree, which we denote a_i. So $a_i = i$ for $i = 1, \dots, m-1$ and $a_i = a_{\mathbf{x}^{(i)}}$ for $i \geqslant m$. Hence the intensity matrices of the reduced urns are

$$A_k := (a_j \, \mathbb{E}\, \xi_{ji}^{(k)})_{i,j=1}^{k}.$$

As in the fringe case, the diagonal values of A_k for $m \geqslant 3$ are easy to determine.

Proposition 10.1. *For $m \geqslant 3$, $k = m-1, \dots, q$ and every type $i = 1, \dots, k$ we have $(A_k)_{ii} = -a_i$.*

Proof. We need to show that whenever a ball of type i is drawn, it is never replaced. For $i \leqslant m$ this is clear from the discussion above (cases 2 and 3), so let us assume $i > m$. Suppose that a key is inserted into a child with less than $m-1$ keys. Then in the original urn a ball of type i is replaced by a ball of higher index j, so in the reduced urn i is replaced with either a ball of type j or some balls with types at most m, none of which coincide with type i. If a key is inserted into a full child of the root (which means $x_m^{(i)} > 0$), then also a ball of type $\mathbf{x}^{(m+1)} = (m-1, 1, 0, \dots, 0)$ is added. Since its last coordinate is zero, it cannot coincide with type i. □

The number of protected nodes is the total number of balls of types $(i, 0, \dots, 0)$, $i = 0, \dots, m-1$. The urn $\mathcal{X}_n^{q-1}$ ignores the dead type $(0, \dots, 0)$. However, writing b_i for the number of keys in a tree of type i, the number of balls of the dead type can be expressed as an affine function of the other types, namely

$$X_{n,q} = \frac{1}{m-1}\left(n - \sum_{i=m+1}^{q-1} b_i X_{n,i}\right).$$

Consider the urn $\mathcal{X}_n^{q-m-1}$ obtained by deleting the roots of trees with at most one child. For $n \geqslant m$, the first m coordinates of $\mathcal{X}_n^{q-m-1}$ (single nodes and the type $(m, 0, \dots, 0)$) are equal to the last m coordinates of $\mathcal{X}_n^{q-1}$ (roots with one child). Consequently, the urn described in Section 9 but with the dead type ignored, and thus $\binom{2m}{m} - 2 = q - m - 1$ types, is isomorphic to the urn $\mathcal{X}_n^{q-m-1}$ and has thus also intensity matrix A_{q-m-1} (up to relabelling the types). In particular, the number of protected nodes is an affine function of $\mathcal{X}_n^{q-m-1}$ and therefore, in order to prove asymptotic normality of the number of protected nodes, it is enough to prove asymptotic normality of $\mathcal{X}_n^{q-m-1}$. To deduce this from Theorem 4.1, it remains to show that $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$, which, for $m \leqslant 26$, follows from the next theorem.

Theorem 10.2. *Let $m \geqslant 2$. The eigenvalues of A_{q-m-1} are the roots of the polynomial $\phi_m(\lambda) = \prod_{i=1}^{m-1}(\lambda + i) - m!$ plus the multiset*

$$\{-a_i : i = m, m+1, \dots, q-m-1\}.$$

Proof. For $m = 2$, the eigenvalues were determined in [22] as $\{1, -2, -3, -4\}$ (when we ignore the dead type), which agrees with the statement.

For $m \geqslant 3$ the proof goes along the same lines as the proof of Theorem 6.2; we again show by induction the corresponding statement for A_k for $m-1 \leqslant k \leqslant q-m-1$. Note that urn $\mathcal{X}_n^{m-1}$ has the same meaning as in the fringe case, since $X_{n,i}^{m-1}$, $i = 1, \dots, m-1$, stands for the number of nodes with $i-1$ keys. Therefore A_{m-1} is defined by (35) and has characteristic polynomial $\phi_m(\lambda)$.

We use Proposition 10.1 instead of Proposition 6.1. The linear map T such that $\mathcal{X}_n^k = T\mathcal{X}_n^{k+1}$ is a $(k+1) \times k$ matrix obtained by appending to an identity matrix a column $(x_1, \dots, x_m, 0, \dots, 0)'$, where $\mathbf{x}^{(k+1)} = (x_1, \dots, x_m)$ is the vector describing type $k+1$. The rest of the proof is identical. □

Proof of Theorem 3.7. This is similar to the other proofs. As in Section 6, for $m \leqslant 26$, the roots $\lambda \neq \lambda_1$ of ϕ_m satisfy $\operatorname{Re}\lambda \leqslant \gamma_m < \frac{1}{2} = \lambda_1/2$, where γ_m is given by (41), and thus Theorem 10.2 shows that (42) holds for all eigenvalues $\lambda \neq \lambda_1$ of A_{q-m-1}. As said above, A_{q-m-1} is also the intensity matrix of the urn in Section 9, without the dead type, and if we reinstate the dead type, we just add one eigenvalue 0 (since the new column in A is identically 0). Hence Theorem 4.1 applies to the urn in Section 9, and implies, since Z_n is the total number of balls of the m types $(i, 0, \dots, 0)$, $0 \leqslant i \leqslant m-1$,

$$n^{-1/2}\big(Z_n - \mu' n\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2) \tag{66}$$

for some μ' and $\sigma^2 \geqslant 0$. Furthermore, it follows from [27] and (42) that (for any m)

$$\mathbb{E}(Z_n) = \mu' n + O\bigl(n^{\max(\gamma_m, 0)}\bigr). \tag{67}$$

Moreover, it follows from [23, Theorems 7.11, 10.1 and 10.3] that $Z_n/n \xrightarrow{\text{a.s.}} \mu_m$ given by (10), and consequently, by dominated convergence (see also [23, Remark 5.19])

$$\mathbb{E}(Z_n) = \mu n + o(n). \tag{68}$$

By (67) and (68), $\mu' = \mu$ (for any m). When $m \leqslant 26$, we have $\gamma_m < \frac{1}{2}$, and thus (67) implies (11), and thus $\mathbb{E}\, Z_n$ can be replaced by $\mu' n = \mu n$ in (9).

To see that $\sigma^2 > 0$, it suffices by [27, Theorem 3.6] to show that $\operatorname{Var}(Z_n) > 0$ for some n. This is easy; for example, with $n = 2m - 1$ keys, we can have either 0 or 1 protected node. □

11 Examples with explicit calculations of variances

We give some examples with explicit calculations (done using Mathematica).

11.1 Example of Theorem 3.5 when $m = 3$ and $k = 4$

We consider the case when we want to evaluate σ_4^2 in Theorem 3.5 in the case of a random ternary search tree ($m = 3$). We use the construction of the Pólya urn in Example 5.1, which gives an urn with the following 6 different (living) types:

1: An empty node.

2: A node with one key.

3: A node with two keys and three external children.

4: A tree with a root holding two keys and one child holding one key, plus two external children.

5: A tree with a root holding two keys and two children holding one key each, plus one external child.

6: A tree with a root holding two keys and one child holding two keys, plus two external children of the root and three external children of the leaf.

See Figure 2 for an illustration of these types.

The activities of the types are $1, 2, 3, 4, 5, 5$. We can easily describe the intensity matrix, first noting that if we draw a type k for $k \leqslant 3$ it is replaced by one of type $k + 1$. If we draw a type 4 it is replaced by one of type 5 with probability $1/2$ and one of type 6 with probability $1/2$. If we draw a type 5 it is replaced by three of type 2 with probability $1/5$, and one of each of the types 1, 2 and 3 with probability $4/5$; see Figure 3 for an

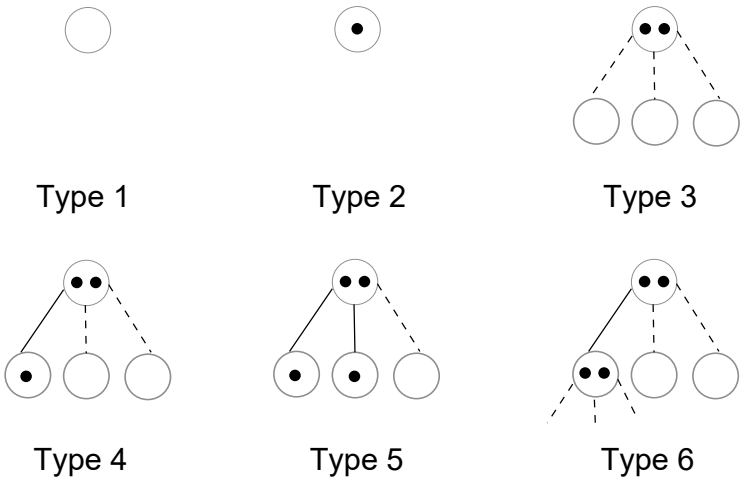

Figure 2: The different types for counting the number of the fringe subtrees with four keys in a ternary search tree.

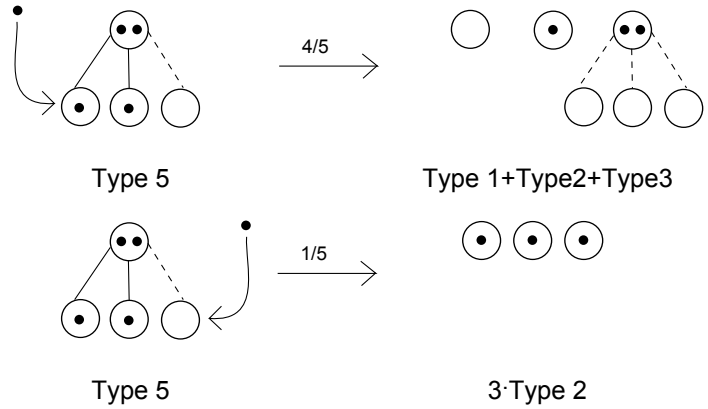

Figure 3: The two possibilities for adding a key to a tree of type 5 in Figure 2.

illustration. Finally if we draw a type 6 it is replaced by one each of the types 1, 2 and 3 with probability 2/5, and two of type 1 and one of type 4 with probability 3/5; see Figure 4 for an illustration.

Thus, we get the intensity matrix A in (19) as

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 4 & 8 \\ 1 & -2 & 0 & 0 & 7 & 2 \\ 0 & 2 & -3 & 0 & 4 & 2 \\ 0 & 0 & 3 & -4 & 0 & 3 \\ 0 & 0 & 0 & 2 & -5 & 0 \\ 0 & 0 & 0 & 2 & 0 & -5 \end{pmatrix}. \tag{69}$$

The eigenvalues are, by direct calculation or by Theorem 6.2,

$$1, -3, -4, -4, -5, -5. \tag{70}$$

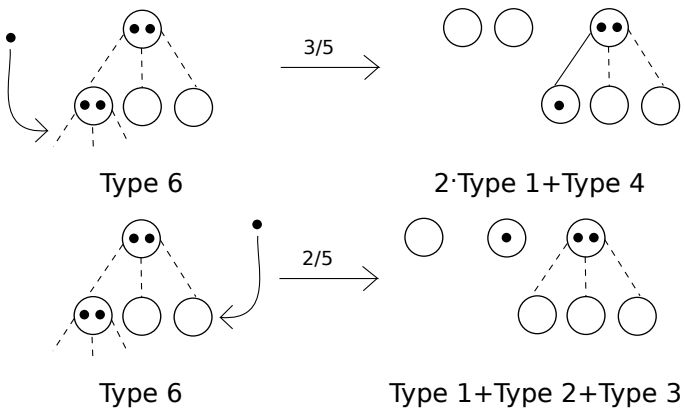

Figure 4: The two possibilities for adding a key to a tree of type 6 in Figure 2

(We know already that $\lambda_1 = 1$, as was noted in Remark 5.2.)

Furthermore, by Remark 5.2, the left eigenvector $u_1 = a = (1, 2, 3, 4, 5, 5)$. The right eigenvector v_1, with the normalization (20), is found to be

$$v_1 = \Big(\frac{3}{25}, \frac{1}{10}, \frac{2}{25}, \frac{3}{50}, \frac{1}{50}, \frac{1}{50}\Big). \tag{71}$$

Note that $\mu = v_1$ in the proof of Theorem 3.2 (of which Theorem 3.5 is a direct consequence), since $\lambda_1 = 1$. The fringe subtrees with 4 keys in the random ternary search tree correspond to the last two types, so $Y_{n,4} = X_n^{T^5} + X_n^{T^6}$. Hence, the expected number of subtrees with 4 keys is, see (45),

$$\mathbb{E}\, Y_{n,4} = (\mu_5 + \mu_6)n + o\big(n^{1/2}\big) = \frac{1}{25}n + o\big(n^{1/2}\big). \tag{72}$$

Note that this gives the same answer as Theorem 3.5 (where results from branching processes were applied to deduce the expectation), since the asymptotic expectation in (8) is

$$\frac{n}{(H_3 - 1)(4 + 1)(5 + 1)} = \frac{n}{25}.$$

To calculate the variance σ_4^2, we calculate the covariance matrix Σ in Theorem 4.1 by Theorem 4.1(ii); thus we first calculate B_i, B and Σ_I in (21)–(23). We describe the calculations briefly; for further details on how the calculations are performed, we refer to [22], where Theorem 4.1(ii) was applied to the urn in Figure 1 in Section 10 above to count the number of protected nodes in a random ternary search tree. Since A is diagonalisable, it is, as an alternative, also possible to calculate Σ by Theorem 4.1(iii).

We thus first calculate $B_i = \mathbb{E}(\xi_i \xi_i')$ in (21). As an example we have (the other cases are analogous)

$$B_5 = \tfrac{1}{5} \cdot b_1 b_1' + \tfrac{4}{5} \cdot b_2 b_2', \tag{73}$$

where

$$b_1 = (0,3,0,0,-1,0)' \qquad \text{and} \qquad b_2 = (1,1,1,0,-1,0)'. \tag{74}$$

We then calculate B by (22) and evaluate the integral in (23), which yields Σ_I. Finally, $\Sigma = \Sigma_I$ by Theorem 4.1 with $c = 1$. The result is

$$\Sigma = \begin{pmatrix} \frac{29017}{259875} & -\frac{117371}{10395000} & -\frac{44311}{5197500} & -\frac{2143}{945000} & -\frac{28289}{5197500} & -\frac{28289}{5197500} \\ -\frac{117371}{10395000} & \frac{7379}{83160} & -\frac{34927}{5197500} & -\frac{3907}{236250} & -\frac{166037}{20790000} & -\frac{166037}{20790000} \\ -\frac{44311}{5197500} & -\frac{34927}{5197500} & \frac{159241}{2598750} & -\frac{4747}{236250} & -\frac{84709}{10395000} & -\frac{84709}{10395000} \\ -\frac{2143}{945000} & -\frac{3907}{236250} & -\frac{4747}{236250} & \frac{39227}{945000} & -\frac{13309}{1890000} & -\frac{13309}{1890000} \\ -\frac{28289}{5197500} & -\frac{166037}{20790000} & -\frac{84709}{10395000} & -\frac{13309}{1890000} & \frac{22613}{1299375} & -\frac{6749}{2598750} \\ -\frac{28289}{5197500} & -\frac{166037}{20790000} & -\frac{84709}{10395000} & -\frac{13309}{1890000} & -\frac{6749}{2598750} & \frac{22613}{1299375} \end{pmatrix}. \tag{75}$$

However, to calculate σ_4^2, we only need the submatrix

$$\Delta = \begin{pmatrix} \sigma_{5,5} & \sigma_{5,6} \\ \sigma_{6,5} & \sigma_{6,6} \end{pmatrix} = \begin{pmatrix} \frac{22613}{1299375} & -\frac{6749}{2598750} \\ -\frac{6749}{2598750} & \frac{22613}{1299375} \end{pmatrix}. \tag{76}$$

Summing the $\sigma_{i,j}$ in (76), which is equivalent to calculating $(1,1)\Delta(1,1)'$, we find

$$\sigma_4^2 = \frac{38477}{1299375}.$$

Note that we can use this urn to calculate the asymptotic variance for the number of leaves in the random ternary search tree, which was evaluated in [22, Theorem 4.1]. We get

$$(0,1,1,1,2,1)\Sigma(0,1,1,1,2,1)' = \frac{89}{2100}. \tag{77}$$

We could also use this urn to evaluate

$$\sigma_3^2 = (0,0,0,1,0,0)\Sigma(0,0,0,1,0,0)' = \frac{39227}{945000}, \tag{78}$$

$$\sigma_2^2 = (0,0,1,0,0,1)\Sigma(0,0,1,0,0,1)' = \frac{131}{2100}, \tag{79}$$

$$\sigma_1^2 = (0,1,0,1,2,0)\Sigma(0,1,0,1,2,0)' = \frac{8}{75}. \tag{80}$$

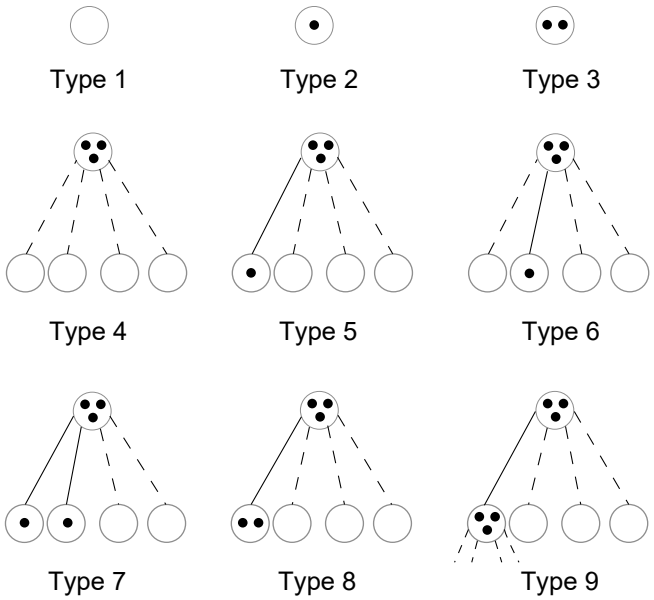

Figure 5: The different types used to find the joint distribution of fringe trees that are isomorphic to type 7 and type 9 in a quaternary search tree.

11.2 Example of Theorem 3.2 when $m = 4$

We consider the random quaternary search tree ($m = 4$) as an ordered tree. Suppose that we consider two fringe subtrees in this tree. Let the first one T^1 consist of 5 keys i.e., $k_1 = 5$, so that the root holds three keys with its first two children holding one key each, and with its remaining two external children to the right. Let the second one T^2 consist of 6 keys i.e., $k_2 = 6$, so that the root holds three keys, and with its first child holding three keys and thus also having four external children, and with the remaining three external children of the root to the right.

We use the construction of the Pólya urn in Section 5 which gives an urn with the following 9 different (living) types, see Figure 5:

1–4 For $k \leqslant 4$, type k is a node with $k-1$ keys; the fourth type also has four external children.

5 A root with three keys with its four children having $(1,0,0,0)$ keys.

6 A root with three keys with its four children having $(0,1,0,0)$ keys.

7 A root with three keys with its four children having $(1,1,0,0)$ keys.

8 A root with three keys with its four children having $(2,0,0,0)$ keys.

9 A root with three keys with its four children having $(3,0,0,0)$ keys, and the first child having four external children.

We get the intensity matrix as in the example in Section 11.1. We describe one example of a transition, the others are similar. If we draw a type 5 it is replaced by one of type 7 with probability 1/5, and one of type 8 with probability 2/5, and two of type 1 and two of type 2 with probability 2/5; see Figure 6.

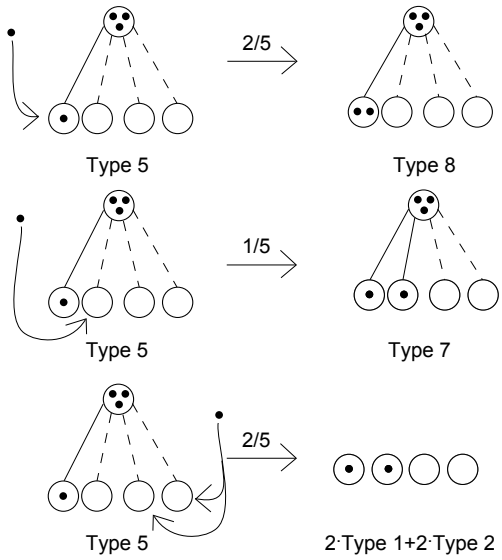

Figure 6: The three possibilities for adding a key to a tree of type 5 in Figure 5.

The intensity matrix is

$$A = \begin{pmatrix} -1 & 0 & 0 & 6 & 4 & 10 & 10 & 6 & 24 \\ 1 & -2 & 0 & 2 & 4 & 4 & 10 & 3 & 5 \\ 0 & 2 & -3 & 0 & 0 & 2 & 4 & 3 & 0 \\ 0 & 0 & 3 & -4 & 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 1 & -5 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & -5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & -6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & -7 \end{pmatrix}. \tag{81}$$

The eigenvalues are, by direct calculation or by Theorem 6.2,

$$1, -\tfrac{7}{2} + \tfrac{\sqrt{23}}{2}i, -\tfrac{7}{2} - \tfrac{\sqrt{23}}{2}i, -4, -5, -5, -6, -6, -7. \tag{82}$$

Note that the activity vector $a = (1, 2, 3, 4, 5, 5, 6, 6, 7)$ and it again follows that the left eigenvector $u_1 = a$. The right eigenvector v_1, with the normalization (20), is calculated as

$$v_1 = \Big(\frac{113}{520}, \frac{61}{455}, \frac{34}{455}, \frac{33}{728}, \frac{1}{130}, \frac{1}{130}, \frac{1}{455}, \frac{1}{455}, \frac{3}{3640}\Big). \tag{83}$$

We see from this vector that the expected number of fringe subtrees of the two types that we consider is $(\frac{1}{455} + \frac{3}{3640})n + o(n) = \frac{11}{3640}n + o(n)$, since the two types are type 7 and

type 9. We can verify that this gives the same answer as Theorem 3.2 (where results from branching processes were applied to deduce the expectation), since the vector $\hat{\boldsymbol{\mu}}$ in (3) has coordinates

$$\frac{\mathbb{P}(\mathcal{T}_5 = T^1)}{(H_4-1)(5+1)(5+2)} = \frac{\frac{3}{5}\cdot\frac{2}{4}\cdot\frac{1}{3}}{(\frac{1}{2}+\frac{1}{3}+\frac{1}{4})\cdot 6\cdot 7} = \frac{1}{455} \tag{84}$$

and

$$\frac{\mathbb{P}(\mathcal{T}_6 = T^2)}{(H_4-1)(6+1)(6+2)} = \frac{\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{1}{4}}{(\frac{1}{2}+\frac{1}{3}+\frac{1}{4})\cdot 7\cdot 8} = \frac{3}{3640}. \tag{85}$$

We proceed by calculating the covariance matrix Σ. It again turns out that A is diagonalisable, and this time we apply Theorem 4.1(iii). We again have to calculate the matrix B in (22), and for this we have to calculate $B_i = \mathbb{E}(\xi_i\xi_i')$ in (21). For example,

$$B_7 = \tfrac{4}{6}\cdot b_1 b_1' + \tfrac{2}{6}\cdot b_2 b_2', \tag{86}$$

where

$$b_1 = (2,1,1,0,0,0,-1,0,0)' \qquad \text{and} \qquad b_2 = (1,3,0,0,0,0,-1,0,0)'. \tag{87}$$

Having calculated $B_1,\dots,B_9$ in this way, we obtain the matrix B from (22) and then the covariance matrix Σ by (26); the result is shown in Appendix A. However, to evaluate the joint distribution of the fringe subtrees T^1 and T^2 described above, we only need the submatrix

$$\Gamma = \begin{pmatrix} \sigma_{7,7} & \sigma_{7,9} \\ \sigma_{9,7} & \sigma_{9,9} \end{pmatrix} = \begin{pmatrix} \frac{157523}{72872800} & -\frac{2884319}{194424630400} \\ -\frac{2884319}{194424630400} & \frac{5681341}{6943736800} \end{pmatrix}. \tag{88}$$

Note that we can also use this urn to calculate the asymptotic variance σ_1^2 for the number of leaves in the random quaternary search tree; we get

$$\sigma_1^2 = (0,1,1,1,1,1,2,1,1)\Sigma(0,1,1,1,1,1,2,1,1)' = \frac{5276}{122525}. \tag{89}$$

We can alternatively obtain σ_1^2 by using the simple Pólya urn with the four types illustrated in Figure 7. (As another simple example of Theorem 3.2 and the construction in Section 5, see also [22, Section 5.2] for a minor variation of this urn.) The intensity matrix is

$$A = \begin{pmatrix} -1 & 0 & 0 & 12 \\ 1 & -2 & 0 & 4 \\ 0 & 2 & -3 & 0 \\ 0 & 0 & 3 & -4 \end{pmatrix}. \tag{90}$$

The eigenvalues are, by direct calculation or by Theorem 6.2,

$$1, -\tfrac{7}{2}+\tfrac{\sqrt{23}}{2}i, -\tfrac{7}{2}-\tfrac{\sqrt{23}}{2}i, -4. \tag{91}$$

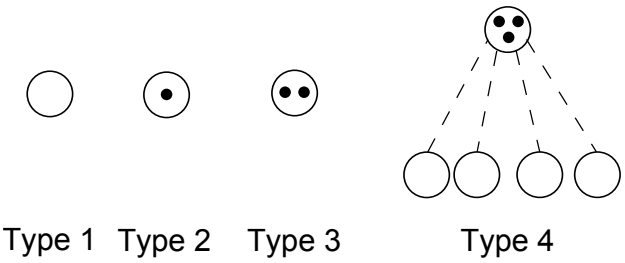

Figure 7: The different types used to study the number of leaves in a quaternary search tree.

Since A is diagonalisable we may again use Theorem 4.1(iii) to calculate Σ. We obtain

$$\Sigma = \begin{pmatrix} \frac{34466}{122525} & \frac{153}{49010} & -\frac{963}{24505} & -\frac{10393}{245050} \\ \frac{153}{49010} & \frac{519}{4901} & -\frac{339}{9802} & -\frac{681}{24505} \\ -\frac{963}{24505} & -\frac{339}{9802} & \frac{276}{4901} & -\frac{57}{3770} \\ -\frac{10393}{245050} & -\frac{681}{24505} & -\frac{57}{3770} & \frac{4391}{122525} \end{pmatrix}. \tag{92}$$

From Σ we see that the asymptotic variance for the number of leaves in a random quaternary search tree is

$$(0,1,1,1)\Sigma(0,1,1,1)' = \frac{5276}{122525}, \tag{93}$$

which equals the result in (89).

11.3 Example of Theorem 3.11 when $k = 3$

We consider the case when we want to evaluate σ_3^2 in Theorem 3.11 in the case of a linear preferential attachment tree. As explained in Section 2.2, only the quotient χ/ρ matters, and thus we may assume that $\chi \in \{-1, 0, 1\}$. For (notational) simplicity, we consider only the case $\chi = 1$ in the formulas below; the cases $\chi = 0, -1$ (and general χ) are similar but are left to the reader. The final results will be expressed in the parameter $\kappa = \rho/(\chi + \rho)$ in (14); these results are valid for all values of χ and ρ. (This follows either by checking the other cases, or by analytic continuation, since the eigenvector v_1 and the integral (23) are analytic functions of (χ, ρ) in a suitable domain.)

We use the construction of the Pólya urn in Section 7.2, with $\mathcal{S}'$ the set of (unordered) trees with at most 3 nodes. (Cf. Example 5.1 for m-ary search trees.) This gives an urn with the following 5 different types, see Figure 8:

1: An empty node.

2: A path with two nodes.

3: A path with three nodes.

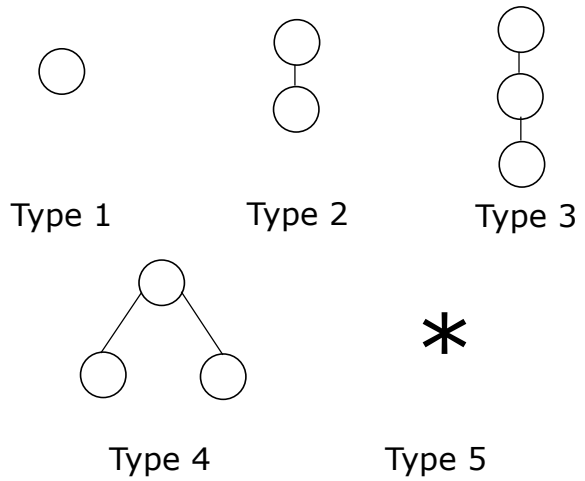

Figure 8: The different types used to study the number of fringe subtrees with three nodes in a preferential attachment tree.

4: A tree consisting of one root with two children.

5: A special type (with activity 1).

We get the intensity matrix as in the examples in Section 11.1 and Section 11.2. We describe one example of a transition, the others are similar. If we draw a type 3 it is replaced by, see Figure 9,

- 1 of type 1, 1 of type 2, and $2+\rho$ of type 5 with probability $\frac{1+\rho}{2+3\rho}$;
- 1 of type 4 and $1+\rho$ of type 5 with probability $\frac{1+\rho}{2+3\rho}$;
- 1 of type 3 and $1+\rho$ of type 5 with probability $\frac{\rho}{2+3\rho}$.

The intensity matrix is

$$A = \begin{pmatrix} -\rho & 0 & \rho+1 & 5\rho+6 & 1 \\ \rho & -2\rho-1 & \rho+1 & 2\rho & 0 \\ 0 & \rho & -2\rho-2 & 0 & 0 \\ 0 & \rho+1 & \rho+1 & -3\rho-2 & 0 \\ 0 & 0 & 3(\rho+1)^2 & 3(\rho+1)(\rho+2) & 1 \end{pmatrix}. \tag{94}$$

The eigenvalues are, simplest by Theorem 8.2,

$$\rho+1, -\rho, -2\rho-1, -3\rho-2, -3\rho-2.$$

It again turns out that A is diagonalisable. To calculate Σ we apply Theorem 4.1(iii). We again have to first calculate B_i and B in (21)–(22).

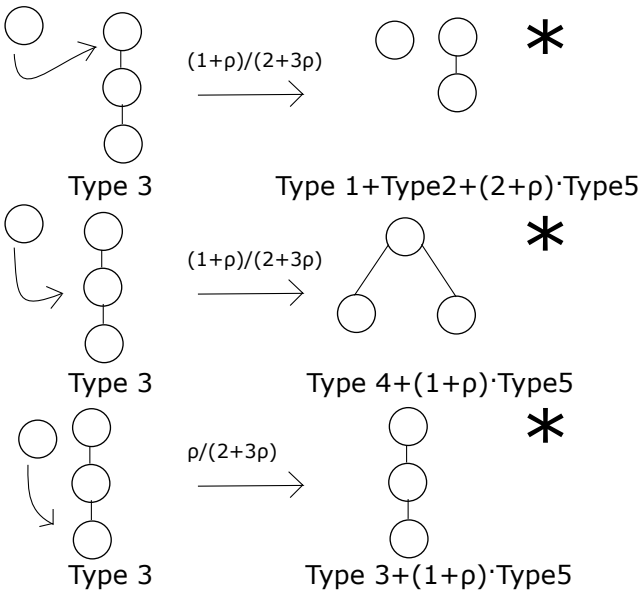

Figure 9: The three possibilities for adding an additional node to a tree of type 3 in Figure 8.

Note that the activity vector $a = (\rho, 1+2\rho, 2+3\rho, 2+3\rho, 1)$ and it again follows that the left eigenvector $u_1 = a$. The eigenvector v_1 (with the normalisation (20)) is calculated to be

$$v_1 \;=\; \frac{1}{(2\rho+1)(3\rho+2)(4\rho+3)}\Big(6(\rho\,+\,1)^2, 3\rho(\rho\,+\,1), \rho^2, \rho(\rho\,+\,1), 6(\rho\,+\,1)^3\Big). \quad (95)$$

We express this using κ in (14), which yields

$$v_1 \;=\; \frac{1}{(\kappa+1)(\kappa+2)(\kappa+3)}\Big(6(1\,-\,\kappa), 3(1\,-\,\kappa)\kappa, (1\,-\,\kappa)\kappa^2, (1\,-\,\kappa)\kappa, 6\Big). \quad (96)$$

Recall from the proof of Theorem 3.9 (of which Theorem 3.11 is a direct consequence) that $\mu = \lambda_1 v_1$, where $\lambda_1 = \rho + 1 = 1/(1-\kappa)$. The fringe subtrees with three nodes correspond to type 3 and type 4, so $Y_{n,3} = X_{n,3} + X_{n,4}$. Hence, from (61) and (96) we obtain

$$\mathbb{E}\, Y_{n,3} = \frac{1}{1-\kappa}\,\frac{(1-\kappa)\kappa^2 + (1-\kappa)\kappa}{(\kappa+1)(\kappa+2)(\kappa+3)}\, n + O(1) = \frac{\kappa}{(\kappa+2)(\kappa+3)} n + O(1), \quad (97)$$

which agrees with Theorem 3.11 and (18).

We next calculate $B_i = \mathbb{E}(\xi_i \xi_i')$ in (21); we take B_3 as an example, where we get, see Figure 9,

$$B_3 = \tfrac{1+\rho}{2+3\rho} \cdot b_1 b_1' + \tfrac{1+\rho}{2+3\rho} \cdot b_2 b_2' + \tfrac{\rho}{2+3\rho} \cdot b_3 b_3', \quad (98)$$

with

$$b_1 = (1, 1, -1, 0, \rho + 2)', \tag{99}$$
$$b_2 = (0, 0, -1, 1, \rho + 1)', \tag{100}$$
$$b_3 = (0, 0, 0, 0, \rho + 1)'. \tag{101}$$

We then find the matrix B. Finally, the eigenvectors u_i and v_i are calculated for all eigenvalues and the covariance matrix Σ is calculated by (26). The covariances are listed in Appendix B, again using the notion κ in (14). As said above, these formulas are valid for all χ and ρ.

Returning to the number of fringe subtrees of order 3 we thus obtain

$$\begin{aligned}\sigma_3^2 &= (0,0,1,1,0)\Sigma(0,0,1,1,0)' = \Sigma_{3,3} + 2\Sigma_{3,4} + \Sigma_{4,4} \\ &= \frac{3\kappa\,(11\kappa^3 + 52\kappa^2 + 77\kappa + 30)}{2(\kappa+2)(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)}.\end{aligned} \tag{102}$$

We can check that this formula yields previously known results (obtained by other methods) in the three most important special cases, Examples 2.3–2.5.

For the random recursive tree, $\kappa = 1$ and (102) yields $\sigma_3^2 = 17/336$, which equals the result given by Devroye [8, Theorem 4], where σ_k^2 was calculated for general k (using different methods), see also [16] and [21, Proposition 1.13 and (1.20)].

For the binary search tree, $\kappa = 2$ and (102) yields $\sigma_3^2 = 8/175$, which agrees with the result by Devroye [8, Theorem 5] (for general k), see also [21, Proposition 1.10].

For the plane oriented recursive tree, $\kappa = 1/2$ and (102) yields $\sigma_3^2 = 663/15680$. This variance was calculated, for general k, by Fuchs [17, Theorem 1.1] by other methods (generating functions).

Remark 11.1. There is a mistake in the formula for the asymptotic variance in [17, Theorem 1.1]: the numerator $8k^2 - 4k - 8$ should be $8k^2 - 4k$. (The reason is that in the calculation of $\operatorname{Var}(X_{n,k})$ on [17, p. 419], there should be a plus sign in front of $\frac{4}{(4k^2-1)^2}$ instead of a minus sign.) With this correction, (102) (with $\kappa = 1/2$) agrees with the value for $k = 3$ of the formula in [17, Theorem 1.1], and the values for σ_1^2 and σ_2^2 obtained below agree with the values of the formula for $k = 1, 2$.

Note that we can use the Pólya urn in this example to calculate also σ_1^2 and σ_2^2, i.e. the constants in the asymptotic variances for the numbers $Y_{n,1}$ and $Y_{n,2}$ of fringe subtrees of size 1 and 2 in a linear preferential attachment tree. Note that $Y_{n,1}$ is simply the number of leaves, i.e., the number of nodes of out-degree 0. We have $Y_{n,1} = X_{n,1} + X_{n,2} + X_{n,3} + 2X_{n,4}$ and $Y_{n,2} = X_{n,2} + X_{n,3}$, see Figure 8. Thus, again using Appendix B,

$$\sigma_1^2 = (1,1,1,2,0)\Sigma(1,1,1,2,0)' = \frac{\kappa}{(\kappa+1)^2(2\kappa+1)}, \tag{103}$$
$$\sigma_2^2 = (0,1,1,0,0)\Sigma(0,1,1,0,0)' = \frac{\kappa\,(5\kappa^2 + 10\kappa + 6)}{(\kappa+1)(\kappa+2)^2(2\kappa+1)(2\kappa+3)}. \tag{104}$$

However, these values can also be obtained by smaller urns, yielding simpler calculations, for examples by the urn with only types 1, 2 and 5 (the special type) above. For $k = 1$ it suffices to use the urn with two colours and intensity matrix (59) used in the proof of Theorem 8.2, see Example 12.5.

We can again check that the results agrees with known results when $\kappa = 1, \frac{1}{2}, 2$. For the random recursive tree ($\kappa = 1$), (103)–(104) yield $\sigma_1^2 = 1/12$ and $\sigma_2^2 = 7/90$, as shown in [8, Theorem 4], see also [21, Proposition 1.13]; for the plane oriented recursive tree ($\kappa = 1/2$) we obtain $\sigma_1^2 = 1/9$ and $\sigma_2^2 = 49/600$, as shown in [37] and [17, Theorem 1.1] (see Remark 11.1); for the binary search tree ($\kappa = 2$) we obtain $\sigma_1^2 = 2/45$ and $\sigma_2^2 = 23/420$, as shown in [8, Theorem 5], see also [16] and [21, Proposition 1.10].

12 Degrees

By using (simpler) variants of the Pólya urns described above for studying fringe subtrees in m-ary search trees and preferential attachment trees, we can also easily prove normal limit theorems for the out-degrees of the nodes in both of these models.

12.1 Out-degrees in m-ary search trees

We first consider m-ary search trees for which by out-degree we mean the number of internal children. The following theorem was recently proved by Kalpathy and Mahmoud [31] using a Pólya urn (based on gaps instead of nodes) that is equivalent to the one used here. (A simpler version, A_m in the proof below, was used in [22] to study the special case of the number of leaves.) We nevertheless sketch a proof in order to show the connections with the analysis in the previous sections, and in particular the induction argument for the eigenvalues. (In [31], the eigenvalues were calculated numerically.)

Theorem 12.1. *Let $D_{n,k}$ be the number of nodes with out-degree k in the random m-ary search tree $\mathcal{T}_n$. If $m \leqslant 26$, then, as $n \to \infty$,*

$$\frac{D_{n,k} - \mu_k n}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{105}$$

with

$$\mu_k = \begin{cases} \frac{m-1}{2(H_m-1)(m+1)}, & k = 0, \\ \frac{1}{(H_m-1)m(m+1)}, & 1 \leqslant k \leqslant m, \end{cases} \tag{106}$$

where σ_k^2 is some positive constant.

Proof. We construct a Pólya urn by chopping up the m-ary search tree as in Section 5, but we now erase all edges from parents to internal children, keeping only the edges to external children. Hence, the small trees in the resulting forest, which are represented by balls in the urn, are of the following $2m - 1$ types. (We regard the trees as unordered.)

- A single internal node with $1, \dots, m - 2$ keys.

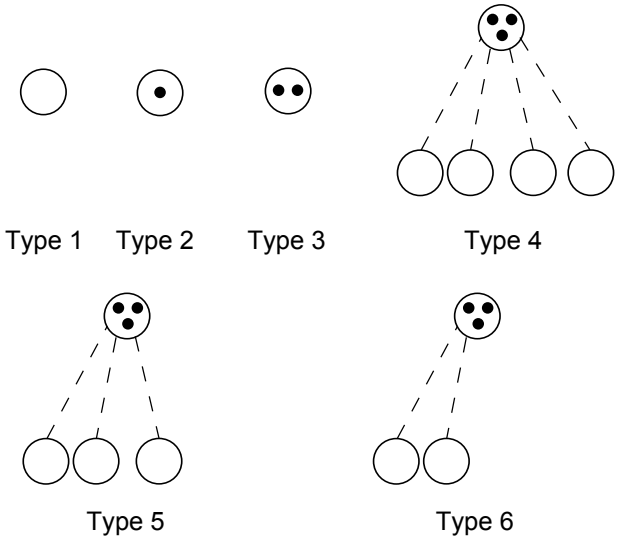

Figure 10: The alternative types used to characterize the out-degrees in a quaternary search tree.

- A root with $m-1$ keys and $0, \dots, m$ external children.

We can simplify a little by noting that the type consisting of a root with 0 external children is dead (activity 0) so it does not affect the evolution of the urn and can be ignored (although it should be included in the final count of node degrees; it represents the nodes of degree m). Moreover, the type with 1 external child can instead be represented by a single external node (although it should be counted as a node of degree $m-1$). This yields a Pólya urn with the following $2m-2$ types. (See Figure 10, which shows such alternative types in the case $m = 4$.)

$1, \dots, m-1$ Type i is a single node with $i-1$ keys. There are i gaps and thus the activity $a_i = i$.

$m, \dots, 2m-2$ Type i is a node with $m-1$ keys and $2m-i$ external children. The activity $a_i = 2m-i$. (The out-degree is $i-m$.)

If we draw a ball of type $i \leqslant m-2$, it is replaced by a ball of type $i+1$. Similarly, a ball of type $m-1$ is replaced by a ball of type m. A ball of type $i \in \{m, \dots, 2m-3\}$ is replaced by a ball of type $i+1$ and a ball of type 2. A ball of type $2m-2$ is replaced by a ball of type 1 and a ball of type 2.

For our induction argument, we also consider a reduced urn with types $1, \dots, k$, for $m-1 \leqslant k \leqslant 2m-2$, obtained by chopping up also all trees of types $j > k$. In other words, we replace a ball of type $j > k$ by $2m-j$ balls of type 1. (Just as we already have cut up trees with a single external child.) This reduced urn thus ignores all nodes with $m-1$ keys and degree $> k-m$.

Let A_k be the intensity matrix of this urn with k types. When a ball is drawn, there will never be a ball of the same type in the set of balls replacing it. Hence every $\xi_{ii} = -1$

and the diagonal elements of A_k are $-a_i$, where a_i is the activity of type i. For $k = m-1$, the reduced urn is the same as the urn with the $m-1$ first types in Section 6, so A_{m-1} is given by (35) and its eigenvalues are the roots of $\phi_m(\lambda) := \prod_{i=1}^{m-1}(\lambda+i) - m!$. It now follows by the same induction argument as in Section 6 that the eigenvalues of A_k are the roots of ϕ_m plus $\{-m, -(m-1), \dots, -(2m-k)\}$. In particular, taking $k = 2m-2$, the eigenvalues of the intensity matrix $A = A_{2m-2}$ are the eigenvalues of ϕ_m plus the negative numbers $-2, \dots, -m$.

Hence, for every eigenvalue $\lambda \neq \lambda_1 = 1$, $\operatorname{Re}\lambda \leqslant \max(\gamma_m, -2) \leqslant \max(\gamma_m, 0)$, cf. (41), and thus Theorem 4.1 applies when $m \leqslant 26$. The rest of the proof is as in the proof of Theorem 3.2 and other proofs above. The constants (106) can be found either by finding the eigenvector v_1 of the intensity matrix A explicitly, as in [31], or by comparison with [23, Theorems 7.11 and 7.14] (proved using branching processes).

Finally, $\sigma_k^2 > 0$ by another application of [27, Theorem 3.6] and the fact that $D_{n,k}$ is not deterministic for all n, as is easily seen. □

12.2 Out-degrees in preferential attachment trees

Mahmoud and Smythe [36] used a Pólya urn to show asymptotic normality for the numbers of nodes of out-degrees 0, 1 and 2 in a random recursive tree, and Mahmoud, Smythe and Szymański [37] did the same for a plane oriented recursive tree; these results were extended to arbitrary degrees by Janson [25]. We can extend this to general linear preferential attachment trees (using essentially the same urn).

In the case $\chi < 0$, when we can assume $\chi = -1$ and $\rho = m$ for some integer m, the resulting tree has no nodes of out-degree $> m$, and we consider only out-degrees $k \leqslant m$ in the following theorem; otherwise k is arbitrary. For the asymptotic proportions μ_k in (108), see also [23, (6.33)].

Theorem 12.2. *Let $\hat{D}_{n,k}$ be number of nodes with out-degree k in the linear preferential attachment tree Λ_n defined by the weights* (1)*. Then, as $n \to \infty$,*

$$\frac{\hat{D}_{n,k} - \mu_k n}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{107}$$

with some $\sigma_k^2 > 0$ and

$$\mu_k = \frac{w_1}{w_0 + w_1} \prod_{i=1}^{k} \frac{w_{i-1}}{w_i + w_1} = \frac{\chi + \rho}{\chi + 2\rho} \prod_{i=1}^{k} \frac{\chi(i-1) + \rho}{\chi(i+1) + 2\rho}. \tag{108}$$

Proof. As in Section 7, we begin by constructing an urn with infinitely many types. The types are $\{0, 1, 2, \dots\}$, and a ball of type i simply represents a node of out-degree i. The activity of type i is thus $a_i = w_i = \chi i + \rho$, see (1). When a ball of type i is drawn, it is replaced by a ball of type $i+1$ and a ball of type 0. (In the case $\chi < 0$, we consider only a finite number of types so we have a finite urn; we leave the minor modifications in this case to the reader.)

To get an urn with finitely many types, we truncate as in Section 7.2 (and [25]); we choose an integer $k \geqslant 0$ and use the $k+2$ types $\{0,1,\dots,k\} \cup \{*\}$, where the new type $*$ represents the activity of the nodes with out-degrees $> k$. Hence, $*$ has activity 1. If we draw a ball of type $i < k$, we replace it by a ball of type $i+1$ and a ball of type 0, as before; if we draw a ball of type k, we replace it by a ball of type 0 and $w_{k+1} = (k+1)\chi+\rho$ balls of type $*$; if we draw a ball of type $*$, it is replaced and we add χ additional balls of type $*$ and a ball of type 0.

Let A_{k+2} denote the $(k+2) \times (k+2)$ intensity matrix of this urn. For $k = 0$ we have the same urn as in the proof of Theorem 8.2, so A_2 is given by (59) with the eigenvalues $\chi+\rho$ and $-\rho$. As in the proof of Theorem 6.2, the eigenvalues of A_k are inherited by A_{k+1}, and a simple induction shows that the eigenvalues of A_{k+2} are $\chi+\rho$ and $-w_j = -(\chi j+\rho)$ for $0 \leqslant j \leqslant k$. In particular, all eigenvalues except $\lambda_1 = \chi+\rho$ are negative. Hence Theorem 4.1 applies and the proof is completed as the other proofs. The constants μ_k can be found by verifying directly that (108) yields an eigenvector of the intensity matrix A for the infinite urn, and thus it is mapped to an eigenvector for A_{k+2} by the truncation above. Alternatively, we can use [23, (6.14)]. The claim that $\sigma_k > 0$ follows, as before, from [27, Theorem 3.6] and the easy fact that $\hat{D}_{n,k}$ is non-deterministic. □

Remark 12.3. We can modify the urn in the proof of Theorem 12.2 to an equivalent one by changing each ball of type $i \leqslant k$ to w_i new balls of type i; the new balls all have activity 1 and can be interpreted as gaps. (This is the urn actually used in [25].) The new intensity matrix A_{k+2} has the same eigenvalues as the old one, but it has the advantage that it is homogeneous of degree 1 in ρ and χ, i.e., each entry is a linear combination of χ and ρ.

Remark 12.4. In both cases, we also obtain asymptotic joint normal distributions of the numbers $D_{n,k}$ and $\hat{D}_{n,k}$ for different k.

Example 12.5. The simplest case is $k = 0$, when $\hat{D}_{n,0}$ is the number of leaves in Λ_n. In this case, (108) yields

$$\mu_0 = \frac{w_1}{w_0+w_1} = \frac{\chi+\rho}{\chi+2\rho} = \frac{1}{\kappa+1}. \tag{109}$$

Moreover, the proof above yields the urn with two colours and intensity matrix A_2 given by (59). The eigenvalues of A_2 are by direct calculation or by Theorem 8.2, $\lambda_1 = \chi+\rho$ and $-\rho$. Simple calculations show that

$$B = \begin{pmatrix} \frac{\rho+\chi}{2\rho+\chi} & \frac{\chi(\rho+\chi)}{2\rho+\chi} \\ \frac{\chi(\rho+\chi)}{2\rho+\chi} & \frac{(\rho+\chi)(\rho^2+\chi\rho+\chi^2)}{2\rho+\chi} \end{pmatrix} \tag{110}$$

and the covariance matrix Σ is, using, for example, Theorem 4.1(iii) again,

$$\Sigma = \begin{pmatrix} \frac{\kappa}{(\kappa+1)^2(2\kappa+1)} & \frac{\kappa^2}{(\kappa-1)(\kappa+1)^2(2\kappa+1)} \\ \frac{\kappa^2}{(\kappa-1)(\kappa+1)^2(2\kappa+1)} & \frac{\kappa^3}{(2\kappa+1)(\kappa^2-1)^2} \end{pmatrix}. \tag{111}$$

Thus, the asymptotic variance of the number of leaves is

$$\sigma_1^2 = (1,0)\Sigma(1,0)' = \frac{\kappa}{(\kappa+1)^2(2\kappa+1)}, \tag{112}$$

which equals the result in (103). (This was shown for the random recursive tree and the binary search tree in [8], and for the plane oriented recursive tree in [37].)

Example 12.6. We also consider the case when $k = 1$. In this case the proof above yields the urn with three colours and intensity matrix A_3 given by

$$A_3 = \begin{pmatrix} 0 & \rho+\chi & 1 \\ \rho & -\rho-\chi & 0 \\ 0 & (\rho+\chi)(\rho+2\chi) & \chi \end{pmatrix}. \tag{113}$$

The eigenvalues of A_3 are $\lambda_1 = \chi+\rho$, $-\rho$ and $-(\chi+\rho)$. Simple calculations show that

$$B = \begin{pmatrix} \frac{\rho+\chi}{2\rho+\chi} & -\frac{\rho}{2(2\rho+\chi)} & \frac{(\rho+\chi)(\rho+2\chi)}{2(2\rho+\chi)} \\ -\frac{\rho}{2(2\rho+\chi)} & \frac{3\rho}{2(2\rho+\chi)} & -\frac{\rho(\rho+2\chi)}{2(2\rho+\chi)} \\ \frac{(\rho+\chi)(\rho+2\chi)}{2(2\rho+\chi)} & -\frac{\rho(\rho+2\chi)}{2(2\rho+\chi)} & \frac{(\rho+\chi)^2(\rho+2\chi)}{2(2\rho+\chi)} \end{pmatrix} \tag{114}$$

and, for example, using Theorem 4.1(iii),

$$\Sigma = \begin{pmatrix} \frac{\kappa}{(\kappa+1)^2(2\kappa+1)} & -\frac{\kappa\left(2\kappa^2+3\kappa+2\right)}{2(\kappa+1)^2(\kappa+2)(2\kappa+1)} & \frac{\kappa(2-\kappa)}{2(\kappa+1)^2(\kappa+2)(2\kappa+1)} \\ -\frac{\kappa\left(2\kappa^2+3\kappa+2\right)}{2(\kappa+1)^2(\kappa+2)(2\kappa+1)} & \frac{\kappa\left(2\kappa^3+25\kappa^2+32\kappa+12\right)}{12(\kappa+1)^2(\kappa+2)(2\kappa+1)} & -\frac{\kappa(2-\kappa)\left(10\kappa^2+13\kappa+6\right)}{12(\kappa+1)^2(\kappa+2)(2\kappa+1)} \\ \frac{\kappa(2-\kappa)}{2(\kappa+1)^2(\kappa+2)(2\kappa+1)} & -\frac{\kappa(2-\kappa)\left(10\kappa^2+13\kappa+6\right)}{12(\kappa+1)^2(\kappa+2)(2\kappa+1)} & \frac{\kappa(2-\kappa)\left(10\kappa^2+7\kappa+6\right)}{12(\kappa+1)^2(\kappa+2)(2\kappa+1)} \end{pmatrix}. \tag{115}$$

In the upper left corner, we find again σ_1^2 in (103) and (112).

The upper left 2×2 submatrix of Σ, giving the asymptotic variances and covariance of the numbers of nodes of out-degrees 1 and 2, was found in [8] for the binary search tree ($\kappa = 2$), in [36] for the random recursive tree ($\kappa = 1$) and in [37] for the plane oriented recursive tree ($\kappa = 1/2$), see also [25].

Acknowledgements

We thank the anonymous referees for very careful reviews and for pointing out additional references.

References

[1] Aldous D., Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.* **1** (1991), no. 2, 228–266.

[2] Barabási A.L. and Albert R., Emergence of scaling in random networks. *Science* 15 October 1999, **286** (1999), no. 5439, 509–512.

[3] Bóna M., k-protected nodes in binary search trees. *Adv. Appl. Math.* **53** (2014), 1–11.

[4] Chauvin B. and Pouyanne N., m-ary search trees when $m \geqslant 27$: a strong asymptotics for the space requirement. *Random Structures Algorithms* **24**, (2004), 133–154.

[5] Cheon G.S. and Shapiro L., Protected points in ordered trees. *Appl. Math. Lett.* **21** (2008), no. 5, 516–520.

[6] Chern H.-H. and Hwang H.-K., Phase changes in random m-ary search trees and generalized quicksort. *Random Structures Algorithms* **19** (2001), no. 3-4, 316–358.

[7] Dennert F. and Grübel R., On the subtree size profile of binary search trees. *Combin. Probab. Comput.* **19** (2010), no. 4, 561–578.

[8] Devroye L., Limit laws for local counters in random binary search trees. *Random Structures Algorithms* **2** (1991), no. 3, 303–315.

[9] Devroye L., Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.* **32** (2002/03), no. 1, 152–171.

[10] Devroye L. and Janson S., Protected nodes and fringe subtrees in some random trees. *Electronic Communications in Probability* **19** (2014), no. 6, 1–10.

[11] Drmota M., *Random Trees*. Springer, Vienna, 2009.

[12] Du R. and Prodinger H., Notes on protected nodes in digital search trees. *Appl. Math. Lett.* **25** (2012), no. 6, 1025–1028.

[13] Feng Q. and Mahmoud H. M., On the variety of shapes on the fringe of a random recursive tree. *J. Appl. Prob.* **47** (2010), no. 01, 191–200.

[14] Fill J.A. and Kapur N., Transfer theorems and asymptotic distributional results for m-ary search trees. *Random Structures Algorithms* **26** (2005), no. 4, 359–391.

[15] Flajolet P., Gourdon X. and Martínez C., Patterns in random binary search trees. *Random Structures Algorithms* **11** (1997), no. 3, 223–244.

[16] Fuchs M., Subtree sizes in recursive trees and binary search trees: Berry–Esseen bounds and Poisson approximations. *Combin. Probab. Comput.* **17** (2008), no. 5, 661–680.

[17] Fuchs M., Limit theorems for subtree size profiles of increasing trees. *Combin. Probab. Comput.* **21** (2012), no. 3, 412–441.

[18] Gopaladesikan M., Mahmoud H. M. and Ward M. D., Asymptotic joint normality of counts of uncorrelated motifs in recursive trees. *Methodology and Computing in Applied Probability* **16** (2014), no. 4, 863–884.

[19] Gut A., *Probability: A Graduate Course*, 2nd ed., Springer, New York, 2013.

[20] Heimbürger A., Asymptotic distribution of two-protected nodes in m-ary search trees. Master thesis, Stockholm University and KTH, 2014. http://www.diva-portal.org/smash/get/diva2:748258/FULLTEXT01.pdf

[21] Holmgren C. and Janson S., Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.* **20** (2015), no. 4, 1–51.

[22] Holmgren C. and Janson S., Asymptotic distribution of two-protected nodes in ternary search trees. *Electron. J. Probab.* **20** (2015), no. 9, 1–20.

[23] Holmgren C. and Janson S., Fringe trees, Crump–Mode–Jagers branching processes and m-ary search trees. *Probability Surveys* **14** (2017) 53–154.

[24] Janson S., Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stoch. Process. Appl.* **110** (2004), 177–245.

[25] Janson S., Asymptotic degree distribution in random recursive trees. *Random Structures Algorithms* **26** (2005), no. 1-2, 69–83.

[26] Janson S., Limit theorems for triangular urn schemes. *Probab. Theory Relat. Fields* **134** (2006), 417–452.

[27] Janson S., Mean and variance of balanced Pólya urns. Preprint, 2016. arXiv:1602.06203.

[28] Janson S. and Pouyanne N., Moment convergence of balanced Pólya processes. Preprint, 2016. arXiv:1606.07022.

[29] Jiřina M., Stochastic branching processes with continuous state space. *Czechoslovak Math. J.* **8 (83)** (1958), 292–313.

[30] Karlin S. and Taylor H. M., *A First Course in Stochastic Processes*, 2nd ed., Academic Press, 1975.

[31] Kalpathy R. and Mahmoud H., Degree profile of m-ary search trees: A vehicle for data structure compression. *Probab. Engrg. Inform. Sci.* **30** (2016), no. 1, 113–123.

[32] Mahmoud H.M., *Evolution of Random Search Trees*. John Wiley & Sons, New York, 1992.

[33] Mahmoud H.M., The size of random bucket trees via urn models. *Acta Inform.* **38** (2002), no. 11-12, 813–838.

[34] Mahmoud H.M., *Pólya Urn Models*. CRC Press, Boca Raton, FL, 2009.

[35] Mahmoud H.M. and Pittel B., Analysis of the space of search trees under the random insertion algorithm. *J. Algorithms* **10**, (1989), no. 1, 52–75.

[36] Mahmoud, H.M. and Smythe, R.T., Asymptotic joint normality of outdegrees of nodes in random recursive trees. *Random Structures Algorithms* **3** (1992), no. 3, 255–266.

[37] Mahmoud, H.M., Smythe, R.T. and Szymański, J., On the structure of random plane-oriented recursive trees and their branches. *Random Structures Algorithms* **4** (1993), no. 2, 151–176.

[38] Mahmoud H.M. and Ward M.D., Asymptotic distribution of two-protected nodes in random binary search trees. *Appl. Math. Lett.* **25** (2012), no. 12, 2218–2222.

[39] Mansour T., Protected points in k-ary trees. *Appl. Math. Lett.* **24** (2011), no. 4, 478–480.

[40] Pouyanne N., An algebraic approach to Pólya processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** (2008), no. 2, 293–323.

[41] Szymański J., On a nonuniform random recursive tree. *Annals of Discrete Math.* **33** (1987), 297–306.

A The covariance matrix Σ in Section 11.2

$$\Sigma = \begin{pmatrix}
\frac{3400704921}{13887473600} & \frac{54821229}{37389935200} & -\frac{55644023}{2991148160} & -\frac{5396373387}{1944246304 00} & -\frac{47473653}{6943736800} & -\frac{19950493}{7776985216} & -\frac{19950493}{7776985216} & -\frac{228203991}{1944246304 00} \\
\frac{54821229}{37389935200} & \frac{7300603}{66040975} & -\frac{8044611}{325124800} & -\frac{21170309 13}{9721231520 0} & -\frac{1469561}{30190160 0} & -\frac{7204037}{42266222 00} & -\frac{7204037}{42266222 00} & -\frac{1334771}{17674966 40} \\
-\frac{55644023}{2991148160} & -\frac{8044611}{325124800} & \frac{118631347}{21133112 00} & -\frac{2729459079}{1944246304 00} & \frac{3532337}{12076064 00} & -\frac{161089}{16256240 0} & -\frac{161089}{16256240 0} & \frac{83883901}{1944246304 00} \\
-\frac{5396373387}{1944246304 00} & -\frac{21170309 13}{9721231520 0} & -\frac{2729459079}{1944246304 00} & \frac{617325}{17359342} & -\frac{7661559}{39678496 00} & -\frac{123349341}{1944246304 00} & -\frac{123349341}{1944246304 00} & -\frac{3764589}{13887473600} \\
-\frac{47473653}{6943736800} & -\frac{1469561}{30190160 0} & \frac{3532337}{12076064 00} & -\frac{7661559}{39678496 00} & \frac{63201}{8625760} & -\frac{3151}{8625760} & -\frac{4847}{41641600} & \frac{193079}{39678496 00} \\
-\frac{19950493}{7776985216} & -\frac{7204037}{42266222 00} & -\frac{161089}{16256240 0} & -\frac{123349341}{1944246304 00} & -\frac{3151}{8625760} & \frac{157523}{72872800} & -\frac{2637}{72872800} & -\frac{2884319}{1944246304 00} \\
-\frac{19950493}{7776985216} & -\frac{7204037}{42266222 00} & -\frac{161089}{16256240 0} & -\frac{123349341}{1944246304 00} & -\frac{4847}{41641600} & -\frac{2637}{72872800} & \frac{157523}{72872800} & -\frac{2884319}{1944246304 00} \\
-\frac{228203991}{1944246304 00} & -\frac{1334771}{17674966 40} & -\frac{83883901}{1944246304 00} & -\frac{3764589}{13887473600} & -\frac{193079}{39678496 00} & -\frac{2884319}{1944246304 00} & -\frac{2884319}{1944246304 00} & \frac{5681341}{69437368 00}
\end{pmatrix}$$

B The covariances in the matrix Σ in Section 11.3

$$\Sigma_{1,1} = \frac{\kappa\,(-14\kappa^5 - 73\kappa^4 + 131\kappa^3 + 1438\kappa^2 + 3018\kappa + 2070)}{(\kappa+1)(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{1,2} = \frac{3\kappa\,(2\kappa^6 + 37\kappa^5 + 124\kappa^4 - 55\kappa^3 - 900\kappa^2 - 1488\kappa - 720)}{4(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{1,3} = \frac{\kappa^2\,(10\kappa^5 - 47\kappa^4 - 664\kappa^3 - 2083\kappa^2 - 2592\kappa - 1080)}{4(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{1,4} = \frac{\kappa\,(10\kappa^5 - 47\kappa^4 - 664\kappa^3 - 2083\kappa^2 - 2592\kappa - 1080)}{4(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{1,5} = -\frac{\kappa\,(20\kappa^6 + 104\kappa^5 - 69\kappa^4 - 1088\kappa^3 - 1307\kappa^2 + 1224\kappa + 2160)}{2(\kappa-1)(\kappa+1)(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{2,2} = \frac{3\kappa\,(17\kappa^6 + 145\kappa^5 + 507\kappa^4 + 929\kappa^3 + 976\kappa^2 + 612\kappa + 180)}{2(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{2,3} = -\frac{\kappa^3\,(80\kappa^4 + 579\kappa^3 + 1558\kappa^2 + 1803\kappa + 720)}{4(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{2,4} = -\frac{\kappa^2\,(80\kappa^4 + 579\kappa^3 + 1558\kappa^2 + 1803\kappa + 720)}{4(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{2,5} = \frac{\kappa\,(28\kappa^7 + 264\kappa^6 + 829\kappa^5 + 816\kappa^4 - 515\kappa^3 - 702\kappa^2 + 1152\kappa + 1080)}{4(\kappa-1)(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{3,3} = \frac{\kappa^2\,(49\kappa^4 + 276\kappa^3 + 539\kappa^2 + 396\kappa + 90)}{2(\kappa+1)^2(\kappa+2)(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{3,4} = -\frac{\kappa^3\,(16\kappa^3 + 87\kappa^2 + 152\kappa + 75)}{2(\kappa+1)^2(\kappa+2)(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{3,5} = -\frac{\kappa^2\,(4\kappa^6 + 64\kappa^5 + 251\kappa^4 + 194\kappa^3 - 729\kappa^2 - 1488\kappa - 720)}{4(\kappa-1)(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{4,4} = \frac{\kappa\,(16\kappa^5 + 120\kappa^4 + 341\kappa^3 + 462\kappa^2 + 321\kappa + 90)}{2(\kappa+1)^2(\kappa+2)(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{4,5} = -\frac{\kappa\,(4\kappa^6 + 64\kappa^5 + 251\kappa^4 + 194\kappa^3 - 729\kappa^2 - 1488\kappa - 720)}{4(\kappa-1)(\kappa+1)^2(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)};$$
$$\Sigma_{5,5} = \frac{\kappa\,(-4\kappa^7 - 4\kappa^6 + 147\kappa^5 + 574\kappa^4 + 610\kappa^3 - 51\kappa^2 + 132\kappa + 630)}{(\kappa-1)^2(\kappa+1)(\kappa+2)^2(\kappa+3)^2(2\kappa+1)(2\kappa+3)(2\kappa+5)}.$$