Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Using Pólya urns to show normal limit laws for fringe subtrees in m-ary search trees

Cecilia Holmgren [1] [†], Svante Janson [1] [‡], Matas Šileikis[2]

[1] *Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden*
[2] *Department of Applied Mathematics of Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 118 00 Praha, Czech Republic*

We study fringe subtrees of random m-ary search trees, by putting them in the context of generalised Pólya urns. In particular we show that for the random m-ary search tree with $m \leq 26$, 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.

Keywords: Random trees; Fringe trees; Normal limit laws; Pólya urns; m-ary search trees

1 Introduction

The main focus of this paper is to consider fringe subtrees of random m-ary search trees; 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 fringe subtrees typically are "small" compared to the whole tree.

We will use (generalised) Pólya urns to analyze vectors of the numbers of fringe subtrees of different types in random m-ary search trees. As a result, we prove multivariate normal asymptotic distributions for these random variables, for m-ary search trees when $m \leq 26$. (It is well known that asymptotic normality does not hold for m-ary search trees for $m > 26$, see [2].)

Pólya urns have earlier been used to study the total number of nodes in random m-ary search trees, see [16, 13, 17]. In that case one only needs to consider an urn with $m-1$ different types, describing the nodes holding i keys, where $i \in \{0, 1, \dots, m-2\}$. Recently, in [10] more advanced Pólya urns were used to describe protected nodes in random m-ary search trees, where the types were further divided depending on characteristics of the different fringe subtrees (however, in [10] only the cases $m = 2, 3$ were treated in detail).

In [10] a simpler urn 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 [10] for analyzing arbitrary fringe subtrees of a fixed size in random m-ary search trees. This paper is an extended abstract of [12], where we also prove similar results for the general class of linear preferential attachment trees, and also extend the methods used in [10] to analyze the number of protected nodes in m-ary search trees for $m \leq 26$.

2 m-ary search trees

We recall the definition of m-ary search trees, see e.g. [15] or [6]. An m-ary search tree, where $m \geq 2$, is constructed recursively from a sequence of n *keys* (ordered 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

[†] Partly supported by the Swedish Research Council
[‡] Partly supported by by the Knut and Alice Wallenberg Foundation

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 i'th child. (The binary search tree, i.e., the case $m = 2$, is the simplest case.) 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 matters, so alternatively, we may assume that the keys are n i.i.d. uniform random numbers in $[0, 1]$.)

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 but are not really part of the tree. Thus, a *leaf* is an internal node that has no internal children, but it may have external children.

We say that a node with $i \le m-2$ 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 $\pm\infty$), and a new key has the same probability $1/(n+1)$ of being inserted into any of the gaps. Thus, the evolution of the random m-ary search tree may be described by choosing a gap uniformly at random at each step, and inserting a new key there.

Note that the construction above yields the m-ary search tree as an ordered tree. Hence, a nonrandom m-ary search tree is an 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 nodes with $m-1$ keys have exactly m children, and the other nodes are leaves. There is a natural partial order on the set of (isomorphism classes of) nonrandom m-ary search trees, such that $T \preceq T'$ if T' can be obtained from T by adding keys (including the case $T' = T$).

In applications where the order of the children of a node does not matter, we can simplify by ignoring the order and regard the m-ary search tree as an unordered tree. (Then, we can also ignore the external nodes.) A partial order $T \preceq T'$ is defined on the set of (isomorphism classes of) unordered m-ary search trees in the same way as in the ordered case.

3 Main results

In this section we state the main results on fringe subtrees in random m-ary search trees. These results are extensions of results that previously have been shown for the specific case of the random binary search tree with the use of other methods, see e.g., [4, 5, 9].

Remark 3.1 As said in the introduction, m-ary search trees can be regarded as either ordered or unordered trees. The most natural interpretation is perhaps the one as ordered trees, and it implies the corresponding result for unordered trees in, e.g., 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.

The following theorem generalises [9, Theorem 1.22], where the specific case of the binary search tree was analyzed.

Let $H_m := \sum_{k=1}^{m} 1/k$ be the m'th harmonic number.

Theorem 3.2 *Assume that $2 \le m \le 26$. Let $T^1, \dots, T^d$ be a fixed sequence of nonrandom m-ary search trees and let $\mathbf{Y}_n = \left(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\right)$, 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*

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

Then

$$n^{-1/2}(\mathbf{Y}_n - \boldsymbol{\mu}_n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{3.1}$$

where $\Sigma = (\sigma_{ij})_{i,j=1}^d$ is some covariance matrix. Furthermore, in (3.1), the vector $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n := n\hat{\boldsymbol{\mu}}$, with

$$\hat{\boldsymbol{\mu}} := \Big(\frac{\mathbb{P}(\mathcal{T}_{k_1} = T^1)}{(H_m-1)(k_1+1)(k_1+2)}, \dots, \frac{\mathbb{P}(\mathcal{T}_{k_d} = T^d)}{(H_m-1)(k_d+1)(k_d+2)}\Big). \tag{3.2}$$

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 That $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n$ means that

$$\mathbb{E}\big(X_n^{T^i}\big) = \frac{\mathbb{P}(\mathcal{T}_{k_i} = T^i)}{(H_m-1)(k_i+1)(k_i+2)} n + o\big(n^{1/2}\big). \tag{3.3}$$

A weaker version of (3.3) with the error term $o(n)$ follows from the branching process analysis of fringe subtrees in [11], see the proof in Section 6. The vector $\hat{\boldsymbol{\mu}}_n$ can also, using (5.2) 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 [14].

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). The results in [14] show also

$$\sigma_{ij} = \lim_{n\to\infty} \frac{1}{n} \operatorname{Cov}\big(X_n^{T^i}, X_n^{T^j}\big). \tag{3.4}$$

The following theorem is an important corollary of Theorem 3.2. It also follows from Fill and Kapur [7, Theorem 5.1]. The special case of the random binary search tree was proved by Devroye [4], and the covariances for $Y_{n,k}$ in that case were given by Dennert and Grübel [3], see also [9, Theorem 1.19 and Proposition 1.10]. For binary search trees also the case when the size k is depending on n has been analyzed; in that case both normal and Poisson limit laws appear, see e.g., Fuchs [8] and [9].

Theorem 3.4 *Assume that $2 \le m \le 26$. Let k 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 $n \to \infty$,*

$$n^{-1/2}\big(Y_{n,k} - \mathbb{E}\, Y_{n,k}\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{3.5}$$

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

$$n^{-1/2}\Big(Y_{n,k} - \frac{n}{(H_m-1)(k+1)(k+2)}\Big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2). \tag{3.6}$$

Remark 3.5 The asymptotic mean $\frac{n}{(H_m-1)(k+1)(k+2)}$ in (3.6) easily follows from (3.3). The constant σ_k^2 can again be calculated explicitly from our proof.

We give one example of Theorem 3.4 in Section 7, where we let $m = 3$ and $k = 4$.

4 Generalised Pólya urns

A (generalised) Pólya urn process is defined as follows, see e.g. [13] or [17]. 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) $a_i \in \mathbb{R}_{\ge 0}$, and a random vector $\xi_i = (\xi_{i1}, \dots, \xi_{iq})$, where $\xi_{ij} \in \mathbb{Z}_{\ge 0}$ and $\xi_{ii} \in \mathbb{Z}_{\ge -1}$. The urn evolves according to a discrete time Markov process. At each time $n \ge 1$, 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.

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

$$A := (a_j \mathbb{E}\, \xi_{ji})_{i,j=1}^q. \tag{4.1}$$

The intensity matrix A with its eigenvalues and eigenvectors is central for proving limit theorems.

We use the basic assumptions (A1)–(A6) on the Pólya urn stated in [13, p. 180] together with the following simplifying assumption, cf. [10]:

(A7) At each time $n \geq 1$, there exists a ball of a dominating type, as defined in [13].

Using the Perron–Frobenius theorem, it is easy to verify all conditions (A1)–(A6) for the Pólya urns used in this paper, and (A7) follows because the urn is irreducible if we ignore balls with activity 0, and there will always be a ball of positive activity, see [13, Lemma 2.1] and the discussion in [12].

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 $u'v$. We let λ_1 denote the largest real eigenvalue of A. (This exists by our assumptions and the Perron–Frobenius theorem.) 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 largest eigenvalue λ_1, i.e., vectors satisfying

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

We assume that v_1 and u_1 are normalised so that

$$a \cdot v_1 = a' v_1 = v_1' a = 1, \qquad\qquad u_1 \cdot v_1 = u_1' v_1 = v_1' u_1 = 1, \tag{4.2}$$

see [13, 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. We define the matrices

$$B_i := \mathbb{E}(\xi_i \xi_i') \tag{4.3}$$

$$B := \sum_{i=1}^q v_{1i} a_i B_i \tag{4.4}$$

$$\Sigma_I := \int_0^\infty P_I e^{sA} B e^{sA'} P_I' e^{-\lambda_1 s} ds, \tag{4.5}$$

where we recall that $e^{tA} = \sum_{j=0}^\infty t^j A^j / j!$. From [13] it follows that when $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$, the integral Σ_I in (4.5) converges.

Furthermore, it is proved in [13] that, under assumptions (A1)–(A7), $\mathcal{X}_n$ is asymptotically normal if $\operatorname{Re}\lambda \leq \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$. We will apply the following result from [13].

Theorem 4.1 ([13, Theorem 3.22 and Lemma 5.4]) *Assume* (A1)–(A7) *and that we have normalised as in* (4.2)*. Also assume that* $\operatorname{Re}\lambda < \lambda_1/2$*, for each eigenvalue* $\lambda \neq \lambda_1$*.*

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

$$n^{-1/2}(\mathcal{X}_n - n\mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{4.6}$$

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

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

$$a \cdot \mathbb{E}(\xi_i) = c, \qquad i = 1, \dots, q. \tag{4.7}$$

Then the covariance matrix in (4.6) *is given by* $\Sigma = c\Sigma_I$*, with* Σ_I *as in* (4.5)*.*

□

Remark 4.2 It is easily seen that (4.7) implies that $\lambda_1 = c$ and $u_1 = a$, see e.g. [13, Lemma 5.4]. There is also an alternative way to evaluate Σ in the case when A is diagonalisable (which is the case at least for many examples of Theorem 3.2 and Theorem 3.4, e.g., the example in Section 7), see [12, Theorem 4.1(iii)] or [13, Lemma 5.3].

Remark 4.3 From (4.6) follows immediately a weak law of large numbers:

$$\mathcal{X}_n/n \xrightarrow{\mathrm{p}} \mu. \tag{4.8}$$

In fact, the corresponding strong law $\mathcal{X}_n/n \xrightarrow{\mathrm{a.s.}} \mu$ holds as well, see [13, Theorem 3.21]. Furthermore, in all applications in the present paper, all ξ_{ij} are bounded and thus each $X_{n,i} \le Cn$ for some deterministic constant; hence (4.8) implies by dominated convergence that also the means converge:

$$\mathbb{E}\,\mathcal{X}_n/n \to \mu. \tag{4.9}$$

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 3.4 for m-ary search trees. We consider either ordered or unordered trees, see Remark 3.1.

Let $T^1, \dots, T^d$ be a fixed sequence of (nonrandom) m-ary search trees and let as in Theorem 3.2 $\mathbf{Y}_n = (X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d})$, 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.)

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 that at least one tree T^i contains at least $m-2$ keys). 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) nonrandom 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

$$\mathcal{S} := \bigcup_{i=1}^{d} \{T : T \preceq T^i\}, \tag{5.1}$$

and the set of all types by $\mathcal{S}^* := \mathcal{S} \cup \{*\}$, where $*$ is the dead type.

When a key is added to the tree $\mathcal{T}_n$, it is added to a leaf with at most $m-2$ keys or 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

$$X_n^{T^i} = \sum_{T\in\mathcal{S}} n(T,T^i) X_{n,T}. \tag{5.2}$$

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 (5.2).

Example 5.1 (a Pólya urn to count fringe subtrees with 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 3.4. In this case, the order of children in the tree does not matter so it is easier to regard the trees as unordered.

Thus, fix $k \ge m-2$ and let T^i, $i \in \{1,\dots,d\}$, be the sequence of all m-ary search trees that can be obtained with at most k keys. Hence, (5.1) yields $\mathcal{S} = \{T^i : 1 \le i \le d\}$.

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, which we ignore.

The replacement rules in the Pólya urn are easy to describe. The types are the m-ary search trees with at most k keys. A type T with j keys has $j+1$ gaps, and is thus given activity $j+1$. Let $T_1,\dots,T_{j+1}$ be the trees obtained by adding a key to one of these gaps in T. (Some of these may be equal.) If we draw a ball of type T and $j<k$, then the drawn ball is replaced by one ball of a type randomly chosen among $T_1,\dots,T_{j+1}$ (with probability $1/(j+1)$ each); note that these trees have $j+1 \le k$ keys and are themselves types in the urns. On the other hand, if $j=k$, then each of these trees has $k+1$ keys so its root is dead; the root contains $m-1$ keys so after removing it we are left with m subtrees with together $k+1-(m-1) \le k$ keys, so 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; this leaves m living subtrees and we add balls of the corresponding types.

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 we similarly, using (5.2), may obtain the number of fringe subtrees with ℓ keys, for any $\ell \le k$, from the same urn. This enables us to obtain joint convergence in Theorem 3.4 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 [16, 13, 17] to study the number of nodes, and the numbers of nodes with different numbers of keys, in an m-ary search tree.

In Section 7 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, (4.7) holds with $c=1$. Hence, $\lambda_1 = 1$ by Remark 4.2.

6 Proofs

As said in Section 4, it is easy to see that the Pólya urns constructed in Section 5 satisfy (A1)–(A7), for example with the help of [13, Lemma 2.1]. To apply Theorem 4.1 it remains to show that $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \ne \lambda_1$. 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 version with unordered trees is 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 $k \le q$, let

$$\mathcal{S}_k := \{T_1, \dots, T_k\}. \tag{6.1}$$

For $k \ge m-1$, 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}. In the result below we assume $m \ge 3$, the case $m = 2$ is similar and we refer to [12] for the corresponding statement and proof in that case.

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

$$\operatorname{tr}(A_k) = -\sum_{i=1}^{k} a_i. \tag{6.2}$$

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 $m \ge 3$, 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 (4.1). □

Theorem 6.2 *Let $m \ge 2$. 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*

$$\{-a_i : i = m, m+1, \dots, q\}. \tag{6.3}$$

Proof: We prove by induction on k that the theorem holds for A_k (with q replaced by k in (6.3)), for any k with $m-1 \le k \le q$. 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., [13, Example 7.8] and [17, 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.,

$$A_{m-1} = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 & m(m-1) \\ 1 & -2 & 0 & \dots & 0 & 0 \\ 0 & 2 & -3 & \dots & 0 & 0 \\ 0 & 0 & 3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & m-2 & -(m-1) \end{pmatrix}. \tag{6.4}$$

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 (6.3) is empty in this case.

We proceed to the induction step. Let $m-1 \le k < q$. By using arguments similar to those that were used in the proof of [10, 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 for the Pólya urn with types in $\mathcal{S}_k$.

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

$$\mathbb{E}(\mathcal{X}_1^{k+1} - \mathcal{X}_0^{k+1}) = \frac{A_{k+1}\mathcal{X}_0^{k+1}}{a^{k+1} \cdot \mathcal{X}_0^{k+1}}, \tag{6.5}$$

$$\mathbb{E}(\mathcal{X}_1^{k} - \mathcal{X}_0^{k}) = \frac{A_{k}\mathcal{X}_0^{k}}{a^{k} \cdot \mathcal{X}_0^{k}}. \tag{6.6}$$

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,

$$TA_{k+1} = A_k T. \tag{6.7}$$

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

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

Then, by (6.7),

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

and thus $u'T = (T'u)'$ is a left generalised eigenvector of A_{k+1} for the eigenvalue λ. Since T is onto, 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,

$$\operatorname{tr} A_{k+1} = \lambda_1 + \dots + \lambda_{k+1} = \operatorname{tr} A_k + \lambda_{k+1} \tag{6.8}$$

and thus by (6.2) (when $m > 2$) or the corresponding result in [12] (when $m = 2$),

$$\lambda_{k+1} = \operatorname{tr}(A_{k+1}) - \operatorname{tr}(A_k) = -a_{k+1}. \tag{6.9}$$

Thus, by induction, Theorem 6.2 holds for every A_k, with $m-1 \le k \le q$, 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); it is well-known that this holds if $m \le 26$, but not for larger m, see [18] and [7].

In the remainder of this section we assume $m \le 26$. Thus $\operatorname{Re}\lambda < \lambda_1/2$ for every eigenvalue $\lambda \ne \lambda_1$, and Theorem 4.1 applies to the urn defined above.

Proof of Theorem 3.2: By Theorem 4.1(i), (4.6) holds, with $\mu = \lambda_1 v_1 = v_1$.

By (5.2), $\mathbf{Y}_n = \bigl(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\bigr) = R\mathcal{X}_n$ for some (explicit) linear operator R. Hence, (4.6) implies

$$n^{-1/2}\bigl(\mathbf{Y}_n - nR\mu\bigr) = R\bigl(n^{-1/2}(\mathcal{X}_n - \mu)\bigr) \xrightarrow{\mathrm{d}} \mathcal{N}\bigl(0, R\Sigma R'\bigr). \tag{6.10}$$

Furthermore, by [14],

$$\mathbb{E}\,\mathcal{X}_n = n\mu + o\bigl(n^{1/2}\bigr), \tag{6.11}$$

and thus

$$\boldsymbol{\mu}_n = \mathbb{E}\,\mathbf{Y}_n = R\bigl(\mathbb{E}\,\mathcal{X}_n\bigr) = nR\mu + o\bigl(n^{1/2}\bigr). \tag{6.12}$$

Hence, (6.10) implies (3.1) (with the covariance matrix $R\Sigma R'$, where Σ is as in (4.6)).

Moreover, as said in Remark 3.3, it follows from [11], to be precise by combining [11, (5.30), Theorem 7.10 and Theorem 7.11], that

$$\mathbb{E}\,\mathbf{Y}_n = n\hat{\mu} + o(n). \tag{6.13}$$

By combining (6.12) and (6.13) we see that $R\mu = \hat{\mu}$ (since neither depends on n), and thus (6.12) yields (3.3).

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 [14, 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 [9, 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 $1 \le k_1 \le k_2 \le \dots$. 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, $i \ge 2$, 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 3.4: This follows from Theorem 3.2; we refer to [12] for details. □

7 Example of Theorem 3.4 when $m = 3$ and $k = 4$

We consider the case when we want to evaluate σ_4^2 in Theorem 3.4 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 childen of the leaf.

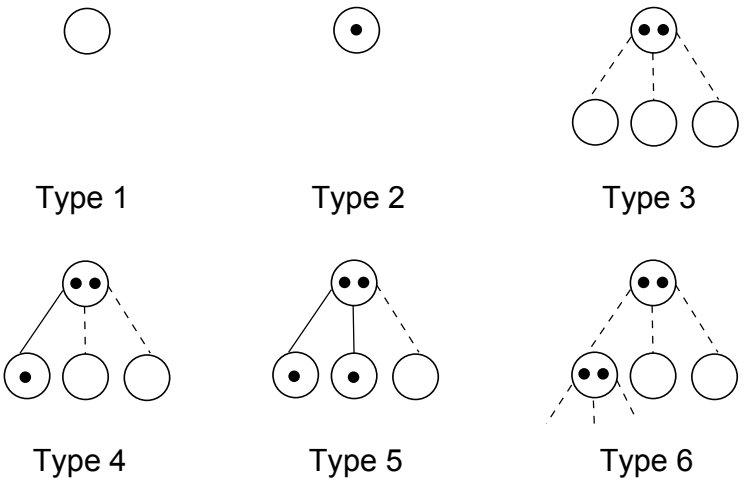

Fig. 1: The different types for counting the number of the fringe subtrees with four keys in a ternary search tree.

See Figure 1 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 $k \le 3$ 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 each of the types 1, 2 and 3 with probability 4/5; see Figure 2 for an 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.

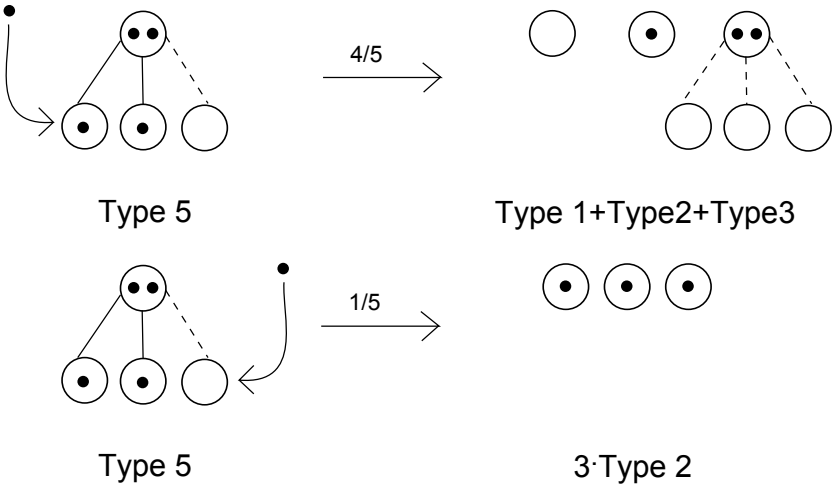

Fig. 2: The two possibilities for adding a key to a node in a tree of type 5 of a ternary search tree.

Thus, we get the intensity matrix A in (4.1) as

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 4 & 8 \\ 1 & -2 & 0 & 0 & 7 & 2 \\ 0 & 2 & -3 & 0 & 4 & 2 \\ 0 & 0 & 3 & -4 & 0 & 3 \\ 0 & 0 & 0 & 2 & -5 & 0 \\ 0 & 0 & 0 & 2 & 0 & -5 \end{pmatrix}. \tag{7.1}$$

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

$$1, -3, -4, -4, -5, -5. \tag{7.2}$$

(We know already that $\lambda_1 = 1$, as was noted in Section 6 as a consequence of Remark 4.2.)

Furthermore, by Remark 4.2, the left eigenvector u_1 is $u_1 = a = (1, 2, 3, 4, 5, 5)$. The right eigenvector v_1, with the normalization (4.2), is $v_1 = (3/25, 1/10, 2/25, 3/50, 1/50, 1/50)$. Note that $\mu = v_1$ in Theorem 4.1, since $\lambda_1 = 1$. Hence, the asymptotic mean in (3.6) for $X_{n,4}$ is $(\mu_5 + \mu_6)n = \frac{n}{25}$. (However, to get the asymptotic expectation in (3.6) for arbitrary k and m we could instead use branching processes, see [11].)

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 in (4.3)–(4.5). Since A is diagonalisable, there is also an alternative way to calculate Σ, see Remark 4.2; see also [12, Theorem 4.1(iii)] and [13, Lemma 5.3].

To calculate B in (4.4) we need to calculate $B_i = \mathbb{E}(\xi_i \xi_i')$ in (4.3). We only describe how to get the matrix B_5 since the other cases are analogous. We get $B_5 = \frac{1}{5} \cdot b_1 b_1' + \frac{4}{5} \cdot b_2 b_2'$, where $b_1 = (0, 3, 0, 0, -1, 0)'$ and $b_2 = (1, 1, 1, 0, -1, 0)'$; see Figure 2. Now we can use Mathematica to evaluate the integral in (4.5), which yields Σ_I. Finally, $\Sigma = \Sigma_I$ by Theorem 4.1 with $c = 1$. The result is given in (7.3).

$$\Sigma = \begin{pmatrix}
\frac{29017}{259875} & -\frac{117371}{10395000} & -\frac{44311}{5197500} & -\frac{2143}{945000} & -\frac{28289}{5197500} & -\frac{28289}{5197500} \\
-\frac{117371}{10395000} & \frac{7379}{83160} & -\frac{34927}{5197500} & -\frac{3907}{236250} & -\frac{166037}{20790000} & -\frac{166037}{20790000} \\
-\frac{44311}{5197500} & -\frac{34927}{5197500} & \frac{159241}{2598750} & -\frac{4747}{236250} & -\frac{84709}{10395000} & -\frac{84709}{10395000} \\
-\frac{2143}{945000} & -\frac{3907}{236250} & -\frac{4747}{236250} & \frac{39227}{945000} & -\frac{13309}{1890000} & -\frac{13309}{1890000} \\
-\frac{28289}{5197500} & -\frac{166037}{20790000} & -\frac{84709}{10395000} & -\frac{13309}{1890000} & \frac{22613}{1299375} & -\frac{6749}{2598750} \\
-\frac{28289}{5197500} & -\frac{166037}{20790000} & -\frac{84709}{10395000} & -\frac{13309}{1890000} & -\frac{6749}{2598750} & \frac{22613}{1299375}
\end{pmatrix}. \tag{7.3}$$

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

$$\Delta = \begin{pmatrix} \sigma_{5,5} & \sigma_{5,6} \\ \sigma_{6,5} & \sigma_{6,6} \end{pmatrix} = \begin{pmatrix} \frac{22613}{1299375} & -\frac{6749}{2598750} \\ -\frac{6749}{2598750} & \frac{22613}{1299375} \end{pmatrix}. \tag{7.4}$$

Summing the $\sigma_{i,j}$ in (7.4), 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 total number of leaves in the random ternary search tree, which was evaluated in [10, Theorem 4.1]. We get

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

We could also use this urn to evaluate

$$\sigma_3^2 = (0, 0, 0, 1, 0, 0)\Sigma(0, 0, 0, 1, 0, 0)' = \frac{39227}{945000}, \tag{7.5}$$

$$\sigma_2^2 = (0, 0, 1, 0, 0, 1)\Sigma(0, 0, 1, 0, 0, 1)' = \frac{131}{2100}, \tag{7.6}$$

$$\sigma_1^2 = (0, 1, 0, 1, 2, 0)\Sigma(0, 1, 0, 1, 2, 0)' = \frac{8}{75}. \tag{7.7}$$

References

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

[2] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] Holmgren, C. and Janson S., Fringe trees, Crump–Mode–Jagers branching processes and m-ary search trees. Preprint, 2016. `arXiv:1601.03691`.

[12] Holmgren, C., Janson S and M. Šileikis., Multivariate normal limit laws for the numbers of fringe subtrees in m-ary search trees and preferential attachment trees. Preprint, 2016. `arXiv:1603.08125`

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

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

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

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

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

[18] 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.