ALEA, Lat. Am. J. Probab. Math. Stat. 13, 605–634 (2016)

B-urns

B. Chauvin, D. Gardy, N. Pouyanne and D.-H. Ton-That

Laboratoire de Mathématiques de Versailles,
UVSQ, CNRS, Université Paris-Saclay
78035 Versailles, France.
E-mail address: brigitte.chauvin@uvsq.fr

Laboratoire DAVID,
UVSQ, Université Paris-Saclay
78035 Versailles, France.
E-mail address: daniele.gardy@uvsq.fr

Laboratoire de Mathématiques de Versailles,
UVSQ, CNRS, Université Paris-Saclay
78035 Versailles, France.
E-mail address: nicolas.pouyanne@uvsq.fr

Laboratoire DAVID,
UVSQ, Université Paris-Saclay
78035 Versailles, France.
E-mail address: tonthatdaihai@gmail.com

Abstract. The fringe of a B-tree with parameter m is considered as a particular Pólya urn with m colors. More precisely, the asymptotic behaviour of this fringe, when the number of stored keys tends to infinity, is studied through the composition vector of the fringe nodes. We establish its typical behaviour together with the fluctuations around it. The well known phase transition in Pólya urns has the following effect on B-trees: for $m \leq 59$, the fluctuations are asymptotically Gaussian, though for $m \geq 60$, the composition vector is oscillating; after scaling, the fluctuations of such an urn strongly converge to a random variable W. This limit is $\mathbb{C}$-valued and it does not seem to follow any classical law. Several properties of W are shown: existence of exponential moments, characterization of its distribution as the solution of a smoothing equation, existence of a density relatively to the Lebesgue measure on $\mathbb{C}$, support of W. Moreover, a few representations of the composition vector for various values of m illustrate the different kinds of convergence.

Received by the editors July 29, 2015; accepted June 28, 2016.
2010 *Mathematics Subject Classification.* 60J80, 68W40, 68Q87.
Key words and phrases. B-tree, fringe analysis, Pólya urn, urn model, martingale, multitype branching process, smoothing transforms, contraction method.

1. Introduction

B-trees are a fundamental structure in computer science, they have been introduced in the early seventies by Bayer (1971); Bayer and McCreight (1972), to store large quantities of data. These particular search trees are conceived in order to have all their leaves at the same level. The nodes at the deepest level are called the *fringe nodes*. A precise description can be found in Section 2 where are presented two classical algorithms giving a B-tree. The actual writing of these algorithms can be found for example in Cormen et al. (2001) for one of them (the so-called *prudent* algorithm in the sequel), in Kruse and Ryba (1999) for the other one (called the *optimistic* algorithm in the sequel).

The fringe analysis of B-trees goes back to Yao (1977/78) and has been developped by many authors (see for example the survey Baeza-Yates, 1995), both for B-trees and B^+-trees (where all the keys are stored in the fringe nodes). In Yao (1977/78) appears the Pólya urn model, which we develop in this article. Indeed, the fringe of a B-tree with parameter m (where m is a positive integer) can be considered as a particular Pólya urn with m colors, so that a lot of information can be obtained concerning the asymptotic behaviour of this fringe, when the number of stored keys tends to infinity. Besides, it turns out that the same Pólya urn mechanism occurs for the insertion into the leaves of a paged binary search tree. Paged binary search trees are a variation of binary search trees where a subtree of size less than a given capacity is replaced by a bucket. Mahmoud (2002) has been the first to consider the evolution of a random bucket tree as an urn model.

Let us describe a Pólya urn process as follows. Consider an urn that contains balls of, say, d different colors. Start with a finite number of different color balls as initial composition (possibly monochromatic). At each discrete time n, draw a ball at random, check its color, put it back into the urn and add balls according to the following rule: if the drawn ball is of color i, add $a_{i,j}$ balls of color j, where the $a_{i,j}$ are integer-valued. Thus, the replacement rule is described by the so-called *replacement matrix*, which is a dimension d matrix, whose coefficients are the $a_{i,j}$, for i and j in $\{1,\dots,d\}$.

Usually, the integers $a_{i,j}$ are assumed to be nonnegative for $i \neq j$ and the integers $a_{i,i}$ are nonpositive or nonnegative. A negative coefficient $a_{i,i}$ means that, if a ball of color i is drawn, then $a_{i,i}$ balls of color i are removed from the urn. In this case, we have to ensure that at least $a_{i,i}$ balls of color i exist in the urn. This quality is called the *tenability* of the urn. To ensure that an urn with a negative coefficient $a_{i,i}$ is tenable, it is necessary and sufficient to have the following arithmetical condition (this can be easily proved by induction on n). Fix an initial composition $(\alpha_1,\dots,\alpha_d)$, meaning that there are α_j balls of colour j at time zero in the urn, then the tenability condition can be written as

$$-a_{i,i} \text{ divides } \alpha_i, a_{1,i}, \dots, a_{d,i}. \tag{1.1}$$

Moreover, in the present paper, the urn is assumed to be *balanced*, which means that the total number of balls added at each step is a constant: there exists an integer S such that, for any i in $\{1,\dots,d\}$, $\sum_{j=1}^{d} a_{i,j} = S$.

Let us emphasize that "drawing a ball at random" means choosing *uniformly* among the balls contained in the urn. That is why this model is related to many situations in mathematics, algorithmics or theoretical physics where a uniform choice

among objects determines the evolution of a process. See Johnson and Kotz (1977); Mahmoud (2009); Flajolet et al. (2006) for many examples. For a general probabilistic treatment of Pólya urns, see Pouyanne (2008), Janson (2004) or Mailler (2014).

In the paper Yao (1977/78), the focus is on the average number of nodes in the B-tree. Nevertheless, the main ideas are already there, namely the dynamics transforming a tree of size n into a tree of size $n+1$, which is the same dynamics as in a Pólya urn process. The recent progresses in Pólya urn processes and their asymptotic behaviour (Pouyanne, 2008; Chauvin et al., 2011, 2015, Janson, 2004, Mailler, 2014) lead to a more complete landscape for the B-trees. Our aim in this article is to present in a hopefully concise form a collection of results about the asymptotic behaviour of the fringe nodes in a B-tree, namely their typical behaviour and the fluctuations around it. Our main interest is focused on these fluctuations, which happen to have a phase transition: for $m \leq 59$, the fluctuations are of order $\sqrt{n}$ and have a Gaussian limit in distribution. But for $m \geq 60$, the fluctuations are of order n^{σ}, where σ is larger than $1/2$ and increases to 1 when m tends to infinity. Moreover, an oscillating and significative phenomenon occurs in the fluctuation term. After scaling, the fluctuations strongly converge (meaning almost surely) to a random limit, here called W. The random variable W is $\mathbb{C}$-valued and does not seem to follow any classical law.

The paper is organized as follows. In Section 2 are presented two classical algorithms allowing to construct a B-tree. In that section is also precised how the insertion dynamics is that of a suitable Pólya urn. In Section 3 are introduced the random vectors which describe the fringe of a B-tree. In Section 4 is established the phase transition, and we get the precise asymptotic behaviour of the fringe nodes, in Corollary 4.4 of Theorem 4.3. For $m \geq 60$, the fluctuations around the drift are expressed via a random variable W, which is studied in the last sections. Thanks to an embedding into continuous time (Section 5), a multitype branching process is put forward. Properties of the continuous-time limit process can be translated to the discrete-time process, via an explicit connection. Several properties of W are proved in Section 6: W admits a density on the whole complex plane; it has exponential moments; it is the unique solution of a certain "smoothing equation" in a convenient probability distribution space. Finally, in Section 7, a few pictures provide a synthetic and concrete illustration of the different kinds of convergence, depending on whether $m \leq 59$ or $m \geq 60$.

2. B-tree algorithms

2.1. *Description of a B-tree.* For a positive integer $m \geq 2$, a B-tree with parameter m is a search[1] tree, where the keys are stored into the internal nodes and the leaves[2] represent insertion possibilities (we call them *gaps*), they do not contain any key; furthermore all the leaves are at the same depth. A *fringe node* is an internal node whose only descendants are leaves. In the literature, these fringe nodes are sometimes called *final internal nodes* or *leaf-nodes*, or *internal leaves*. We try to

[1] A search tree is a tree where internal nodes contain sorted keys and where a node containing k keys $x_1, \ldots, x_k$ defines $k+1$ intervals such that, for $j = 1, \ldots, k+1$, the keys in the j-th subtree belong to the j-th interval.

[2] The leaves, sometimes called external nodes, are the nodes without any descendant.

be non-ambiguous in the following, and use the terms fringe nodes and fringe node process. In the figures below, internal nodes are represented by ellipses and leaves by squares.

As is the case for the leaves, the fringe nodes of a B-tree are at the same depth. Moreover, each internal node (fringe or otherwise) has a capacity; the root contains between 1 and $C(m)$ keys, and the other internal nodes between $c(m)$ and $C(m)$ keys. When a node contains $C(m)$ keys, we say that the node is *saturated*.

The minimal – $c(m)$ – and maximal – $C(m)$ – values depend both on the parameter m and on the precise definition of the B-tree, which is itself closely related to the exact insertion algorithm, of which we present two versions below. Let us just state that $c(m) = m - 1$ and $C(m) = 2m - 1$ for the first algorithm, and $c(m) = m$ and $C(m) = 2m$ for the second one. In both cases we want to insert a new key into a tree of size n, i.e. having already n keys in its internal nodes, and consequently $n + 1$ leaves, or insertion possibilities.

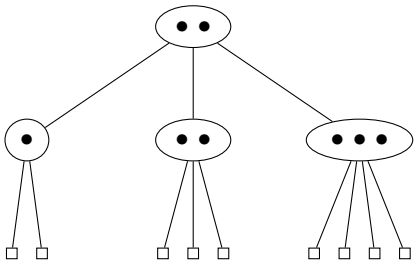

FIGURE 2.1. *An example of B-tree of size* 8. *Here* $m = 2$, *nodes contain between* 1 *and* 3 *keys. There are* 3 *fringe nodes and* 9 *leaves. One fringe node is saturated.*

2.2. *The prudent algorithm.* In what we call here the *prudent algorithm* for insertion into a B-tree with parameter m, the nodes contain between $m - 1$ and $2m - 1$ keys. An insertion of a new key concerns a given leaf, so that a branch (the nodes between the root and this leaf) is determined for this insertion. The algorithm proceeds by *going down* from the root to the leaf, along this branch. We begin by checking the root: if it is saturated, it is split, a new root is created with a single key which is the median of the keys of the old root (remember that it has an odd number of keys, hence the median is defined without any ambiguity) and two sons, and the height of the tree increases by 1. If the root is not saturated, we do not modify it. We then proceed along the branch to the insertion gap. When we meet a (non-root) saturated node, the median key of that node moves to the parent node (which is not saturated – if it initially was, we have already taken care of it) and the saturated node is split. Then, when we finally arrive at a fringe node, we split it when necessary, and the insertion of the new key always takes place into a non-saturated fringe node: the saturated nodes are dealt with *before* we find the node in which the insertion of the new key will take place. This algorithm, which can be presented both recursively and iteratively (there being only a descent from the root to a leaf), is found, e.g., in the book Cormen et al. (2001). If we consider the fringe nodes, insertion on a saturated node (with $2m - 1$ keys) gives rise to 2 new fringe nodes with respectively m and $m - 1$ keys. See Figure 2.2.

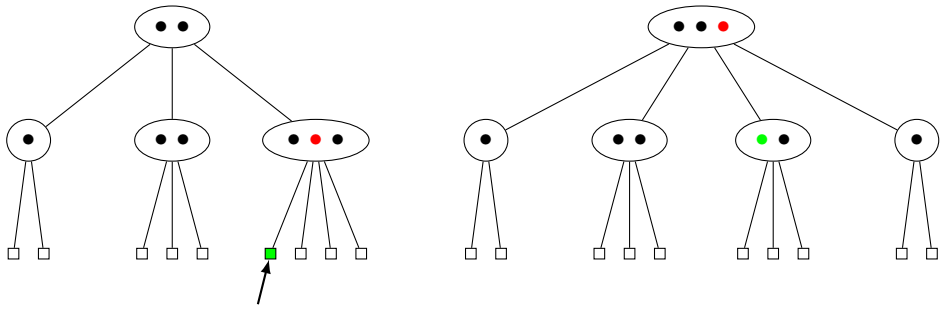

FIGURE 2.2. *An example of insertion in a B-tree, for the prudent algorithm. Here $m = 2$, nodes contain between 1 and 3 keys. The middle key in red moves to the parent node.*

2.3. *The optimistic algorithm.* In what we call here the *optimistic algorithm*, for insertion into a B-tree with parameter $m-1$, the nodes contain between $m-1$ and $2m-2$ keys. Here the saturated nodes are dealt with *after* we have found the place for insertion. An insertion of a new key concerns a given leaf. If the corresponding fringe node is not saturated, the insertion occurs in this node; if it is saturated, the algorithm has to create a non-saturated node into which we can insert the new key. It needs to find what would be the place of the new key among the (already sorted) $2m-2$ keys; the middle key among these $2m-2+1 = 2m-1$ keys moves to the parent node, and the saturated node is split into 2 new fringe nodes with $m-1$ keys. If the parent node is saturated, a key is pushed up into the grandparent node, etc... all the way up to the root if necessary; if the root is saturated, it is split as well and the height of the tree increases by 1. This algorithm proceeds by going down from the root to the gap of insertion, and then up to (some node on) the branch from that leaf to the root, and is possibly best understood recursively. Figure 2.3 illustrates an insertion on a saturated node for a B-tree with parameter $m = 3$.

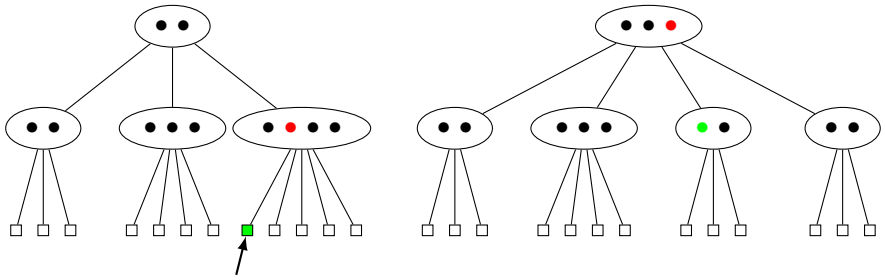

FIGURE 2.3. *An example of insertion in a B-tree, for the optimistic algorithm. Here $m = 3$, nodes contain between 2 and 4 keys. The middle key in red is determined among the 4 keys of the saturated node and the new key, and moves to the parent node.*

2.4. *Insertion as the evolution of a Pólya urn.* For both the prudent and the optimistic algorithms, let us define different types of fringe nodes: we say that a fringe

node is *of type* k, when it contains $m+k-2$ keys and has thus $m+k-1$ gaps. For the prudent algorithm, k varies between 1 and $m+1$; for the optimistic algorithm, k varies between 1 and m.

We analyse the fringe of the tree through the so-called *composition vector* L_n, which counts the number of fringe nodes of each type in a B-tree of parameter m, at time n, i.e., assuming that we start from an empty tree and add the keys one by one, when the tree contains n keys. Thus, $L_n^{(k)}$, the k-th coordinate of L_n, is the number of fringe nodes of type k. For the prudent algorithm, L_n is a vector of dimension $m+1$, whereas it is a vector of dimension m for the optimistic algorithm.

For both algorithms, we define G_n as the composition vector of *gaps* at time n. We say that a gap is *of type* k, when it is attached to a fringe node of type k. Thus $G_n^{(k)}$, the k-th coordinate of G_n, is the number of gaps of type k. In other words:

$$(m+k-1)L_n^{(k)} = G_n^{(k)}. \tag{2.1}$$

For both algorithms, the process $(G_n)_{n\in\mathbb{N}}$ is a Pólya urn process, as defined in the Introduction, where the balls are the gaps and the colors are the different types. Indeed, when the keys are randomly chosen under the so-called random permutation model, meaning that the keys are independently identically distributed (i.i.d.), then the insertion of a new key in a B-tree of size n occurs *uniformly* on any of the $n+1$ gaps of the tree.

• In the prudent algorithm, the number of keys in a fringe node ranges from $m-1$ to $2m-1$, there are $m+1$ types, and the vector L_n is of dimension $m+1$. The replacement matrix of the gap process is of dimension $m+1$ and equal to[3]

$$r_m = \begin{pmatrix} -m & (m+1) & & & & \\ & -(m+1) & (m+2) & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & 2m & \\ m & (m+1) & & & -2m \end{pmatrix}.$$

Figure 2.4 illustrates the same insertion as in Figure 2.2, taking into account the different types (colors) of the fringe nodes.

• In the optimistic algorithm, the number of keys in a fringe node ranges from $m-1$ to $2m-2$, there are m types, and the vector L_n is of dimension m. The replacement matrix of the gap process is of dimension m and equal to

$$R_m = \begin{pmatrix} -m & m+1 & & & \\ & -(m+1) & m+2 & & \\ & & \ddots & & \\ & & & -(2m-2) & 2m-1 \\ 2m & & & & -(2m-1) \end{pmatrix}.$$

This matrix appears in the paper Mahmoud (2002) as a replacement matrix in the so-called paged binary search trees model. Indeed, in the bucket tree, the insertion mechanism induces the same dynamics on the leaves.

Figure 2.5 illustrates the same insertion as in Figure 2.3, taking into account the different types (colors) of the fringe nodes.

[3] An empty entry stands for a zero in all the matrices of this article.

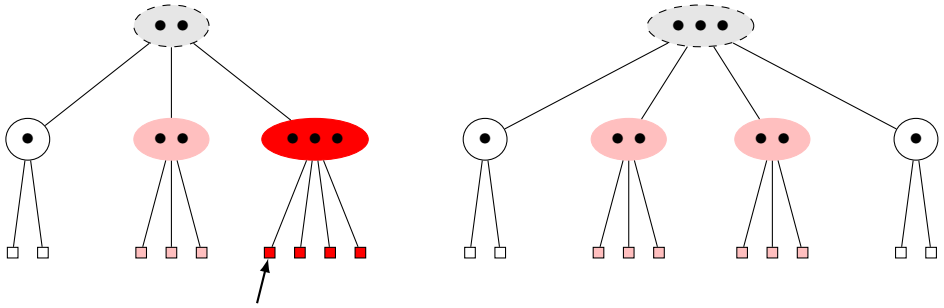

FIGURE 2.4. *An example of insertion in a B-tree, for the prudent algorithm. Here $m = 2$, nodes contain between 1 and 3 keys, there are 3 colors, white, pink and red.*

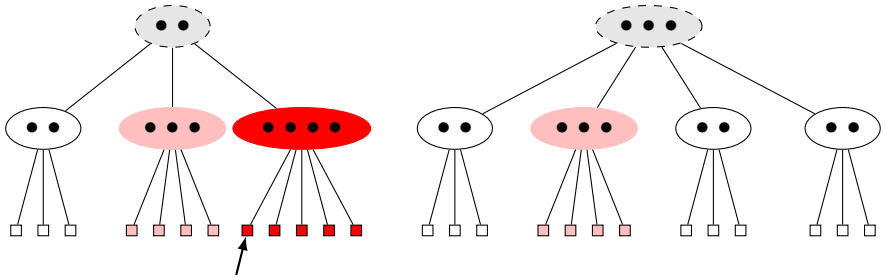

FIGURE 2.5. *An example of insertion in a B-tree, for the optimistic algorithm. Here $m = 3$, nodes contain between 2 and 4 keys, there are 3 colors, white, pink and red.*

Observe that both replacement matrices r_m and R_m are balanced (any row sums to 1), which is an immediate consequence of the dynamics, since one key (one ball) is added at each unit of time.

All the results in this paper hold for both algorithms, including the phase transition depending on whether $m \leq 59$ or $m \geq 60$. However, the proofs and results for the optimistic algorithm being somewhat simpler than those for the prudent algorithm, we choose to present them and to leave the other case to the reader: from now on,

we consider a B-tree constructed by the optimistic algorithm.

3. Gaps of a B-tree as a Pólya urn

Let us remind from Section 2 that a fringe node of a B-tree contains from $m-1$ to $2m-2$ keys and from m to $2m-1$ gaps. For any $k \in \{1, \ldots, m\}$, a fringe node that contains $m+k-2$ keys is called *of type* k. We are interested in the *fringe node composition vector* L_n of a B-tree at time n, whose k-th coordinate counts the number of fringe nodes of type k.

Notation

Let $(e_k)_{1\leq k\leq m}$ be the canonical basis of $\mathbb{R}^m$. Denote by $w_1,\ldots,w_m$ the vectors defined by

$$\begin{cases} w_k = -e_k + e_{k+1},\ 1\leq k\leq m-1 \\ w_m = 2e_1 - e_m. \end{cases} \tag{3.1}$$

The w_k's are the increment vectors of the fringe node dynamics: when a key is inserted in a fringe node of type $k\in\{1,\ldots,m-1\}$, the fringe node is replaced by a fringe node of type $k+1$ (addition of vector w_k) and when a key is inserted in a fringe node of type m, the fringe node is replaced by two fringe nodes of type 1 (addition of vector w_m). When the keys are randomly drawn under the permutation model, the insertion is *uniform* on the gaps and the fringe node composition process of a B-tree is modelized by the $\mathbb{R}^m$-valued Markov chain $(L_n)_{n\in\mathbb{N}}$ defined as follows by its transition probabilities.

Definition of the (discrete-time) fringe node process

For any $k\in\{1,\ldots,m\}$,

$$\mathbb{P}\left(L_{n+1} = L_n + w_k \middle| L_n\right) = \frac{(m+k-1)\langle L_n, e_k\rangle}{K_n} \tag{3.2}$$

where the scalar product $\langle L_n, e_k\rangle = L_n^{(k)}$ is the k-th coordinate of L_n and where K_n denotes the total number of gaps at time n. Of course, when the process starts initially with N_0 keys at time 0, then $K_n = 1 + N_0 + n$. Note that, considering the B-tree, the number of gaps in a fringe node of type k is $m+k-1$ whereas the total number of gaps in the tree at time n is exactly K_n so that Formulae (3.2) completely define a Markov process; it reflects the uniform insertions of the keys in the gaps.

Alternatively, one can consider the gap process. A gap is called of type k when it is contained in a fringe node of type k. Note once more that a fringe node of type k contains $m+k-1$ gaps. Expressed in terms of gaps, the dynamics of key insertion in the B-tree is the following: when the key is inserted in a gap of type $k\in\{1,\ldots,m-1\}$, then $m+k-1$ gaps of type k disappear and are replaced by $m+k$ gaps of type $k+1$; when the key is inserted in a gap of type m, then $2m-1$ gaps of type m disappear and are replaced by $2m$ gaps of type 1. Moreover, under the random permutation model, *the gaps are drawn uniformly*. In other words, the gap composition process of a B-tree is modelized by the following m-color Pólya urn process $(G_n)_{n\in\mathbb{N}}$, having R_m as replacement matrix and $(m,0,\ldots,0)$ as initial composition.

Definition of the (discrete-time) gap process

Denote by $(G_n)_{n\in\mathbb{N}}$ the m-color Pólya urn process defined by the m-dimensional replacement matrix

$$R_m = \begin{pmatrix} -m & m+1 & & & & \\ & -(m+1) & m+2 & & & \\ & & \ddots & & & \\ & & & & -(2m-2) & 2m-1 \\ 2m & & & & & -(2m-1) \end{pmatrix}.$$

Its balance, namely its common row sum, equals 1. Note that the diagonal entries are negative. Nevertheless, the urn is tenable because, in any column, all entries

are multiple of the diagonal coefficient: when a ball of color 2 is drawn, $m+1$ extra balls must be withdrawn from the urn which is always possible because balls of type 2 are put $m+1$ by $m+1$ in the urn as can be seen on R_m's second column. The same phenomenon occurs for any color. Of course, these negative diagonal entries imply that one must necessarily take an initial composition that satisfies such divisibility conditions as well: for any $k \in \{1, \dots, m\}$, the initial number of balls of type k satisfies

$$m + k - 1 \ \text{ divides } \ \langle L_0, e_k \rangle.$$

The symbol $\langle ., . \rangle$ denotes here the standard scalar product on $\mathbb{R}^m$. Thus, the condition (1.1) is fulfilled.

Both Markov processes $(L_n)_n$ and $(G_n)_n$ are related by Relation (3.4), stated hereunder. Let P be the m-dimensional diagonal invertible matrix

$$P = \text{Diag}\,(m, m+1, \dots, 2m-1)\,. \tag{3.3}$$

When $V \in \mathbb{N}^m \setminus \{0\}$, denote by $\left(L_n^V\right)_{n \geq 0}$ the fringe node process starting with $L_0 = V$ and by $\left(G_n^V\right)_{n \geq 0}$ the gap process starting from $G_0 = V$. Then, for any $V \in \mathbb{N}^m \setminus \{0\}$, one has[4] immediatly from (2.1)

$$\left(G_n^{PV}\right)_{n \in \mathbb{N}} \overset{\mathcal{L}}{=} \left(PL_n^V\right)_{n \in \mathbb{N}}. \tag{3.4}$$

In particular, denoting by $|V|$ the sum of V's coordinates, the total number of gaps in the B-tree at time n is

$$K_n = n + |PV| = \sum_{k=1}^{m} (m + k - 1) \langle L_n^V, e_k \rangle = \sum_{k=1}^{m} \langle G_n^{PV}, e_k \rangle.$$

In the following, when no confusion is possible, we lighten the notation G_n^{PV} into G_n, and L_n^V into L_n, like in Theorems 4.1 and 4.3 below.

4. Phase transition: *small* and *large* B-trees

With regard to the asymptotics of their composition vector, Pólya urns are subject to a well known phase transition. See the above references on Pólya urns for a general treatment. Let us translate the phenomenon for our B-urns, looking at the spectral properties of the replacement matrix.

4.1. *Spectral decomposition of the ambient space.* In this section, we state notations relative to the spectral decomposition of the matrix R_m. These notations are used all along the paper. The (unitary) characteristic polynomial of R_m turns out to be

$$\chi_m(X) = \prod_{k=m}^{2m-1} (X + k) - \frac{(2m)!}{m!}. \tag{4.1}$$

Since the matrix R_m belongs to a particular family of matrices studied in Annexe B of the thesis Hennequin (1991), one can see that its complex roots are all simple, the one having the largest real part one being 1. Furthermore, two distinct eigenvalues that have the same real part are conjugated. We denote by

$$\lambda_2 = \sigma_2 + i\tau_2 \tag{4.2}$$

[4] When no confusion is possible, we denote by PV the product of the square matrix P by the vector $V \in \mathbb{R}^m$ instead of the correct form $P^t V$.

the eigenvalue of R_m having the second largest real part σ_2 and a positive imaginary part τ_2. We adopt also the following notations:

$$\begin{cases} H_{m+1}(X) = \displaystyle\sum_{k=1}^{m} \frac{1}{X+k} \\ \\ v(\lambda) = \dfrac{1}{(m+\lambda)H_{m+1}(m+\lambda-1)} \times \\ \\ \quad \left(1, \dfrac{m+1}{m+1+\lambda}, \dfrac{(m+1)(m+2)}{(m+1+\lambda)(m+2+\lambda)}, \ldots, \dfrac{(m+1)\ldots(2m-1)}{(m+1+\lambda)\ldots(2m-1+\lambda)}\right) \\ \\ \langle v(\lambda), e_k \rangle = \dfrac{1}{(m+\lambda)H_{m+1}(m+\lambda-1)} \displaystyle\prod_{j=1}^{k-1} \frac{m+j}{m+j+\lambda} \\ \\ u(\lambda)\,(x_1, \ldots, x_m) = \displaystyle\sum_{k=1}^{m} \left(\prod_{j=0}^{k-2} \frac{\lambda+m+j}{1+m+j} \right) x_k \\ \\ \qquad\qquad = x_1 + \dfrac{\lambda+m}{1+m}\, x_2 + \dfrac{(\lambda+m)(\lambda+m+1)}{(1+m)(2+m)}\, x_3 + \cdots \\ \\ \qquad\qquad + \dfrac{(\lambda+m)\ldots(\lambda+2m-2)}{(1+m)\ldots(2m-1)}\, x_m. \end{cases} \tag{4.3}$$

When λ is an eigenvalue of R_m, the vector $v(\lambda)$ is an eigenvector of ${}^tR_m$ associated with λ. The linear form $u(\lambda)$ is an eigenform of ${}^tR_m$ associated with λ, which means that for any (column) vector V, $u(\lambda)\left({}^tR_m V\right) = \lambda u(\lambda)(V)$. Moreover, if λ and μ are eigenvalues of R_m, then $u(\lambda)\,[v(\mu)] = \delta_{\lambda,\mu}$ (Kronecker). In other words, $(v(\lambda))_{\lambda \in \mathrm{Sp}(R_m)}$ and $(u(\lambda))_{\lambda \in \mathrm{Sp}(R_m)}$ are dual basis of respectively eigenvectors and eigenforms of ${}^tR_m$.

In the sequel, for more simplicity, we denote

$$\begin{cases} v_1 = v(1) = \dfrac{1}{H_{m+1}(m)} \left(\dfrac{1}{m+1}, \dfrac{1}{m+2}, \ldots, \dfrac{1}{2m} \right) \\ \\ u_1\,(x_1, \ldots, x_m) = u(1)\,(x_1, \ldots, x_m) = \displaystyle\sum_{k=1}^{m} x_k \\ \\ v_2 = v(\lambda_2) \text{ and } u_2 = u(\lambda_2). \end{cases} \tag{4.4}$$

The complex vector space $\mathbb{C}^m$ admits the decomposition as direct sum of ${}^tR_m$-stable lines

$$\mathbb{C}^m = \bigoplus_{\lambda \in \mathrm{Sp}(R_m)} \mathbb{C} v(\lambda)$$

and the corresponding projection on any line $\mathbb{C}v(\lambda)$ is $u(\lambda)v(\lambda)$. In the real field, we use the decomposition

$$\mathbb{R}^m = \mathbb{R}v_1 \oplus \mathcal{V}_1 \tag{4.5}$$

where $\mathcal{V}_1$ is the only subspace which is simultaneously ${}^tR_m$-stable and supplementary to $\mathbb{R}v_1$. It is generated by the vectors respectively constituted by the real parts and the imaginary parts of the coordinates of the complex vectors $v(\lambda)$, $\lambda \in \mathrm{Sp}\,(R_m)\setminus\{1\}$. In the same vein, we denote by $\mathcal{V}_2$ the only ${}^tR_m$-stable subspace of $\mathbb{R}^m$ that satisfies

$$\mathbb{R}^m = \mathbb{R}v_1 \oplus \mathbb{R}\Re\,(v_2) \oplus \mathbb{R}\Im\,(v_2) \oplus \mathcal{V}_2.$$

4.2. *Phase transition for gaps and fringe nodes.* The phase transition on urns is expressed on the gap process $(G_n)_n$ in the following result. Note the two very different convergence modes: a weak one for small phases *vs* a strong one with periodic phenomena for large phases. See simulations in Section 7 for an illustration.

Theorem 4.1. *Let $m \geq 2$. Let $V \in \mathbb{N}^m$ be a non zero vector and let $(G_n)_{n\geq 0}$ be the (discrete-time) gap process starting with the initial condition $G_0 = PV$. Then, with notations (4.3) and (4.4),*

(i) (Small phases)
if $m \leq 59$, as n tends to infinity, $\dfrac{G_n - nv_1}{\sqrt{n}}$ converges in distribution to a centered Gaussian vector;

(ii) (Large phases)
if $m \geq 60$, as n tends to infinity,

$$G_n = nv_1 + 2\Re\left(n^{\lambda_2} W^{DT} v_2\right) + o\left(n^{\sigma_2}\right), \tag{4.6}$$

almost surely and in any L^p, $p \geq 1$, where W^{DT} is a complex-valued random variable with expectation $\dfrac{\Gamma\left(|PV|\right)}{\Gamma\left(|PV|+\lambda_2\right)} u_2(PV)$ (Γ denotes Euler Gamma function).

Proof: These results come from the general theory of balanced Pólya urn processes. See Janson (2004) or Pouyanne (2008). The numerical values of σ_2 leading to the phase transition are given in the Appendix. □

Remark 4.2. The phase transition at $m = 59$ is stated in Mahmoud (2002), section 8.1 for the same Pólya urn, seen as the evolution model of the so called paged binary search trees.

In Mahmoud's paper is given the mean of the composition vector G_n as well as the Gaussian asymptotic behaviour of a scalar random variable derived from the vector G_n, in the case $m \leq 59$.

In Theorem 4.1(ii), for any $p \in]0,1[$, the asymptotics is also valid in the (not locally convex) complete metric space L^p defined by the usual quasi-norm. This is true for Theorem 4.3, Corollary 4.4, Theorem 5.1, Theorem 5.2 and Remark 5.3 as well.

When V is a complex vector, $\Re(V)$ denotes the vector made of real parts of V's coordinates. The random variable W^{DT}, which is more deeply studied below, appears as a martingale limit in the field of urn theory. It can be also described as the almost sure limit of G_n after normalisation and projection along the principal direction defined by v_2:

$$W^{DT} = \lim_{n\to\infty} \frac{1}{n^{\lambda_2}} u_2\left(G_n\right).$$

Of course, the phase transition and the asymptotics can be straightforwardly translated on the fringe node process $(L_n)_n$ *via* the diagonal matrix P defined in (3.3). The random variable W^{DT} that appears for large phases in Theorem 4.3 has the same law as the one of the gap process in Theorem 4.1.

Theorem 4.3. *Let $m \geq 2$. Let $V \in \mathbb{N}^m$ be a non zero vector and let $(L_n)_{n\geq 0}$ be the (discrete-time) fringe node process starting with the initial condition $L_0 = V$. Then, with notations (4.3), (4.4) and (3.3),*

(i) (Small phases)

if $m \leq 59$, as n tends to infinity, $\dfrac{L_n - nP^{-1}v_1}{\sqrt{n}}$ converges in distribution to a centered Gaussian vector;

(ii) (Large phases)
if $m \geq 60$, as n tends to infinity,

$$L_n = nP^{-1}v_1 + 2\Re\left(n^{\lambda_2} W^{DT} P^{-1} v_2\right) + o\left(n^{\sigma_2}\right), \tag{4.7}$$

almost surely and in any L^p, $p \geq 1$, where W^{DT} is a complex-valued random variable which has the same distribution as in the variable named the same way in Theorem 4.1.

Geometrically speaking, expansion (4.6) (and expansion (4.7) as well) can be understood as follows. Notice that an analogous explanation holds for expansions (5.4) and (5.5) in Theorem 5.1 and Theorem 5.2 respectively.

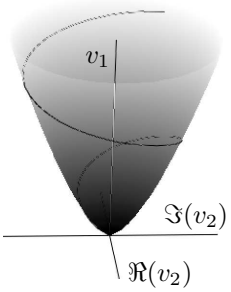

FIGURE 4.6. Spiral

Let us denote by φ any argument of the complex number W^{DT}. The trajectory of the random vector G_n, projected in the 3-dimensional real vector space spanned by the vectors $(\Re(v_2), \Im(v_2), v_1)$ is almost surely asymptotic to the (random) spiral

$$\begin{cases} x_n = 2|W| n^{\sigma_2} \cos(\tau_2 \log n + \varphi), \\ y_n = -2|W| n^{\sigma_2} \sin(\tau_2 \log n + \varphi), \\ z_n = n, \end{cases}$$

drawn on the (random) revolution surface

$$4|W|^2 z^{2\sigma_2} = x^2 + y^2,$$

when n tends to infinity (see Figure 4.6).

As is well known in the field of Pólya urn processes, the phase transition is due to the number σ_2. When $\sigma_2 < 1/2$, the Pólya urn is *small* and admits a weak Gaussian

asymptotics. On the contrary, when $\sigma_2 > 1/2$, the urn is *large* and has a strong and oscillating (λ_2 is nonreal) asymptotic behaviour. Considering the replacement matrix R_m, it turns out that σ_2 is an increasing function of m and that:

- when $m = 59$, $\lambda_2 = (0.49534...) + (9.10305...)i$

while

- when $m = 60$, $\lambda_2 = (0.50378...) + (9.10270...)i$.

These numerical values have been computed by a Newton approximation algorithm, which can be found in the Appendix. The monotonicity of σ_2 as a function of m (it increases to 1 when m tends to infinity) has been evoked by Hennequin (1991) in a figure. Qualitatively, let us emphasize the fact that for large values of m (which is the actual use in computer science, since m amounts to several hundreds), the fluctuation term with W^{DT} is highly significant.

We deduce from these theorems the asymptotic behaviour of the composition vector of the fringe nodes of different types in a B-tree, which is a particular case of Theorem 4.3 with the initial condition $V = (1, 0, \dots, 0)$. We stated the theorems above for arbitrary initial conditions because of the further study of the limit law W^{DT} that requires these wider statements.

Corollary 4.4. *Let $m \geq 2$. Let $\mathcal{L}_n$ be the composition vector at time n of the fringe nodes of different types in a B-tree with minimum degree m. Then, as n goes of to infinity, with notations (3.3) and (4.4),*

(i) when $m \leq 59$, $\dfrac{\mathcal{L}_n - nP^{-1}v_1}{\sqrt{n}}$ converges in distribution to a centered Gaussian vector;

(ii) when $m \geq 60$, $\mathcal{L}_n = nP^{-1}v_1 + 2\Re\left(n^{\lambda_2} W_m^{\mathrm{B-tree}} P^{-1}v_2\right) + o\left(n^{\sigma_2}\right)$, almost surely and in any L^p, $p \geq 1$, where $W_m^{\mathrm{B-tree}}$ is a complex-valued random variable with expectation $\dfrac{m!}{\Gamma(m + \lambda_2)}$.

5. Embeddings into continuous time

Going further and obtaining significant properties of the random limit W^{DT} is not so easy. As can be seen in this section, the classical method of embedding in continuous time turns out to be very fruitful: this idea of embedding discrete urn models in continuous-time branching processes goes back at least to Athreya and Karlin (1968). A description is given in the book Athreya and Ney (1972, Section 9). The method has been recently revisited and developed by Janson (2004), it is the core of recent results on Pólya urns in Chauvin et al. (2011, 2015).

5.1. *Definition of the continuous-time fringe node process.* Denote by $(L(t))_{t \in \mathbb{R}_{\geq 0}}$ the $\mathbb{N}^m \setminus \{0\}$-valued continuous time Markov process having $\mathcal{G}$ as infinitesimal generator, where $\mathcal{G}$ is defined, for any function $f : \mathbb{N}^m \setminus \{0\} \to \mathcal{V}$ ($\mathcal{V}$ is any real or complex vector space) and for any nonzero $X = (x_1, \dots, x_m) \in \mathbb{N}^m$, by

$$\mathcal{G}(f)(X) = \sum_{k=1}^{m} (m + k - 1) x_k \Big[f(X + w_k) - f(X) \Big]$$

where the increment vectors w_k have already be defined by (3.1).

This process is a multitype branching process, embedding of the Markov chain $(L_n)_n$ into continuous time, as classically done (see for example Bertoin, 2006). One

can think of it the following way. At each (real) time $t \geq 0$, one gets particles of m different types named $1, 2, \ldots, m$. Each particle is equipped with a clock that rings at random times. The clock of any particle of type k is exponentially distributed, with parameter $m+k-1$ and all the clocks are independent. The dynamics of the process is the same as in discrete time: for any $k \in \{1, \ldots, m-1\}$, when the clock of a particle of type k rings, the particle disappears and is replaced by a particle of type $k+1$; when the clock of a particle of type m rings, the particle disappears and is replaced by two particles of type 1.

Having the same dynamics, the distributions of the processes $\big(L_n\big)_{n\in\mathbb{N}}$ and $\big(L(t)\big)_{t\in\mathbb{R}_{\geq 0}}$ are as usual related by the finite-time connection

$$(L_n)_{n\in\mathbb{N}} \stackrel{\mathcal{L}}{=} \big(L\left(\tau_{(n)}\right)\big)_{n\in\mathbb{N}} \tag{5.1}$$

where $\tau_{(n)}$ denotes the n-th splitting time (the n-th ringing time). This relation allows us to transfer results on one process to the other. In particular, the results below strongly rely on the fact that $(e^{-({}^tR_m)t}L(t))_{t\in\mathbb{R}_{\geq 0}}$ is a vector-valued martingale (see Janson, 2004 or Athreya and Ney, 1972 for this).

5.2. *Definition of the continuous-time gap process.*

Define the vector-valued continuous-time Markov process $(G(t))_{t\in\mathbb{R}_{\geq 0}}$ as being the embedding into continuous time of the discrete-time urn process $(G_n)_{n\in\mathbb{N}}$. It takes its values in the set of vectors of the form PV where $V \in \mathbb{N}^m \setminus \{0\}$. With notations as above, its infinitesimal generator is given by

$$\mathcal{H}(f)(X) = \sum_{k=1}^{m} x_k \Big[f\left(X + Pw_k\right) - f\left(X\right) \Big],$$

the increment vectors Pw_k being the rows of the urn replacement matrix R_m.

One can think of this process the following way. Take an urn that contains clocks of m different colors named $1, \ldots, m$. Each clock rings at a random time, exponentially distributed with parameter 1 and all the clocks are independent. As soon as a clock rings, the following replacement mechanism occurs: if the ringing clock has color $k \in \{1, \ldots, m-1\}$, then it disappears together with $m+k-2$ other clocks of color k and $m+k$ clocks of color $k+1$ arise in the urn; if the ringing clock has color m, then it disappears together with $2m-2$ other clocks of color k and $2m$ clocks of color 1 arise in the urn. The fact that the ringing times are exponentially distributed allows to think as if all clocks were restarted as soon as one of them rings. Note that the fact that many clocks disappear at the same time prevents $(G(t))_{t\in\mathbb{R}_{\geq 0}}$ from being a multitype branching process.

As in the preceding case, the processes $(G_n)_{n\in\mathbb{N}}$ and $(G(t))_{t\in\mathbb{R}_{\geq 0}}$ have the same dynamics, so that

$$(G_n)_{n\in\mathbb{N}} \stackrel{\mathcal{L}}{=} \Big(G\left(\tau'_{(n)}\right)\Big)_{n\in\mathbb{N}}, \tag{5.2}$$

where $\tau'_{(n)}$ denotes the n-th ringing time.

When $V \in \mathbb{N}^m \setminus \{0\}$, denote by $\big(L(t)^V\big)_{t\geq 0}$ the fringe node process starting with $L(0) = V$ and by $\big(G(t)^V\big)_{t\geq 0}$ the gap process starting from $G(0) = V$. Then, as in the discrete-time case in (3.4), for any $V \in \mathbb{N}^m \setminus \{0\}$,

$$\big(G(t)^{PV}\big)_{t\geq 0} \stackrel{\mathcal{L}}{=} \big(PL(t)^V\big)_{t\geq 0}, \tag{5.3}$$

where P is the diagonal matrix defined in (3.3).

5.3. *Asymptotics of both continuous-time processes.*

The asymptotics of the continuous-time processes admit the same kind of phase transition as in discrete time. We state this asymptotics for both continuous-time processes in Theorems 5.2 and 5.1. Since G is the image of L by P, any of these theorem implies the other one. Nevertheless, as explained below, we prove both of them together using results on branching processes and results on Pólya urns.

Theorem 5.1. *Let $m \geq 2$. Let $V \in \mathbb{N}^m$ be a non zero vector and let $(G(t))_{t\in\mathbb{R}_{\geq 0}}$ be the continous-time gap process that satisfies $G(0) = PV$. Then, with notations (4.3),*

(i) (Small phases)

when $m \leq 59$, as t tends to infinity, $e^{-t}G(t)$ converges almost surely and in any L^p, $p \geq 1$, to ξv_1 where ξ is a positive random variable which is Gamma-distributed with parameter $|PV|$. Furthermore, if one writes $G(t) = G_1(t) + G'_1(t)$ where the random vector $G_1(t)$ is proportional to v_1 and where $G'_1(t)$ is $\mathcal{V}_1$-valued (see(4.5)), then $e^{-t}G_1(t)$ converges almost surely and in any L^p to ξv_1 while $e^{-t/2}G'_1(t)$ converges in distribution to $\sqrt{\xi}N$ where N is a centered $\mathcal{V}_1$-valued Gaussian vector independent of ξ.

(ii) (Large phases)

when $m \geq 60$, as t tends to infinity,

$$G(t) = e^t \xi v_1 \left(1 + o(1)\right) + 2\Re\left(e^{\lambda_2 t} W^{CT} v_2\right)\left(1 + o(1)\right) + o\left(e^{\sigma_2 t}\right) \tag{5.4}$$

almost surely and in any L^p, $p \geq 1$, where W^{CT} is a complex-valued random variable with expectation $u_2(PV)$ and ξ a positive random variable that is Gamma distributed with parameter $|PV|$. The almost sure remainder $o\left(e^{\sigma t}\right)$ is a $\mathcal{V}_2$-valued random vector.

Theorem 5.2. *Let $m \geq 2$. Let $V \in \mathbb{N}^m$ be a non zero vector and let $(L(t))_{t\in\mathbb{R}_{\geq 0}}$ be the continous-time fringe node process that satisfies $L(0) = V$. Then, with notations (4.3) and (3.3),*

(i) (Small phases)

when $m \leq 59$, as t tends to infinity, $e^{-t}L(t)$ converges almost surely and in any L^p, $p \geq 1$, to $\xi P^{-1} v_1$ where ξ is a positive random variable which is Gamma-distributed with parameter $|PV|$. Furthermore, if one writes $L(t) = L_1(t) + L'_1(t)$ where the random vector $L_1(t)$ is proportional to $P^{-1}v_1$ and where $L'_1(t)$ is $P^{-1}\mathcal{V}_1$-valued (see(4.5)), then $e^{-t}L_1(t)$ converges almost surely and in any L^p to $\xi P^{-1}v_1$ while $e^{-t/2}L'_1(t)$ converges in distribution to $\sqrt{\xi}N'$ where N' is a centered $P^{-1}\mathcal{V}_1$-valued Gaussian vector independent of ξ.

(ii) (Large phases)

when $m \geq 60$, as t tends to infinity,

$$L(t) = e^t \xi P^{-1} v_1 \left(1 + o(1)\right) + 2\Re\left(e^{\lambda_2 t} W^{CT} P^{-1} v_2\right)\left(1 + o(1)\right) + o\left(e^{\sigma_2 t}\right) \tag{5.5}$$

almost surely and in any L^p, $p \geq 1$, where W^{CT} is a complex-valued random variable with expectation $u_2(PV)$ and ξ a positive random variable that is Gamma distributed with parameter $|PV|$. The almost sure remainder $o\left(e^{\sigma t}\right)$ is a $P^{-1}\mathcal{V}_2$-valued random vector.

Note that the random variables ξ and W^{CT} that appear in both theorems have been denoted the same way because their distributions are the same in both cases. This comes immediately from (5.3).

Proof of Theorems 5.1 and 5.2: Despite the fact that similar results can be found in Janson (2004) and Mailler (2014), the particular case of our processes is not properly contained in their statement. The proofs are essentially made the same way as in both papers Chauvin et al. (2014, 2012), which deal with m-ary search trees. Despite m-ary search trees are a close model, where replacement matrices belong to the same family of matrices studied by Hennequin (1991), precise results contained in Theorems 5.1 and 5.2 as well as smoothing equations (6.2) and (6.3) below cannot be directly deduced from analogous results in Chauvin et al. (2014, 2012).

We give hereunder the general scheme of the argumentation. The first tool comes from the fact that the normalised projection $(e^{-t}u_1(G(t)))_{t\geq 0}$ is always a convergent positive martingale. The random variable ξ is its limit.

(i) Small phases. The process $(L(t))_{t\in\mathbb{R}_{\geq 0}}$ is a multitype branching process so that *(i)* in Theorem 5.2 is covered by Athreya and Ney (1972) and Janson (2004). Relation (5.3) thus implies *(i)* in Theorem 5.1.

(ii) Large phases. As for the first projection, $\left(e^{-\lambda_2 t}u_2(G(t))\right)_{t\geq 0}$ is a martingale, which is convergent if, and only if $\sigma_2 > 1/2$, *i.e.* when $m \geq 60$. The complex-valued random variable W^{CT} is its limit. The oscillating term $\Re\left(e^{\lambda_2 t}W^{CT}P^{-1}v_2\right)$ in Theorem 5.2 is a consequence of Athreya and Ney (1972) and Janson (2004)'s results. In order to establish the almost sure remainders $o\left(e^{\sigma_2 t}\right)$, we use results on discrete-time Pólya urns shown in Pouyanne (2008). The work is done on the gap process $(G(t))_t$ viewed as an embedded urn into continuous time. For any $t \geq 0$, decompose $G(t)$ as the sum $G(t) = G_1(t) + G_2(t) + G_\ell(t) + G_s(t)$ of its respective following projections on the described supplementary subspaces:

- $G_1(t)$ is the projection on $\mathbb{R}v_1$ as before;
- $G_2(t)$ is the projection on the real plane generated by the real part and the imaginary part of v_2;
- $G_\ell(t)$ is the projection on the subspace of $\mathbb{R}^m$ generated by the real and imaginary parts of the eigenvectors $v(\lambda)$ for all eigenvalues λ different from 1 and λ_2 such that $\Re(\lambda) > 1/2$ (*large* projections);
- finally, $G_s(t)$ is the projection on the subspace of $\mathbb{R}^m$ generated by the real and imaginary parts of the eigenvectors $v(\lambda)$ for all eigenvalues λ such that $\Re(\lambda) \leq 1/2$ (*small* projections).

As seen before, $e^{-t}G_1(t)$ converges to ξv_1 almost surely and in L^p, $p \geq 1$, by martingale techniques; this gives rise to the first term $e^t\xi v_1$ in the asymptotics of $G(t)$. Since $G_2(t) = 2\Re\left[u_2\left(G_2(t)\right)v_2\right]$ and because of the convergence in L^p, $p \geq 1$, of the complex martingale $\left(e^{-\lambda_2 t}u_2(G(t))\right)_{t\geq 0}$ mentioned above, one gets the second term $\Re\left(e^{\lambda_2 t}W^{CT}v_2\right)$ of $G(t)$'s asymptotics. The remainder $o\left(e^{\sigma_2 t}\right)$ is obtained from G_ℓ and G_s asymptotics. As for G_2, if λ is an eigenvalue of R_m such that $\Re(\lambda) > 1/2$, by martingale arguments, the complex projection of $G(t)$ on any eigenline $\mathbb{C}v(\lambda)$ is equivalent to $e^{\lambda t}W_\lambda$ almost surely and in any L^p where W_λ is a complex-valued random variable. In particular, the whole projection $G_\ell(t)$ is $o\left(e^{\sigma_2 t}\right)$, almost surely and in any L^p.

To make the proof complete, it remains to show that $G_s(t)$ is $o\left(e^{\sigma_2 t}\right)$ as well. To prove this fact, we use the technique detailed in Chauvin et al. (2014) (Theorem 4.1 and Lemma 4.2). It consists in considering the same projection for the discrete-time urn process $(G_n)_n$, in using the moment bounds proven in Pouyanne (2008) for small projections of discrete-time Pólya urns and in coming back to continuous time by Relation (5.2). By this means, one shows after some probabilistic arguments that for any $\eta > 0$, the whole projection G_s satisfies that $e^{-\left(\eta+\frac{1}{2}\right)t} G_s(t)$ is bounded, almost surely and in L^p, $p \geq 1$, implying the expected result on $G(t)$. The corresponding asymptotics of $L(t)$ is obtained by taking the image of $G(t)$ by P^{-1}. □

Remark 5.3. For $m \geq 60$, we deduce from these theorems the asymptotic behaviour of the continuous-time fringe node process, denoted by $(\mathcal{L}(t))_t$, starting from the B-tree initial condition $V = (1, 0, \ldots, 0)$:

$$\mathcal{L}(t) = e^t \xi P^{-1} v_1 \left(1 + o(1)\right) + 2\Re\left(e^{\lambda_2 t} \mathcal{W}^{CT} P^{-1} v_2\right)\left(1 + o(1)\right) + o\left(e^{\sigma_2 t}\right) \tag{5.6}$$

almost surely and in any L^p, $p \geq 1$, where $\mathcal{W}^{CT}$ is a complex-valued random variable with expectation m and ξ a positive random variable that is Gamma distributed with parameter m. The almost sure remainder $o\left(e^{\sigma t}\right)$ is a $P^{-1}\mathcal{V}_2$-valued random vector.

For large phases, the finite time connections (5.2) or (5.1) lead to a relation between the random variables W in discrete and continuous times. This relation, commonly named *martingale connection* will be stated and used below in the article. We indicate hereafter how one can get it. Take for instance Relation (5.2) concerning the gap processes $(G_n)_n$ and $(G(t))_t$ starting with the same initial condition $G_0 = G(0) = PV$. Using Theorems 4.1 and 5.1, since $\tau_{(n)}$ tends almost surely to $+\infty$ as n goes of to infinity, one gets successively $\xi = \lim_{t\to\infty} e^{-t} u_1\left(G(t)\right) = \lim_{n\to\infty} e^{-\tau_{(n)}} u_1(G_n) = \lim_{n\to\infty} n e^{-\tau_{(n)}}$ on one hand. On the other hand, $W^{CT} = \lim_{t\to\infty} e^{-\lambda_2 t} u_2\left(G(t)\right) = \lim_{n\to\infty} e^{-\lambda_2 \tau_{(n)}} u_2(G_n) = \lim_{n\to\infty} \left[n e^{-\tau_{(n)}}\right]^{\lambda_2} \left[n^{-\lambda_2} u_2(G_n)\right]$. This entails the martingale connection

$$W^{CT} \overset{\mathcal{L}}{=} \xi^{\lambda_2} W^{DT}. \tag{5.7}$$

We just recall here that the random variable ξ is Gamma-distributed with expectation $|PV|$.

6. Limit law of large B-trees

In this section appear the benefits of the embedding in continuous time. Indeed, the branching property applied to the fringe node process $(L(t))_t$, together with the asymptotics proved in Theorem 5.2, allow us to see the limit W^{CT} as a solution of a distributional equation. This is detailed in Section 6.1. It is the starting point to deduce several properties of W^{CT}: its distribution is the unique solution of such an equation in a convenient space of probability distributions (Theorem 6.2 in Section 6.3); it admits exponential moments in a neighborhood of 0 (Theorem 6.5 in Section 6.4); up to a change of function, its Laplace transform is a solution of the quite simple (but unsolvable!) differential equation $y^{(m)} = y^2$ (Theorem 6.6 in Section 6.5); it admits a density relatively to Lebesgue measure on $\mathbb{C}$ and its support is the whole complex plane (Theorem 6.7 in Section 6.6). Thanks to connection (5.7) between W^{CT} and W^{DT}, corresponding results are true for W^{DT} and consequently for $W_m^{\mathrm{B-tree}}$.

6.1. *Dislocation equations in continuous time.* In this section, using the branching property of the continuous-time process $(L(t))_t$, we show that the complex-valued random variable W^{CT} is solution of a very simple distributional equation.

In order to simplify the notations, for any $k \in \{1, \dots, m\}$, denote by W_k the limit random variable W^{CT} (or its distribution) of the continuous-time fringe node process $(L(t)^{e_k})_t$ that starts with one particle of type k, which means that its initial composition $L(0)$ is the k-th vector e_k of $\mathbb{R}^m$ canonical basis. Denote also by τ_k the *first* splitting time of the process $(L(t)^{e_k})_t$; its is exponentially distributed, with parameter $m + k - 1$.

Because of the branching property of the process $(L(t))_t$, for any time $t \geq \tau_1$, the processes $(L(t)^{e_1})_{t\geq 0}$ and $(L(t)^{e_2})_{t\geq 0}$ are related by the distributional equation

$$L(t)^{e_1} \stackrel{\mathcal{L}}{=} L(t - \tau_1)^{e_2}.$$

In the asymptotic form given by Theorem 5.2, consider the second order term on both sides of the equality, which consists in projecting, normalizing and letting t tend to infinity. This leads to the distributional equality

$$W_1 = e^{-\lambda_2 \tau_1} W_2,$$

the random variables W_2 and τ_1 being independent. Doing the same for all values of $k \in \{1, \dots, m\}$ leads to the distributional system:

$$\left\{ \begin{array}{rcl} W_1 & \stackrel{\mathcal{L}}{=} & e^{-\lambda_2 \tau_1} W_2 \\ W_2 & \stackrel{\mathcal{L}}{=} & e^{-\lambda_2 \tau_2} W_3 \\ & \vdots & \\ W_{m-1} & \stackrel{\mathcal{L}}{=} & e^{-\lambda_2 \tau_{m-1}} W_m \\ W_m & \stackrel{\mathcal{L}}{=} & e^{-\lambda_2 \tau_m} \left(W_1^{(1)} + W_1^{(2)} \right) \end{array} \right. \tag{6.1}$$

where

- for any $k \in \{1, \dots, m-1\}$, the random variables τ_k and W_{k+1} of the k-th equation's right-hand sides are independent;
- in the right-hand side of the last equation, the random variables $W_1^{(1)}$ and $W_1^{(2)}$ are independent copies of W_1, both being independent of τ_m as well.

We recall that for any $k \in \{1, \dots, m\}$, the random variable τ_k is exponentially distributed, with parameter $m+k-1$ (see Section 5). In particular, W_1 is a solution of the following distributional equation, sometimes called fixed point equation or smoothing equation in some branching processes contexts (see Liu, 1998 or Biggins and Kyprianou, 2005).

$$W_1 \stackrel{\mathcal{L}}{=} B^{\lambda_2} \left(W_1^{(1)} + W_1^{(2)} \right) \tag{6.2}$$

where

- the random variables $W_1^{(1)}$ and $W_1^{(2)}$ are independent copies of W_1;
- B is a random variable, independent of $W_1^{(1)}$ and $W_1^{(2)}$, Beta distributed with parameters (m, m) which means that it admits $t^{m-1}(1-t)^{m-1}\mathbb{1}_{[0,1]}(t)$ as a density ($\mathbb{1}_A$ denotes the indicatrix function of the set A).

The distribution of B is computed the following way. By immediate computation from System (6.1), one sees that $B = e^{-(\tau_1+\tau_2+\cdots+\tau_m)}$, the variables τ_k being mutually independent. To recognize the Beta(m, m) law, one can make a direct

computation of its density or compute its moments (a Beta distribution is characterized by its moments because its support is compact).

6.2. *Smoothing equation in discrete time.* In a general setting of m-color Pólya urns, including the case of negative entries on the diagonal of the replacement matrix, Mailler (2014) proves that W^{DT} is a solution of a distributional equation which turns to be in our case

$$W \overset{\mathcal{L}}{=} B_1^{\lambda_2} W^{(1)} + B_2^{\lambda_2} W^{(2)}, \tag{6.3}$$

where

- the random variables $W^{(1)}$ and $W^{(2)}$ are independent copies of W;
- (B_1, B_2) is a random vector, independent of $W^{(1)}$ and $W^{(2)}$, Dirichlet distributed with parameters (m, m), which means that $B_1 + B_2 = 1$ and that B_1 and B_2 are Beta distributed with parameters (m, m).

Remark 6.1. One proof of this result in Mailler (2014) uses the tree structure of the urn. Nevertheless, we do not actually understand what kind of "divide-and-conquer" type argument, applied to B-trees, could lead to this equation. Indeed, in other cousin models, like m-ary search trees (see Fill and Kapur, 2004), such a backward decomposition leads to a finite time decomposition equation and passing to the limit, it gives the distributional equation.

6.3. *Contraction methods.* The question of existence and unicity of solutions of equations like (6.2) or (6.3) is classically solved using the Banach fixed point theorem. One point of view, frequent in analysis of algorithms, consists in starting from a decomposition property of the algorithm at finite time, deduce a distributional equation on a cost variable, and pass to the limit to get a smoothing equation on the limit random variable. See Knape and Neininger (2014) for Pólya urns, and also the general paper by Neininger and Rüschendorf (2004) or the survey Neininger and Rüschendorf (2006) for many examples of this so-called contraction method. Another point of view (in this article) consists in taking advantage of the dynamics of the algorithm and exhibiting a martingale limit, solution of a smoothing equation. Thus, the existence is automatically achieved. In both points of view, to get the unicity, the contraction property has to be established, in a convenient space of probability distributions, classically equipped with a Wasserstein distance to get a complete metric space of measures.

We do not prove here the theorem below, since for equation (6.2), it is done in a general framework by Mailler (2014, Prop 1 in Section 5.2) and for equation (6.3), a similar proof is done in Janson (2004, proof of Th 3.9 (iii)). Remark that conditions (tenability, irreducibility and balance of the replacement matrix R_m) needed in these results are fullfilled, as already noticed in Section 3. The same kind of results can be found in Knape and Neininger (2014) even if the only case $a_{i,i} \geq -1$ is considered there. See also Chauvin et al. (2014, 2012).

Theorem 6.2. *When A is a complex number, let $\mathcal{M}_2(A)$ be the space of probability distributions on $\mathbb{C}$ that have A as expectation and a finite second moment, endowed with a complete metric space structure by the Wasserstein distance. Let $\lambda \in \mathbb{C}$ be any root of the characteristic polynomial (4.1) such that $\Re(\lambda) > \frac{1}{2}$. Then,*

(i) *Each of the two equations*

$$W \overset{\mathcal{L}}{=} B_1^{\lambda} W^{(1)} + B_2^{\lambda} W^{(2)}$$

where $W^{(1)}$ and $W^{(2)}$ are independent copies of W, and where (B_1, B_2) is a random vector, independent of $W^{(1)}$ and $W^{(2)}$, Dirichlet distributed with parameters (m, m), and

$$W \stackrel{\mathcal{L}}{=} B^{\lambda}\left(W^{(1)} + W^{(2)}\right)$$

where $W^{(1)}$ and $W^{(2)}$ are independent copies of W, and where B is independent of $W^{(1)}$ and $W^{(2)}$, Beta distributed with parameters (m, m),

have a unique solution in $\mathcal{M}_2(A)$.

(ii) *For $m \geq 60$, the variable $W_m^{\mathrm{B-tree}}$, defined in Corollary 4.4, is the unique solution of (6.3) having $\dfrac{m!}{\Gamma(m+\lambda_2)}$ as expectation and a finite second moment.*

(iii) *For $m \geq 60$, the variable $\mathcal{W}^{CT}$, defined in (5.6), is the unique solution of (6.2) having m as expectation and a finite second moment.*

6.4. *Cascades and exponential moments.* Let $\lambda \in \mathbb{C}$ be any root of the characteristic polynomial (4.1) such that $\Re(\lambda) > \frac{1}{2}$ and let B be a Beta distribution with parameters (m, m). A simple computation leads to $2\mathbb{E}\left(B^{\lambda}\right) = 1$. This is coherent with equation

$$W \stackrel{\mathcal{L}}{=} B^{\lambda}\left(W^{(1)} + W^{(2)}\right). \tag{6.4}$$

Moreover, for any positive real s,

$$2\mathbb{E}\left(B^{s}\right) = \frac{(2m)\ldots(m+1)}{(2m-1+s)\ldots(m+s)} < 1 \Longleftrightarrow s > 1.$$

Consequently, when $2\Re(\lambda) > 1$, one has

$$2\mathbb{E}\left(|B^{\lambda}|^2\right) < 1. \tag{6.5}$$

Theorem 6.5 below states that any solution W of Equation (6.4) admits exponential moments in a neighbourhood of 0, so that the moment exponential generating series of W defines an analytic function in a neighbourhood of the origin. Another consequence is that the law of W is determined by its moments.

The proof relies on a Mandelbrot's cascade here defined in a complex setting (see Barral et al., 2010 for complex Mandelbrot's cascades).

To lighten the notations, denote for a while $A := B^{\lambda}$ and let $A_u, u \in U$ be independent copies of A, indexed by all finite sequences of 0 and 1:

$$u = u_1 \ldots u_n \in U := \bigcup_{n \geq 1} \{0,1\}^n.$$

Let $Y_0 = m$, $Y_1 = 2mA$ and for $n \geq 2$,

$$Y_n = \sum_{u_1 \ldots u_{n-1} \in \{0,1\}^{n-1}} 2mAA_{u_1}A_{u_1u_2}\ldots A_{u_1\ldots u_{n-1}}.$$

By the branching property, and using $2\mathbb{E}A = 1$, it is easy to see that $(Y_n)_n$ is a martingale with expectation m. This martingale has been studied by many authors in the real-valued random variable case, especially in the context of Mandelbrot's cascades, see for example Liu (2001) and the references therein. It can be easily seen that

$$Y_{n+1} = B^{\lambda}\left(Y_n^{(1)} + Y_n^{(2)}\right) \tag{6.6}$$

where $Y_n^{(1)}$ and $Y_n^{(2)}$ are independent of each other and independent of B^λ and each has the same distribution as Y_n. Therefore for $n \geq 1$, Y_n is square-integrable and

$$\operatorname{Var} Y_{n+1} = 2\mathbb{E}|B^\lambda|^2 \operatorname{Var} Y_n + (4\mathbb{E}|B^\lambda|^2 - 1)$$

where $\operatorname{Var} X = \mathbb{E}\left(|X - \mathbb{E}X|^2\right)$ denotes the variance of X. Since $2\mathbb{E}|B^\lambda|^2 < 1$, the martingale $(Y_n)_n$ is bounded in L^2, so that the following result holds.

Lemma 6.3. *Let $\lambda \in \mathbb{C}$ be any root of the characteristic polynomial (4.1) such that $\Re(\lambda) > \frac{1}{2}$ and let B be a Beta distribution with parameters (m, m). When $n \to +\infty$,*

$$Y_n \to Y_\infty \text{ a.s. and in } L^2,$$

where Y_∞ is a (complex-valued) random variable with variance

$$\operatorname{Var}(Y_\infty) = \frac{4\mathbb{E}|B^\lambda|^2 - 1}{1 - 2\mathbb{E}|B^\lambda|^2}.$$

Notice that, passing to the limit in (6.6) gives a new proof of the existence of a solution W of Eq. (6.4) with a given expectation and finite second moment whenever $\Re(\lambda) > 1/2$. From Section 6.3, we have the uniqueness of solution of this equation so that Theorem 6.5 below will be proved as soon as it holds for Y_∞.

Lemma 6.4. *There exist some constants $C > 0$ and $\varepsilon > 0$ such that for all $t \in \mathbb{C}$ with $|t| \leq \varepsilon$, we have*

$$\mathbb{E}e^{\langle t, Y_\infty \rangle} \leq e^{m\Re(t) + C|t|^2}. \tag{6.7}$$

Proof: By Fatou lemma, it is sufficient to prove the existence of $C > 0$ and $\varepsilon > 0$ such that for all $t \in \mathbb{C}$ with $|t| \leq \varepsilon$, and for every integer n,

$$\mathbb{E}e^{\langle t, Y_n \rangle} \leq e^{m\Re(t) + C|t|^2}. \tag{6.8}$$

Denote $\varphi_n(t) := \mathbb{E}e^{\langle t, Y_n \rangle}$ and notice that $\varphi_{n+1}(t) = \mathbb{E}\left(\varphi_n^2(tB^{\overline{\lambda}})\right)$ thanks to Equation (6.6), allowing to prove (6.8) by recursion on $n \geq 0$. For $n = 0$,

$$\varphi_0(t) := \mathbb{E}e^{\langle t, Y_0 \rangle} = e^{m\Re(t)}$$

and by the recursion assumption,

$$\varphi_{n+1}(t) \leq \mathbb{E}\left(e^{2C|t|^2|B^\lambda|^2 + 2m\Re\left(tB^{\overline{\lambda}}\right)}\right) = e^{m\Re(t) + C|t|^2} f(t_1, t_2)$$

where for any $t \in \mathbb{C}$, written $t = t_1 + it_2$ with $t_1, t_2 \in \mathbb{R}$,

$$f(t_1, t_2) = \mathbb{E}\left(e^{C|t|^2(2|B^\lambda|^2 - 1) + 2m\Re\left(tB^{\overline{\lambda}}\right) - m\Re(t)}\right),$$

so that it is sufficient to prove that $(0, 0)$ is a local maximum of f. Writing $\lambda = \sigma + i\tau$, with $\sigma, \tau \in \mathbb{R}$,

$$f(t_1, t_2) = \mathbb{E}\exp\left[C(t_1^2 + t_2^2)(2B^{2\sigma} - 1) + 2mB^\sigma(t_1\cos(\tau) + t_2\sin(\tau)) - mt_1\right].$$

Remembering $2\mathbb{E}\left(B^\lambda\right) = 1$, which means $2\mathbb{E}\left(B^\sigma\cos(\tau)\right) = 1$ and $\mathbb{E}\left(B^\sigma\sin(\tau)\right) = 0$, we get that the first derivatives vanish at $(0, 0)$ which is a critical point. Moreover,

the calculation of the second partial derivatives gives

$$\begin{aligned}\frac{\partial^2 f}{\partial t_1}(0,0) &= \mathbb{E}\left[(2mB^\sigma\cos(\tau)-m)^2+2C\left(2B^{2\sigma}-1\right)\right],\\ \frac{\partial^2 f}{\partial t_2}(0,0) &= \mathbb{E}\left[(2mB^\sigma\sin(\tau))^2+2C\left(2B^{2\sigma}-1\right)\right],\\ \frac{\partial^2 f}{\partial t_1\partial t_2}(0,0) &= \mathbb{E}\left(2mB^\sigma\cos(\tau)-m\right)\left(2mB^\sigma\sin(\tau)\right).\end{aligned}$$

By (6.5), $\mathbb{E}\left(2B^{2\sigma}-1\right)<0$, so that the Hessian matrix at $(0,0)$ is definite negative for $C>0$ large enough which implies that $(0,0)$ is a local maximum of f. □

The following theorem is a direct consequence of Lemma 6.4, like in Chauvin et al. (2014).

Theorem 6.5. *Let $\lambda\in\mathbb{C}$ be a root of the characteristic polynomial (4.1) with $\Re(\lambda)>1/2$ and let W be a solution of Eq. (6.4). There exist some constants $C>0$ and $\varepsilon>0$ such that for all $t\in\mathbb{C}$ with $|t|\le\varepsilon$,*

$$\mathbb{E}e^{\langle t,W\rangle}\le e^{m\Re(t)+C|t|^2}\quad\text{and}\quad\mathbb{E}e^{|tW|}\le 4e^{m|t|+2C|t|^2}. \tag{6.9}$$

6.5. *Laplace transform.* Theorem 6.5 above concerning W^{CT} and Theorem 7 (ii) in Mailler (2014) (which establishes that the Laplace series of W^{DT} has an infinite radius of convergence) answer the question of the convergence of the Laplace series of W^{DT} and W^{CT}. Nevertheless, a natural investigation consists in searching more information about these Laplace transforms coming from the smoothing equations.

Indeed, the dislocation equations (6.1) lead to a system of differential equations on the Laplace transforms

$$\forall k=1,2,\dots,m,\qquad \varphi_k(z):=\mathbb{E}\left(e^{<z,W_k>}\right), \tag{6.10}$$

where we recall that W_k is the limit random variable W^{CT} of the continuous-time fringe node process $(L(t)^{e_k})_t$ that starts with one particle of type k. Using the independence between the splitting times τ_k (which are exponentially distributed, with parameter $m+k-1$) and the W_k, for $k=1,2,\dots,m-1$,

$$\varphi_k(z)=\int_0^{+\infty}\varphi_{k+1}\left(ze^{-\overline{\lambda_2}t}\right)(m+k-1)e^{-(m+k-1)t}dt,$$

and after a change of variable, and derivation, for $k=1,2,\dots,m-1$,

$$\frac{m+k-1}{\overline{\lambda_2}}\varphi_k(z)+z\varphi_k'(z)=\varphi_{k+1}(z),$$

and for $k=m$,

$$\frac{2m-1}{\overline{\lambda_2}}\varphi_m(z)+z\varphi_m'(z)=\varphi_1^2(z).$$

Thanks to a convenient change of function, a simple calculation gives the following theorem about the Laplace transforms of the W_k, for $k=1,2,\dots,m$.

Theorem 6.6. *For $k=1,2,\dots,m$, let φ_k be the Laplace transform of W_k defined in (6.10), and let*

$$\psi_k(z):=\left(-\overline{\lambda_2}\right)^{m+k-1}\frac{\varphi_k\left(z^{-\overline{\lambda_2}}\right)}{z^{m+k-1}},$$

(for any determination of the logarithm). Then the functions ψ_k satisfy the simple differential system

$$\begin{cases} \psi_k' = \psi_{k+1}, & \forall k \in \{1, \dots, m-1\}, \\ \psi_m' = \psi_1^2. \end{cases}$$

In particular, ψ_1 is a solution of the differential equation

$$y^{(m)} = y^2.$$

6.6. *Density and support.* Liu's method has been developped in Liu (1999, 2001) for positive real-valued random variables solution of smoothing equations of the same type as (6.2) or (6.3). Adapting this method to $\mathbb{C}$-valued random variables, Mailler (2014) gets the support and the existence of a density for the limit law of a d-color Pólya urn. The theorem below is a particular case.

Theorem 6.7. *Let $m \geq 2$. Let $V \in \mathbb{N}^m$ be a non zero vector and let W^{DT} and W^{CT} be the W-distributions of the respective discrete-time and continuous-time fringe node processes having V as initial composition (see (ii) in Theorem 4.3 and Theorem 5.2). Then*

(i) the supports of W^{DT} and W^{CT} are both the whole complex plane $\mathbb{C}$;

(ii) W^{DT} and W^{CT} are absolutely continuous relatively to Lebesgue's measure on $\mathbb{C}$;

(iii) as $|t| \to \infty$, $\mathbb{E}e^{i\langle t, W^{DT}\rangle} = O(|t|^{-a})$ for each $a \in \left]0, \dfrac{m}{\Re(\lambda_2)}\right[$, and the same is true for the Fourier transform of W^{CT} as well.

6.7. *Perspectives.* Some open questions remain about W^{DT} and W^{CT}, let us say W:

- can the W distribution be expressed by means of usual distributions? Same question for $|W|$ and $\mathrm{Arg}(W)$?
- how heavy are the tails of W?
- what is the order of magnitude of W's p-th moment as p tends to $+\infty$?

7. Simulations

Let us here summarize and illustrate the asymptotic results concerning the fringe of a random B-trees. We only show the behaviour of the gap process $(G_n)_n$ and illustrate Theorem 4.1; nevertheless, the simulations would be analogous for the fringe node process $(\mathcal{L}_n)_n$ to illustrate Corollary 4.4.

In all the simulations below, sequences of 10^7 random keys have been inserted in a B-tree for different value of the parameter m. Notice that in "real life" computer science implementations, m is most of the time taken around 100 or more.

7.1. *Simulations of G_n.* Figures 7.7 and 7.8 represent the trajectories of three coordinates of the random vector G_n: for any given value $m = 10$, 30, 55, 65, 100 or 237 of the parameter, we make one random drawing of a sequence of 10^7 keys and insert them in a B-tree. On the pictures, the x-axis represents the time $n \in \{0, \dots, 10^7\}$ while the y-axis represents the number $G_n^{(k)}$ of gaps of type k for $k = 1$, $\lfloor m/2 \rfloor$ and m. In each case, the picture illustrates the almost sure asymptotics $G_n \sim nv_1$ when n tends to infinity (remember that v_1 is a non random m-dimensional vector).

In Figure 7.7, m is *small* ($m = 10, 30, 55$). One can already catch sight of the Gaussian fluctuations around the deterministic vector nv_1. Notice that the variance of the Gaussian limit increases with m, so that the amplitude of the fluctuation becomes more visible for $m = 30$ and even more for $m = 55$.

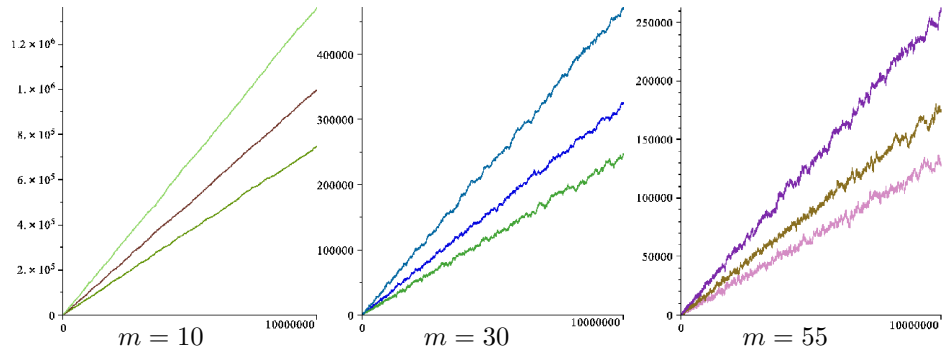

FIGURE 7.7. Simulations for 3 coordinates of the gap process $(G_n)_n$ for *small* m.

In Figure 7.8, m is *large* ($m = 65, 100, 237$). On can see the almost sure oscillations around nv_1 appear and become more visible when m grows. Notice that they are particularly clear for $m = 237$, which is the threshold value when the *third* largest real part of the roots of χ_m becomes larger than $\frac{1}{2}$. See the Appendix for more details.

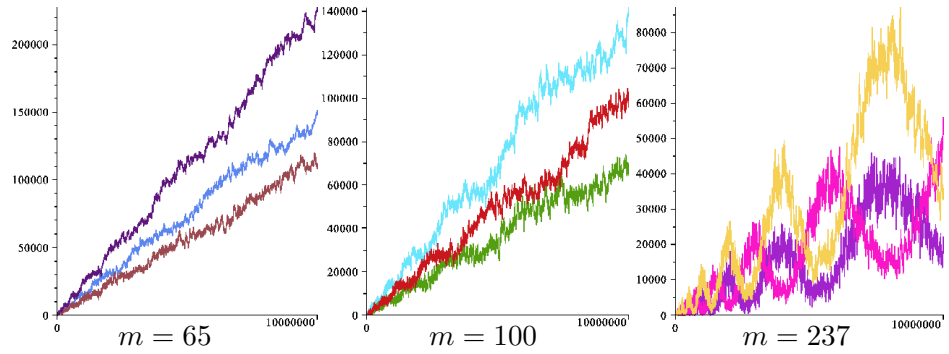

FIGURE 7.8. Simulations for 3 coordinates of the gap process $(G_n)_n$ for *large* m.

Of course, one can make similar graphs for trajectories of the vector $\frac{G_n}{n}$ which converges to the deterministic vector v_1. This is done in Figure 7.9 where the convergence can be seen on the three drawn coordinates. Once more, the fluctuations around the limit v_1 are of different nature depending on $m \leq 59$ or $m \geq 60$, which is also illustrated on this figure. In particular, on can see the "$\cos \log n$" almost sure oscillations arise when $m \geq 60$ and become more evident when m increases.

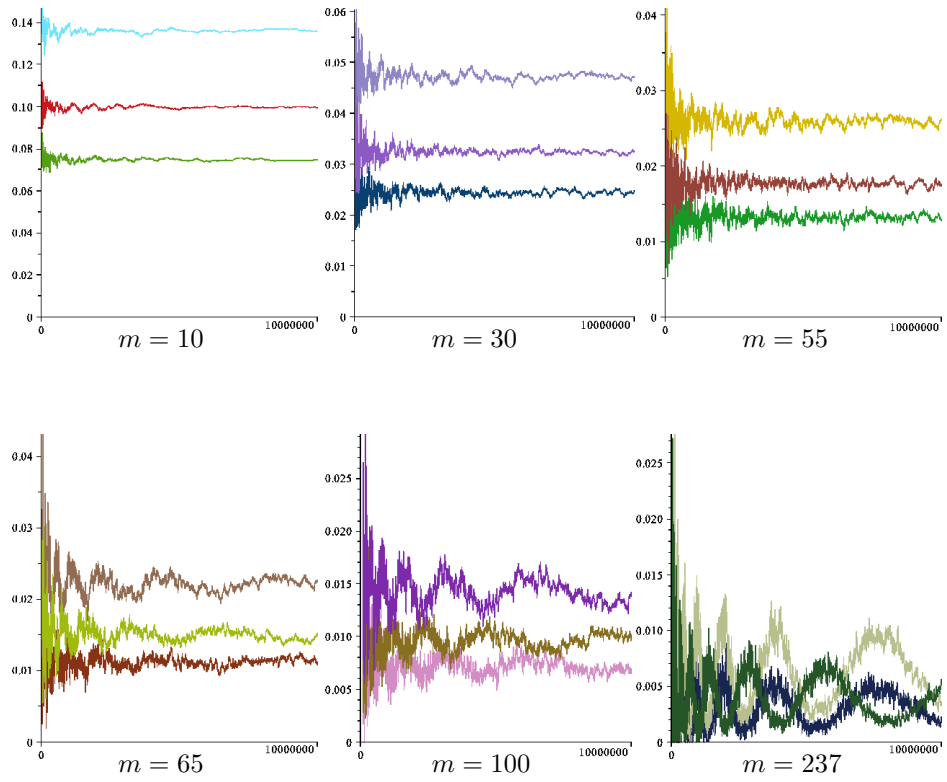

FIGURE 7.9. Simulations of 3 coordinates of $\frac{G_n}{n}$ for *small* and *large* values of m.

7.2. *Simulations of G_n after scaling.* A second kind of simulations focus on the possible scalings of the centered gap process $(G_n - nv_1)_n$. In order to get convergence, according to Theorem 4.1, one has to divide $G_n - nv_1$ by $\sqrt{n}$ when $m \leq 59$ and by n^{σ_2} when $m \geq 60$. Figures 7.10 and 7.11 represent trajectories of the median coordinate (the $\lfloor m/2 \rfloor$-th) of the normalized vector process. Hereunder, X_n denotes this median coordinate $X_n = G_n^{\lfloor m/2 \rfloor}$.

Figure 7.10 deals with *small* values of m, namely $m = 10, 30, 55$ again. On the x-axis, time $n \in \{0, \ldots, 10^7\}$; on the y-axis, the normalized coordinate $\dfrac{X_n - nv_1^{\lfloor m/2 \rfloor}}{\sqrt{n}}$ which converges in distribution to a normal law. Note that even if the random vector $\dfrac{G_n - nv_1}{\sqrt{n}}$ converges in distribution, it almost surely diverges, which is illustrated by its brownian-like trajectory. One can refer to Gouet (1993) for more details on this continuous type process limit.

Figure 7.11 deals with $m = 65, 100, 237$ which are *large* values of m. On the y-axis: the normalized coordinate $\dfrac{X_n - nv_1^{\lfloor m/2 \rfloor}}{n^{\sigma_2}}$, which is almost surely equivalent

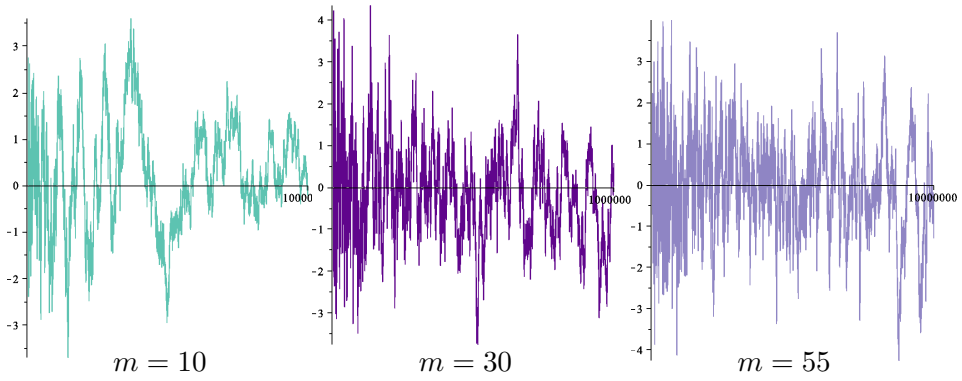

FIGURE 7.10. Simulations of one coordinate of G_n after normalisation, for *small* values of m.

to some $\rho \cos(\tau_2 \log n + \varphi)$ when n tends to infinity, where ρ is a positive random variable (random amplitude), φ a $[0, 2\pi[$-valued random variable (random phase) and τ_2 the imaginary part of the complex eigenvalue $\lambda_2 = \sigma_2 + i\tau_2$. The random variables ρ and φ are proportional to the module and the argument of the complex-valued random variable W_m in Theorem 4.1.

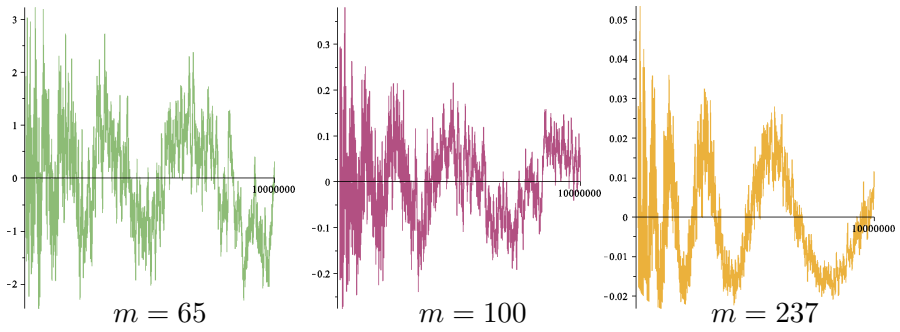

FIGURE 7.11. Simulations of one coordinate of G_n after normalization, for *large* values of m.

8. Appendix. The phase transition and $\sigma_2(m)$

The phase transition that occurs for B-trees with parameter m relies on the roots of the characteristic polynomial

$$\chi_m(X) = \prod_{k=m}^{2m-1} (X + k) - \frac{(2m)!}{m!}.$$

Denote by $\lambda_2 = \lambda_2(m)$ the root of χ_m having the second largest real part and a positive imaginary part. Denote also $\sigma_2 = \sigma_2(m)$ the real part of $\lambda_2(m)$.

As shown in Section 4, the B-tree admits a Gaussian central limit theorem when $m \leq 59$ (small régime) whereas it admits an almost sure nonnormal fluctuation

term of order $n^{\sigma_2(m)}$ around the drift when $m \geq 60$ (large régime). Coming from Pólya urn theory, this asymptotic behaviour depends on whether $\sigma_2(m) < 1/2$ (small régime) or $\sigma_2(m) > 1/2$ (large régime).

Let F the two-variable meromorphic function defined by

$$F(x,y) = \frac{\Gamma(x+2y)\,\Gamma(1+y)}{\Gamma(1+2y)\,\Gamma(x+y)}$$

where Γ denotes Euler's Gamma function. For a given $m \geq 2$, $\lambda_2 = \lambda_2(m) = \sigma_2 + i\tau_2$ is the root of equation $F(X,m) = 1$ having the the second largest real part σ_2 (the first one being reached by the evident root 1) and a positive imaginary part τ_2. Denote by ψ the classical Digamma function, the logarithmic derivative of Euler's Gamma. Since $\frac{\partial}{\partial x}F(x,1/y) = \psi(x+2/y) - \psi(x+1/y) = \log 2 + O(y)$ as y tends to 0, the analytic implicit function theorem shows that $\lambda_2(m)$ is an analytic function of $1/m$ as m tends to $+\infty$. Using the expansion

$$\log\Gamma(z) = z\log z - z - \frac{1}{2}\log z + \frac{1}{2}\log 2\pi + O\left(\frac{1}{z}\right) \mod 2i\pi$$

as $|z|$ tends to infinity (the *mod* coming from the determination of the logarithm), writing $\lambda_2(m)$ as a power series in $1/m$ and putting the first terms of this expansion of in the equation $\log F(\lambda_2(m), m) = 0 \mod 2\mathrm{i}\pi$, one gets by identification

$$\begin{cases} \sigma_2(m) = 1 - \dfrac{\pi^2}{\log^3 2} \times \dfrac{1}{m} + O\left(\dfrac{1}{m^2}\right) \\ \\ \tau_2(m) = \dfrac{2\pi}{\log 2} + \dfrac{\pi}{2\log^2 2} \times \dfrac{1}{m} + O\left(\dfrac{1}{m^2}\right) \end{cases}$$

as m tends to infinity. The graph of the function $m \mapsto \sigma_2(m)$ is given in Figure 8.12. Numerical values of the expansions give $\sigma_2 \approx 1 - 29.63/m + \ldots$ while $\tau_2 \approx 9.06 + 3.27/m + \ldots$

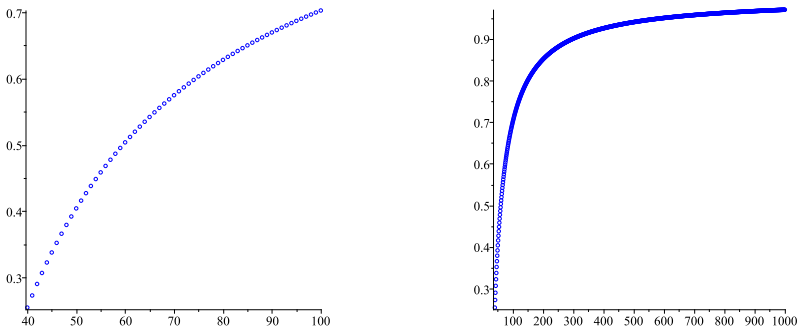

FIGURE 8.12. The graph of the sequence $m \mapsto \sigma_2(m)$.

Moreover, the numerical values of σ_2 around $m = 60$ are the following ones in the next table, showing more accurately that $\sigma_2(m) < 1/2$ if, and only if $m \leq 59$. These numerical values have been computed applying the Newton method to the χ_m function starting from the point $0.5 + 9.0i$, as suggested by the above expansions of σ_2 and τ_2.

m	$\sigma_2(m)$
57	0.4775726941
58	0.4866133472
59	0.4953467200
60	0.5037882018
61	0.5119521623
62	0.5198520971

In order to justify the choice of $m = 237$ in our drawings, denote by $\sigma_3(m)$ the third largest real part of the roots of χ_m. The threshold value when $\sigma_3(m)$ becomes larger than $\frac{1}{2}$ is $m = 237$. It is a matter of fact that, using general statements on Pólya urns, a second almost sure phenomenon with magnitude n^{λ_3} is added to the one we describe in Theorem 4.1 as soon as $m \geq 238$. That is the reason why the above figures have been selected for $m = 237$; indeed, for m increasing from 60 until 237, the asymptotic expansion of G_n contains the oscillating term of amplitude $n^{\sigma_2(m)}$ more and more visible compared to Brownian terms in $n^{\frac{1}{2}}$. For $m > 237$, the second oscillating term of amplitude $n^{\sigma_3(m)}$ appears, making the $n^{\sigma_2(m)}$ oscillation less visible. The numerical values of $\sigma_3(m)$ around $m = 237$ are the following ones.

m	$\sigma_3(m)$
236	0.4971039325
237	0. 4992277960
238	0.5013338161
239	0.5034221856

The threshold value $m = 237$ could be called a "second phase transition" as in Chern and Hwang (2001, Section 9), where they conjectured "further phase transitions" or different "convergence rates" about m-ary search trees.

Acknowledgements

The authors kindly thank Karine Zeitouni for valuable discussions on B-trees and Andrea Sportiello for insightful comments.

References

K. B. Athreya and S. Karlin. Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.* **39**, 1801–1817 (1968). MR0232455.

K. B. Athreya and P. E. Ney. *Branching processes*. Springer-Verlag, New York-Heidelberg (1972). Die Grundlehren der mathematischen Wissenschaften, Band 196. MR0373040.

R. A. Baeza-Yates. Fringe analysis revisited. *ACM Comput. Surv.* **27** (1), 109–119 (1995). DOI: 10.1145/214037.214103.

J. Barral, X. Jin and B. Mandelbrot. Convergence of complex multiplicative cascades. *Ann. Appl. Probab.* **20** (4), 1219–1252 (2010). MR2676938.

R. Bayer. Binary B-trees for virtual memory. In *Proceedings of the 1971 ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control*, pages 219–235. ACM (1971). DOI: 10.1145/1734714.1734731.

R. Bayer and E. M. McCreight. Organization and maintenance of large ordered indexes. *Acta Informatica* **1** (3), 173–189 (1972). DOI: 10.1007/BF00288683.

J. Bertoin. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge (2006). ISBN 978-0-521-86728-3; 0-521-86728-2. MR2253162.

J. D. Biggins and A. E. Kyprianou. Fixed points of the smoothing transform: the boundary case. *Electron. J. Probab.* **10**, no. 17, 609–631 (2005). MR2147319.

B. Chauvin, Q. Liu and N. Pouyanne. Support and density of the limit m-ary search trees distribution. In *23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12)*, Discrete Math. Theor. Comput. Sci. Proc., AQ, pages 191–199. Assoc. Discrete Math. Theor. Comput. Sci., Nancy (2012). MR2957331.

B. Chauvin, Q. Liu and N. Pouyanne. Limit distributions for multitype branching processes of m-ary search trees. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** (2), 628–654 (2014). MR3189087.

B. Chauvin, C. Mailler and N. Pouyanne. Smoothing equations for large Pólya urns. *J. Theoret. Probab.* **28** (3), 923–957 (2015). MR3413961.

B. Chauvin, N. Pouyanne and R. Sahnoun. Limit distributions for large Pólya urns. *Ann. Appl. Probab.* **21** (1), 1–32 (2011). MR2759195.

H.-H. Chern and H.-K. Hwang. Phase changes in random m-ary search trees and generalized quicksort. *Random Structures Algorithms* **19** (3-4), 316–358 (2001). MR1871558.

T. Cormen, C. Leiserson, R. Rivest and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA; McGraw-Hill Book Co., Boston, MA, second edition (2001). ISBN 0-262-03293-7. MR1848805.

J. A. Fill and N. Kapur. Limiting distributions for additive functionals on Catalan trees. *Theoret. Comput. Sci.* **326** (1-3), 69–102 (2004). MR2094243.

P. Flajolet, P. Dumas and V. Puyhaubert. Some exactly solvable models of urn process theory. In *Fourth Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities*, Discrete Math. Theor. Comput. Sci. Proc., AG, pages 59–118. Assoc. Discrete Math. Theor. Comput. Sci., Nancy (2006). MR2509623.

R. Gouet. Martingale functional central limit theorems for a generalized Pólya urn. *Ann. Probab.* **21** (3), 1624–1639 (1993). MR1235432.

P. Hennequin. *Analyse en moyenne d'algorithmes: Tri rapide et arbres de recherche*. Ph.D. thesis, L'École Polytechnique (1991).

S. Janson. Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.* **110** (2), 177–245 (2004). MR2040966.

N. L. Johnson and S. Kotz. *Urn models and their application*. John Wiley & Sons, New York-London-Sydney (1977). An approach to modern discrete probability theory, Wiley Series in Probability and Mathematical Statistics. MR0488211.

M. Knape and R. Neininger. Pólya urns via the contraction method. *Combin. Probab. Comput.* **23** (6), 1148–1186 (2014). MR3265841.

Robert L. Kruse and Alexander J. Ryba. *Data Structures and Program Design in C++*. Prentice-Hall, Inc. (1999). ISBN 0-13-768995-0.

Q. Liu. Fixed points of a generalized smoothing transformation and applications to the branching random walk. *Adv. in Appl. Probab.* **30** (1), 85–112 (1998). MR1618888.

Q. Liu. Asymptotic properties of supercritical age-dependent branching processes and homogeneous branching random walks. *Stochastic Process. Appl.* **82** (1), 61–87 (1999). MR1695070.

Q. Liu. Asymptotic properties and absolute continuity of laws stable by random weighted mean. *Stochastic Process. Appl.* **95** (1), 83–107 (2001). MR1847093.

H. M. Mahmoud. The size of random bucket trees via urn models. *Acta Inform.* **38** (11-12), 813–838 (2002). MR1943199.

H. M. Mahmoud. *Pólya urn models*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL (2009). ISBN 978-1-4200-5983-0. MR2435823.

C. Mailler. Describing the asymptotic behaviour of multicolour Pólya urns via smoothing systems analysis. *ArXiv Mathematics e-prints* (2014). arXiv: 1407.2879v1.

R. Neininger and L. Rüschendorf. A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.* **14** (1), 378–418 (2004). MR2023025.

R. Neininger and L. Rüschendorf. A survey of multivariate aspects of the contraction method. *Discrete Math. Theor. Comput. Sci.* **8** (1), 31–56 (electronic) (2006). MR2247515.

N. Pouyanne. An algebraic approach to Pólya processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** (2), 293–323 (2008). MR2446325.

A. C. C. Yao. On random $2-3$ trees. *Acta Informat.* **9** (2), 159–170 (1977/78). MR0660704.