CONTINUOUS TIME DIGITAL SEARCH TREE AND A BORDER AGGREGATION MODEL

SVANTE JANSON AND DEBLEENA THACKER

Abstract. We consider the continuous-time version of the random digital search tree, and construct a coupling with a border aggregation model as studied in Thacker and Volkov (2018), showing a relation between the height of the tree and the time required for aggregation. This relation carries over to the corresponding discrete-time models. As a consequence we find a very precise asymptotic result for the time to aggregation, using recent results by Drmota et al. (2020) for the digital search tree.

1. Introduction

A *digital search tree* $\mathcal{T}_n$ is a binary tree constructed from a sequence of n binary strings (called *items* or *keys*). (See Section 2 for details, as well as for definitions of other concepts used below.) We consider here only the case when the items are i.i.d. (independent, identically distributed) random infinite binary strings, and furthermore, in each string the digits are independent $\mathrm{Be}(1/2)$ random variables, i.e., 0 or 1 with probability $\frac{1}{2}$ each. (See Section 6 for the b-ary case.) Digital search trees are among the fundamental objects of study in computer science algorithms and have been studied by many authors, see e.g. [1; 3; 4; 5; 6; 8; 10; 11; 15].

Our main concern is with a continuous-time version of the digital search tree, studied also by Aldous and Shields [1]. This can be defined by assuming that an infinite sequence (W_n) of items arrive at random times that are given by a Poisson process; we then let $\mathfrak{T}_t$ be the digital search tree defined by the strings arriving up to time t. The continuous-time version is thus a Poissonization of the standard version. A simple but central result (Theorem 3.5 and [1]) is that the continuous-time digital search tree $\mathfrak{T}_t$ also can be defined in two other ways that turn out to be equivalent; in particular, the continuous-time digital search tree is equivalent to *first-passage percolation* on the infinite binary tree, with the passage times of the edges exponentially distributed such that the passage time of an edge between nodes of depth $k-1$ and k has expectation 2^k.

Our main result couples the continuous-time digital search tree and a *border aggregation model* on a binary tree studied by Thacker and Volkov [17]. In this model, we fix $K \geqslant 1$ and consider the complete binary tree T_K of height K. We recursively define a collection of randomly growing subset of *sticky* nodes S_n, such that S_0 is the set of the 2^K nodes of depth K. S_n is obtained from S_{n-1} as follows: A particle is released from the root, and

Date: 28 April, 2020; revised 1 March, 2021.
Partly supported by the Knut and Alice Wallenberg Foundation.

performs a symmetric (directed) random walk until it comes to a neighbour v_n of S_{n-1}. The random walk now stops, and the node v_n becomes "sticky"; in other words, $S_n := S_{n-1} \cup \{v_n\}$. This is repeated until the root o is sticky. Let ξ_K be the random number of particles to be released until the root o is sticky. We define also a continuous-time version of the border aggregation model by assuming that particles start from the root at times given by a Poisson process (and that the random walk itself takes no time); let Ξ_K be the random time that the root gets sticky in the continuous-time border aggregation model.

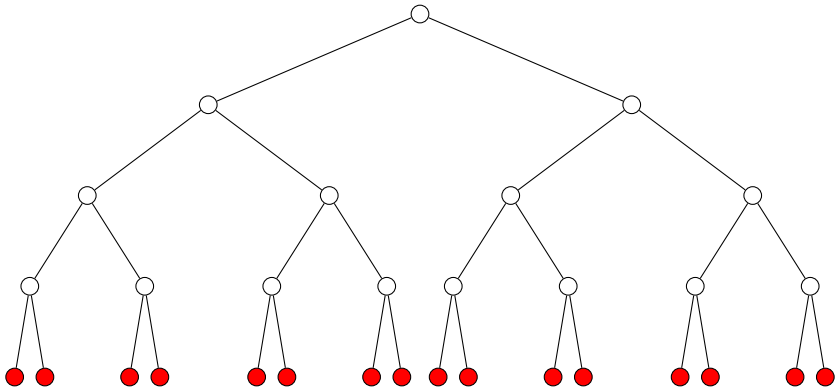

FIGURE 1. Binary tree T_K with $K = 4$, and the red nodes denoting S_0

Note that the digital search tree and the border aggregation model grow in opposite directions: the digital search tree grows from the root downwards, while the border aggregation model grows from the starting boundary at depth K up towards the root. Nevertheless, they are connected by a kind of duality, and we show that the time ξ_K or Ξ_K taken by the border aggregation model equals in distribution the time the (discrete or continuous-time, respectively) digital search tree reaches (external) height K (Theorem 4.1). Equivalently, we have the following results, where $h_{\mathsf{e}}(T)$ denotes the external height of a tree T as defined in (2.1).

Theorem 1.1. *The following equalities hold.*

(i) *(Discrete time.) For any $K \geqslant 1$ and $n \geqslant 0$,*

$$\mathbb{P}\big(\xi_K \leqslant n\big) = \mathbb{P}\big(h_{\mathsf{e}}(\mathcal{T}_n) \geqslant K\big). \tag{1.1}$$

(ii) *(Continuous time.) For any $K \geqslant 1$ and $t \geqslant 0$,*

$$\mathbb{P}\big(\Xi_K \leqslant t\big) = \mathbb{P}\big(h_{\mathsf{e}}(\mathfrak{T}_t) \geqslant K\big). \tag{1.2}$$

We show this using the continuous-time versions; the result then easily transfers to discrete time too.

Asymptotic properties of the height $h_{\mathsf{e}}(\mathcal{T}_n)$ of digital search trees have been studied by several authors [1; 3; 5; 10]. In particular, very precise results are proved by Drmota, Fuchs, Hwang and Neininger [5]. We use these results and Theorem 1.1 to obtain the following result on the distribution of ξ_K, which improves on the bounds $2^{K-2\sqrt{K}+\mathcal{O}(K^{-1/2})} \leqslant \xi_K \leqslant 2^{K-1+o(1)}$ w.h.p. shown in [17, Theorem 5].

Theorem 1.2. *As $K \to \infty$,*

$$\log_2 \xi_K = K - \sqrt{2K} + \frac{1}{2}\log_2 K - \frac{1}{\log 2} + \frac{\log_2 K}{4\sqrt{2K}} + \mathcal{O}_{\mathrm{p}}\Big(\frac{1}{\sqrt{K}}\Big), \tag{1.3}$$

where $\mathcal{O}_{\mathrm{p}}(\cdot)$ is as defined in Section 2.1.

For convenience, let

$$m_K = 2^{K-\sqrt{2K}+\frac{1}{2}\log_2 K - \frac{1}{\log 2} + \frac{\log_2 K}{4\sqrt{2K}}}. \tag{1.4}$$

Then, Theorem 1.2 says that $\log_2 \xi_K = \log_2 m_K + \mathcal{O}_{\mathrm{p}}\big(1/\sqrt{K}\big)$, or, equivalently,

$$\xi_K = m_K\big(1 + \mathcal{O}_{\mathrm{p}}\big(K^{-1/2}\big)\big). \tag{1.5}$$

Conjecture 1.3. *We conjecture that also*

$$\mathbb{E}\,\xi_K = m_K\big(1 + \mathcal{O}\big(K^{-1/2}\big)\big) = 2^{K-\sqrt{2K}+\frac{1}{2}\log_2 K - \frac{1}{\log 2} + \frac{\log_2 K}{4\sqrt{2K}} + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)}. \tag{1.6}$$

We have not been able to prove (1.6), see Remark 5.1, but as a corollary of Theorem 1.2 and tail estimates by Drmota [3], we show the following cruder estimate.

Theorem 1.4. *As $K \to \infty$,*

$$\mathbb{E}\,\xi_K = \mathbb{E}\,\Xi_K = \big(1 + o(1)\big)m_K = 2^{K-\sqrt{2K}+\frac{1}{2}\log_2 K - \frac{1}{\log 2} + o(1)}. \tag{1.7}$$

The border aggregation model was introduced as *internal erosion* by Levine and Peres [13]. In [17], the border aggregation model was studied on a variety of graphs, and several interesting results were obtained. One reason of the interest in the border aggregation model is its possible connections to other interesting models in statistical physics, in particular the classical *diffusion limited aggregation* (DLA) [18; 9; 2], and *internal diffusion limited aggregation* (IDLA) [12; 14; 16]. It is conjectured [13] that on $\mathbb{Z}^2$ DLA and the border aggregation model are "inversions" of each other in some sense; however, no rigorous results are known. (Nevertheless, [17] uses bounds obtained in [9] for DLA to obtain results for the border aggregation model.) Note that the digital search tree can be regarded as IDLA on the infinite binary tree (see Section 2.5); moreover, it can also be regarded as DLA on the same infinite binary tree, see Barlow, Pemantle, Perkins [2, Lemma 1.3]. Thus, our results show a connection between the border aggregation model and IDLA or DLA on trees. (Note that on $\mathbb{Z}^d$, IDLA is very different from DLA and the border aggregation model, with asymptotically a round shape [12].)

The rest of the paper is organized as follows. Section 2 contains definitions and other preliminaries. Section 3 gives the equivalence of the different constructions of the continuous-time digital search tree. Section 4 contains the coupling of the digital search tree and the border aggregation model, leading to the proof of Theorem 1.1, and then Section 5 gives the proofs of Theorems 1.2 and 1.4. Section 6 discusses briefly extensions to the b-ary case.

2. Preliminaries

We recall some standard notation, adding some perhaps less standard details.

2.1. **General.** $\mathrm{Exp}(\lambda)$ denotes an exponential distribution with *rate* λ, i.e., with the density function $\lambda e^{-\lambda x}$, $x>0$, and thus the expectation $1/\lambda$.

$\mathcal{O}_{\mathrm{p}}(a_n)$, where a_n is a given positive sequence, denotes some sequence of random variables X_n such that the family$\{X_n/a_n\}$ is bounded in probability, i.e., $\lim_{C\to\infty}\sup_n \mathbb{P}(|X_n|>Ca_n)=0$.

$\omega(1)$ denotes a sequence tending to $+\infty$.

$x\wedge y$ denotes $\min\{x,y\}$.

2.2. **Binary trees.** An *(extended) binary tree* is a rooted tree where each node has either 0 or two children; in the latter case there is one left child and one right child. Nodes with 0 children (leaves) are called *external nodes* and nodes with 2 children are called *internal nodes*.

Let $V_{\mathsf{i}}(T)$ denote the set of internal nodes of T, and $V_{\mathsf{e}}(T)$ the set of external nodes.

The root of a binary tree is denoted o. The depth $d(v)$ of a node in a binary tree is the distance from v to the root o; thus $d(o)=0$.

If v and w are nodes in a binary tree T, then $v\preceq w$ means that v is on the path from the root to w (including the endpoints).

Unless we say otherwise, we consider only finite binary trees. However, we let T_∞ denote the infinite binary tree where each node has two children. Thus, T_∞ has 2^k nodes of depth k, $k\geqslant 0$. Every finite binary tree can be regarded as a subtree of T_∞.

The size $|T|:=|V_{\mathsf{i}}(T)|$ of an extended binary tree is the number of internal nodes. Thus an extended binary tree of size n has n internal and $n+1$ external nodes.

A binary tree is *empty* if it has size 0, i.e., if there is no internal node and only a single external node (the root).

The (external) height $h_{\mathsf{e}}(T)$ of a binary tree T is the maximum depth of an external node, i.e., (with $\max\emptyset:=-1$ for the empty tree)

$$h_{\mathsf{e}}(T):=\max\bigl\{d(v):v\in V_{\mathsf{e}}(T)\bigr\}=\max\bigl\{d(w):w\in V_{\mathsf{i}}(T)\bigr\}+1. \tag{2.1}$$

T_K is the complete binary tree of height K; it has 2^K external nodes, all at depth K, and thus 2^K-1 internal nodes.

Remark 2.1. It is also common to study binary trees without external nodes; we may call them *reduced binary tree*. The subtree of internal nodes in an extended binary tree is a reduced binary tree (including the case of a reduced empty tree with no nodes), and this gives an obvious 1–1 correspondence between extended and reduced binary trees. In the present papers, all binary trees are extended binary trees as defined above.

2.3. **A random walk.** Given an extended binary tree T, consider the random walk defined by starting at the root, and then moving repeatedly from the current node to one of its children, chosen at random with probability 1/2 each (independently of previous choices), until we reach an external node.

For an external node v, let p_v be the probability that this random walk ends in v. Thus $(p_v)_{v\in V_{\mathsf{e}}(T)}$ is a probability distribution on V_{e}, which we call the *harmonic measure* of T. Obviously, the harmonic measure is given by

$$p_v = 2^{-d(v)}, \qquad v \in V_{\mathsf{e}}(T). \tag{2.2}$$

By construction, the harmonic measure is a probability measure, and thus, for any finite binary tree T,

$$\sum_{v\in V_{\mathsf{e}}(T)} 2^{-d(v)} = 1. \tag{2.3}$$

(Alternatively, (2.3) is easily seen by induction on the size $|T|$.)

2.4. **Boundaries.** We say that a finite set B of nodes in $V_{\mathsf{i}}(T_\infty)$ is a *boundary*, if every infinite path from the root contains exactly one element of B. The set of external nodes $V_{\mathsf{e}}(T)$ of a finite binary tree is a boundary; conversely, given a boundary B, there exists exactly one binary tree T with $B = V_{\mathsf{e}}(T)$. (The internal nodes of T are the nodes v that are strict ancestors of some node $w \in B$.) Hence there is a 1–1 correspondence between (finite) binary trees and boundaries, given by $T \leftrightarrow V_{\mathsf{e}}(T)$.

Given a boundary B, the harmonic measure (2.2) on the corresponding tree is a probability measure on B, which we also call the *harmonic measure on* B.

2.5. **Digital search trees.** A *digital search tree* is a binary tree constructed recursively from a sequence of $n \geqslant 0$ infinite binary strings $W_1, \dots, W_n$ (called *items*) as follows; the digital search tree has size n and each internal node stores one of the items. See e.g. [11, Section 6.3], [15, Section 6.1], [4, Section 1.4.3], [8, Section 6.4].

Definition 2.2. The digital search tree is constructed as follows.

(i) Start with an empty binary tree, containing only the root as an external node.
(ii) The items W_i arrive one by one, in order; each item comes first to the root of the tree.
(iii) When an item comes to an external node, it is stored there. The node becomes internal and two new external nodes are added as children to it.
(iv) When an item W_i comes to an internal node v at depth d, it is passed to the left [right] child of v if the $(d+1)$th bit of W is 0 [1]. The construction proceeds recursively until an external node is reached.

We shall only consider the random case, where each string W_i is a random string of independent bits, each with the symmetric Be(1/2) distribution, and furthermore the strings are independent. We let $\mathcal{T}_n$ denote the random digital search tree constructed from such strings, and we consider the sequence $(\mathcal{T}_n)_0^\infty$ constructed from an infinite sequence of items $(W_n)_1^\infty$.

It is obvious from the definitions, that when constructing the random digital search tree $\mathcal{T}_n$, the ith string W_i performs a random walk on $\mathcal{T}_{i-1}$ as described in Section 2.3. (Hence, the digital search tree equals IDLA for this directed random walk, as said in the introduction.) Consequently, the

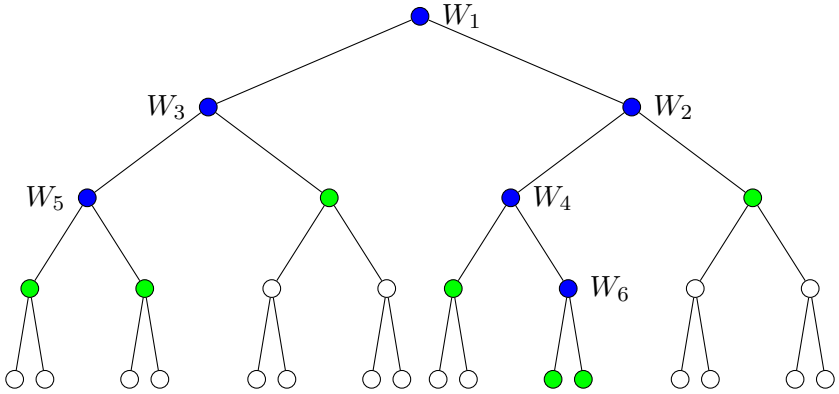

FIGURE 2. Digital search tree for 6 items; $W_1 = \{01011\ldots\}$, $W_2 = \{10011\ldots\}, W_3 = \{00101\ldots\}, W_4 = \{10110\ldots\}, W_5 = \{00011\ldots\}, W_6 = \{10100\ldots\}$. The green nodes are the external nodes and the blue nodes are the internal nodes.

sequence of random digital search trees $(\mathcal{T}_n)_0^\infty$ can also be defined as follows, without explicitly using random strings.

Definition 2.3. The random digital search trees $\mathcal{T}_n$, $n \geqslant 0$, are constructed recursively, starting with $\mathcal{T}_0$ empty. $\mathcal{T}_{n+1}$ is obtained from $\mathcal{T}_n$ by choosing an external node v in $\mathcal{T}_n$ at random according to the harmonic measure (2.2) and converting this node v to an internal node by adding two (external) children to it.

2.6. **Continuous-time digital search trees.** We think of item W_i as arriving at time i, and $(\mathcal{T}_n)_0^\infty$ as a stochastic process of trees in discrete time. It is, as often in similar problems, useful to consider also the corresponding process in continuous time, with items arriving according to a Poisson process with rate 1. This means that item W_n arrives at a random time $\tau(n)$, where the waiting times $\eta_n := \tau(n) - \tau(n-1)$ (with $\tau(0) = 0$) are i.i.d. Exp(1).

Definition 2.4. Let the sequence $(W_n)_n^\infty$ of random items arrive according to a Poisson process with rate 1 on $[0, \infty)$. (As above, the strings W_n are independent, with independent Be(1/2) bits.) The continuous-time digital search tree $\mathfrak{T}_t$ is the digital search tree constructed from the items W_i that have arrived until time t.

Equivalently, we can use Definition 2.3, adding new nodes at times given by a Poisson process.

Let $N(t)$ be the number of items that have arrived up to time t; thus $N(t) \sim \mathrm{Po}(t)$, and $\mathfrak{T}_t$ is the random digital search tree constructed from a random number $N(t)$ items. More precisely, the discrete and continuous-time processes $(\mathcal{T}_n)_n$ and $(\mathfrak{T}_t)_t$ are related by

$$\mathfrak{T}_t = \mathcal{T}_{N(t)} \quad (t \geqslant 0), \qquad\qquad \mathcal{T}_n = \mathfrak{T}_{\tau(n)} \quad (n \geqslant 0). \tag{2.4}$$

In other words, $\mathfrak{T}_t$ is obtained from $\mathcal{T}_n$ by Poissonization.

Note that $\tau(n)$ is the stopping time when the size $|\mathfrak{T}_t|$ becomes n.

2.7. **The border aggregation model.** *Border aggregation models* on finite connected graphs were studied by Thacker and Volkov [17]. In general, consider any (directed or undirected) finite, connected graph with a fixed vertex o, the *origin*, and a non-empty *boundary* set denoted by B. As in the introduction, we recursively define a randomly growing sequence of sets of *sticky* vertices S_n as follows.

Definition 2.5. Construct random *sticky sets* S_n, $n \geqslant 0$, as follows.

(i) $S_0 = B$, the given boundary.
(ii) At times $n = 1, 2, \dots$, given S_{n-1}, let a particle start at o and perform a random walk until it reaches a neighbour v_n of the sticky set. Then it stops, and the node v_n is added to the sticky set, i.e., $S_{n+1} := S_{n-1} \cup \{v_n\}$.
(iii) This is repeated until some time ξ_K when the root becomes sticky; then the process stops.

Thus, ξ_K is the number of particles required to build a path from the boundary to the origin by this aggregation process. We are (as Thacker and Volkov [17]) interested in the distribution of ξ_K.

This model was introduced as *internal erosion* by Levine and Peres [13]. In the present paper we consider only the case described in the introduction, when the graph is the binary tree T_K and the random walk is the directed random walk in Section 2.3.

We use also a continuous-time version of the border aggregation model.

Definition 2.6. The continuous-time border aggregation model is defined as in Definition 2.5, but with particles arriving according to a Poisson process with rate 1. Let Ξ_K be the time this process stops. (We assume that the random walk takes no time.)

Thus, with the notation in Section 2.6,

$$\Xi_K = \tau(\xi_K) = \sum_{i=1}^{\xi_K} \eta_i, \tag{2.5}$$

where $\eta_i \sim \mathrm{Exp}(1)$ are i.i.d. and independent of ξ_K. In particular,

$$\mathbb{E}\,\Xi_K = \mathbb{E}\,\xi_k. \tag{2.6}$$

3. More on continuous-time digital search trees

We give first an alternative construction of the continuous-time digital search tree $\mathfrak{T}_t$ and then show that it agrees with Definition 2.4.

Definition 3.1. Equip each node v in the infinite binary tree T_∞ with a random variable $X_v \sim \mathrm{Exp}(2^{-d(v)})$, with all X_v independent. Let

$$Y_v := \sum_{w \preceq v} X_w, \qquad v \in V(T_\infty), \tag{3.1}$$

and let $\mathfrak{T}_t$ be the extended binary tree with

$$V_{\mathrm{i}}(\mathfrak{T}_t) := \{v \in V(T_\infty) : Y_v \leqslant t\}. \tag{3.2}$$

Remark 3.2. We may interpret the internal nodes in $\mathfrak{T}_t$ as infected; then Definition 3.1 describes an infection that spreads randomly on T_∞ from parents to children, starting with the root o being infected from the outside, where X_v is the time it takes for node v to become infected once its parent is. (Imagine the root having an outside parent that is infected at time 0.) In other words, $\mathfrak{T}_t$ can be seen as first-passage percolation on T_∞, but note that different edges have different distributions of the infection times X_v.

To see the equivalence of Definition 2.4 and Definition 3.1, we introduce a third definition, and then show that all three are equivalent.

Definition 3.3. Equip each node $v \in T_\infty$ with an exponential clock that rings with rate $2^{-d(v)}$, independently of all other clocks. Start with $\mathfrak{T}_0$ empty. Ignore all clocks that are not currently in an external node. When a clock rings in an external node v, then v becomes an internal node of $\mathfrak{T}_t$ and its two children become new external nodes.

Remark 3.4. More generally, Aldous and Shields [1] studied a process defined as in Definition 3.3 but with rates $c^{-d(v)}$ for some constant $c > 1$. (See [2] for $c < 1$.) They noted that this is equivalent to Definition 3.1 (with these rates), and that that the process is a random time change of the corresponding discrete-time process defined as in Definition 2.3, but using instead of the harmonic measure (2.2) on the external nodes the measure where p_v is proportional to $c^{-d(v)}$. Note that the simple relation (2.3) is special for the case $c = 2$, and thus the relation between the discrete and continuous-time models is in general more complicated than in Definition 2.4.

Theorem 3.5 (Essentially Aldous and Shields [1])**.** *Definitions 2.4, 3.1 and 3.3 define the same stochastic process of trees $(\mathfrak{T}_t)_{t\geqslant 0}$. (In the sense of all having the same distribution.)*

Proof. In Definition 3.3, the total rate of the clocks in the external nodes is always 1, by (2.3). Hence, new internal nodes are created with rate 1. Furthermore, if v is an external node, then the clock at v rings with rate $2^{-d(v)}$, and thus the probability that the clock at v is the next clock in an external node that rings is also $2^{-d(v)}$. In other words, when a new internal node is added, it is chosen randomly among the existing external nodes according to the harmonic measure (2.2), just as in Definition 2.3. Hence the process $(\mathfrak{T}_t)$ constructed in Definition 3.3 has the same distribution as the one defined in Definitions 2.2–2.4.

Furthermore, in Definition 3.3, consider for each node $v \in T_\infty$ the stopping time, τ_v say, when v becomes an external node, and let X_v be the waiting time until the next time the clock at v rings. Then X_v, $v \in T_\infty$, are independent exponential random variables with the rates in Definition 3.1. Furthermore, since τ_v is the time the parent of v becomes an internal node (with $\tau_0 = 0$ for the root), it follows by induction that the time $\tau_v + X_v$ when the clock rings and v becomes an internal node equals Y_v defined in (3.1), and thus (3.2) holds and the process $(\mathfrak{T}_t)_t$ coincides with the one defined by Definition 3.1. □

In particular, this gives a description of the height $h_{\mathsf{e}}(\mathfrak{T}_t)$ of $\mathfrak{T}_t$, and thus indirectly also of $h_{\mathsf{e}}(\mathcal{T}_n)$. Use Definition 3.1 and let, for $k \geqslant 0$,

$$Y_k^* := \min_{v:\, d(v)=k} Y_v. \tag{3.3}$$

In other words, Y_k^* is the smallest sum $\sum_w X_w$ along a path from the root to a node of depth k; in the language of Remark 3.2, Y_k^* is the time the infection reaches depth k. (I.e., it reaches external height $k+1$.)

Corollary 3.6. *We have the equality in distribution, for all $t \geqslant 0$,*

$$h_{\mathsf{e}}(\mathfrak{T}_t) \stackrel{\mathrm{d}}{=} \min\bigl\{k \geqslant 0 : Y_k^* > t\bigr\} = \max\bigl\{k \geqslant 0 : Y_k^* \leqslant t\bigr\} + 1. \tag{3.4}$$

Equivalently, for any $t \geqslant 0$ and $k \geqslant 0$,

$$\mathbb{P}\bigl(h_{\mathsf{e}}(\mathfrak{T}_t) > k\bigr) = \mathbb{P}\bigl(Y_k^* \leqslant t\bigr). \tag{3.5}$$

Proof. Definition 3.1 and (3.3) yield the relation, for $k \geqslant 0$,

$$\{h_{\mathsf{e}}(\mathfrak{T}_t) \leqslant k\} = \{v \notin V_{\mathsf{i}}(\mathfrak{T}_t) \text{ when } d(v) = k\} = \{Y_k^* > t\}. \tag{3.6}$$

Hence, using Definition 3.1, (3.4) holds with actual equality of the random variables. By Theorem 3.5, we have equality in distribution for any of the definitions. □

4. Connection with the border aggregation model

Theorem 4.1. *For any $K \geqslant 0$, $\Xi_{K+1} \stackrel{\mathrm{d}}{=} Y_K^*$.*

We give two proofs of this theorem. The first uses a simple induction. The second is longer but perhaps gives more insight; it is more combinatorial and is based on a study of the aggregation process. The second proof also provides a coupling of the two processes.

First proof of Theorem 4.1. The claim is trivially true for $K = 0$: Ξ_1 is the time of arrival of the first particle, so $\Xi_1 \sim \mathrm{Exp}(1)$ and $\Xi_1 \stackrel{\mathrm{d}}{=} X_o = Y_0^*$.

Denote the two children of the root by o_{L} and o_{R}. Consider the continuous-time border aggregation model on T_{K+1}, and let Ξ_{K+1}^* be the time o_{L} or o_{R} becomes sticky. Then the next particle stops at the root, and thus

$$\Xi_{K+1} = \Xi_{K+1}^* + X, \tag{4.1}$$

where $X \sim \mathrm{Exp}(1)$ is independent of Ξ_{K+1}^*.

Up to time Ξ_{K+1}^*, the particles proceed to o_{L} or o_{R}, with probability $1/2$ each and independently of each other and of the arrival times of the particles. By a standard property of Poisson processes, this means that o_{L} and o_{R} are fed particles by two independent Poisson processes with rates $1/2$. Let both these processes continue beyond Ξ_{K+1}^*, and let Ξ_{L} and Ξ_{R} be the times o_{L} and o_{R}, respectively, then become sticky. Then

$$\Xi_{K+1}^* = \Xi_{\mathsf{L}} \wedge \Xi_{\mathsf{R}}. \tag{4.2}$$

Moreover, the two processes beneath o_{L} and o_{R} are independent copies of the original process on the smaller tree T_K, with time running at half speed. Hence, $\Xi_{\mathsf{L}} \stackrel{\mathrm{d}}{=} \Xi_{\mathsf{R}} \stackrel{\mathrm{d}}{=} 2\Xi_K$, and thus by (4.1) and (4.2),

$$\Xi_{K+1} \stackrel{\mathrm{d}}{=} 2\bigl(\Xi_K \wedge \Xi_K'\bigr) + X \tag{4.3}$$

with $\Xi'_K \stackrel{\mathrm{d}}{=} \Xi_K$, $X \sim \mathrm{Exp}(1)$ and Ξ_K, Ξ'_K, X independent.

Similarly, recalling the definition (3.3) of Y^*_K, let Y^*_{L} and Y^*_{R} be the smallest sum $\sum X_v$ along a path from o_{L} or o_{R}, respectively, to a node of depth K. Then

$$Y^*_K = (Y^*_{\mathsf{L}} + X_o) \wedge (Y^*_{\mathsf{R}} + X_o) = Y^*_{\mathsf{L}} \wedge Y^*_{\mathsf{R}} + X_o. \tag{4.4}$$

Moreover, Y^*_{L} and Y^*_{R} are independent and both have the same distribution as $2Y^*_{K-1}$, since the subtree of descendants of o_{L} (or o_{R}), equipped with their X_v is isomorphic to the full tree with root o, but given the variables $2X_v$. Hence, (4.4) yields

$$Y^*_K \stackrel{\mathrm{d}}{=} 2(Y^*_{K-1} \wedge Y^{*\prime}_{K-1}) + X_o, \tag{4.5}$$

with $Y^{*\prime}_{K-1} \stackrel{\mathrm{d}}{=} Y^*_{K-1}$, $X_o \sim \mathrm{Exp}(1)$, and Y^*_{K-1}, $Y^{*\prime}_{K-1}$ and X_o independent.

Comparing (4.3) and (4.5), we see that the distributions of Ξ_{K+1} and Y^*_K satisfy the same recursive equation, and thus they are equal by induction. □

Second proof of Theorem 4.1. In the (discrete or continuous-time) border aggregation model, define, at any given time t, the *absorption set* A_t as the set of all internal nodes v such that v is a neighbour of the sticky set S_t, but no ancestor of v is. Consider only the process (A_t) of absorption sets; A_t evolves by letting a new particle perform the random walk until it hits A_t, say at v. Then v becomes sticky, which means that the parent v' of v is added to A_t, while v and all other descendants of v' are removed. (If v is the root, then instead the process stops.)

Note that the absorption set A_t is a boundary in the sense of Section 2.4, and that given the boundary A_t at some time t, the next node that becomes sticky is chosen randomly from A_t according to the harmonic measure on A_t, see Sections 2.3 and 2.4. Furthermore, $(A_t)_t$ is a Markov process.

From now on we consider the continuous-time version; furthermore, we consider the tree T_{K+1} with external nodes at depth $K+1$. Equip the nodes $v \in V_K := \{v : d(v) \leqslant K\} = V_{\mathsf{i}}(T_{K+1})$ with exponential clocks as in Definition 3.3. Define a process A'_t of subsets of V_K as follows:

(i) $A'_0 := A_0 = \{v : d(v) = K\}$.

(ii) Clocks outside the current A'_t are ignored. When a clock at a node $v \in A'_t$ rings, A'_t is updated as above; i.e., the parent v' of v is added to A'_t, while v and all other descendants of v' are removed. (If v is the root, then instead the process stops.)

Given A'_t, the next clock in A'_t that rings is random with a distribution given by the harmonic measure on A'_t. Hence, the process A'_t just constructed has the same distribution as the process A_t in the aggregation process, and we may assume that $A_t = A'_t$ for all $t \geqslant 0$.

For each node $v \in V_K$, let now $\tau_v := \inf\{t \geqslant 0 : u \in A_t \text{ for some } u \preceq v\}$, i.e., the first time that either v or one of its ancestors belongs to the absorption set, and let X_v be the waiting time from τ_v to the next time that the clock at v rings. Then the random variables X_v, $v \in V_K$, are independent and have the exponential distributions given in Definition 3.1. (We may define X_v also for $d(v) > K$ for completeness, but these variables will not matter.) Define Y_v by (3.1).

For a node $v \in V_K$, let

$$Z_v := \min_{w \succeq v,\, d(w)=K} \{Y_w - Y_v\}. \tag{4.6}$$

This is the minimum over the paths from v to the boundary $V_{\mathsf{e}}(T_{K+1})$ of the sum $\sum_u X_u$ for all nodes u in the path, excluding the endpoints. In particular, $Z_v = 0$ when $d(v) = K$.

We claim that at any time $t \geqslant 0$ with $t \leqslant \Xi_{K+1}$,

$$A_t = \{v \in V_K : Z_v \leqslant t \text{ but } Z_u > t \text{ for all } u \prec v\}, \tag{4.7}$$

and furthermore

$$\tau_v = Z_v \qquad \text{for every } v \in A_t. \tag{4.8}$$

We prove this claim by induction; it is evidently true for $t = 0$, and it then suffices to consider the finite number of times that A_t changes.

Suppose that the claim holds for some time t. If $v \in A_t$, then the next time that the clock at v rings is, letting again v' be the parent of v and noting that $Y_v = Y_{v'} + X_v$ (with $Y_{o'} := 0$),

$$\tau_v + X_v = Z_v + X_v = \min_{w \succeq v,\, d(w)=K} \{Y_w - Y_{v'}\}. \tag{4.9}$$

Let v be the node in the current A_t such that the time $Z_v + X_v$ in (4.9) is minimal. Then v is the next node to become sticky, and its parent v' is the next node added to A_t; this happens at time $\tau_{v'} = Z_v + X_v$, which by (4.9) equals the minimum over all paths from v' to $V_{\mathsf{e}}(T_{K+1})$ that pass through v of the sum $\sum_u X_u$ for u in the path, excluding the endpoints. A path from v' to $V_{\mathsf{e}}(T_{K+1})$ that does not pass through v must pass through some other node $v'' \in A_t$, and since $Z_{v''} + X_{v''} \geqslant Z_v + X_v$, it follows that $\sum_u X_u$ for u in this path is $\geqslant Z_v + X_v$. Hence, using (4.9) and (4.6), $\tau_{v'} = Z_v + X_v = Z_{v'}$; moreover (4.7) holds up to time $Z_v + X_v$. This completes the induction step, and thus the proof of the claim (4.7)–(4.8).

Obviously, $o \in A_t$ for some t, and thus (4.8) applies to $v = o$. Consequently, the time Ξ_{K+1} that the root becomes sticky is, using the definitions of τ_o and X_o together with (4.8), (3.1), (4.6) and (3.3),

$$\Xi_{K+1} = \tau_o + X_o = Z_o + X_o = Z_o + Y_o = Y^*_K. \tag{4.10}$$

□

Proof of Theorem 1.1. (ii): Theorem 4.1 and Corollary 3.6 yield, for $K \geqslant 1$ and $t \geqslant 0$,

$$\mathbb{P}\big(\Xi_K \leqslant t\big) = \mathbb{P}\big(Y^*_{K-1} \leqslant t\big) = \mathbb{P}\big(h_{\mathsf{e}}(\mathfrak{T}_t) \geqslant K\big). \tag{4.11}$$

(i): By (3.6),

$$Y^*_k = \min\{t \geqslant 0 : h_{\mathsf{e}}(\mathfrak{T}_t) > k\}. \tag{4.12}$$

Define analoguously, for the discrete time process,

$$Y^{**}_k := \min\{n \geqslant 0 : h_{\mathsf{e}}(\mathcal{T}_n) > k\}. \tag{4.13}$$

Then, see the relations (2.4),

$$Y^*_k = \tau(Y^{**}_k) = \sum_{i=1}^{Y^{**}_k} \eta_i, \tag{4.14}$$

where as in (2.5), η_i are i.i.d. Exp(1) and independent of the discrete time process. Hence, (2.5), Theorem 4.1 and (4.14) yield

$$\sum_{i=1}^{\xi_{K+1}} \eta_i = \Xi_{K+1} \overset{\mathrm{d}}{=} Y_K^* = \sum_{i=1}^{Y_K^{**}} \eta_i. \tag{4.15}$$

If we take the Laplace transforms of the left-hand side, we obtain by conditioning on ξ_{K+1}, for any $s \geqslant 0$,

$$\mathbb{E}\exp\Bigl(-s\sum_{i=1}^{\xi_{K+1}}\eta_i\Bigr) = \mathbb{E}\Bigl(\bigl(\mathbb{E}\,e^{-s\eta}\bigr)^{\xi_{K+1}}\Bigr) = \mathbb{E}\Bigl((1+s)^{-\xi_{K+1}}\Bigr). \tag{4.16}$$

This and an identical calculation for the right-hand side show that, taking $s = x^{-1} - 1$, $\mathbb{E}\bigl(x^{\xi_{K+1}}\bigr) = \mathbb{E}\bigl(x^{Y_K^{**}}\bigr)$ for every $x \in (0,1)$. In other words, ξ_{K+1} and Y_K^{**} have the same probability generating function, and thus the same distribution.

Consequently, using the definition (4.13), for $K \geqslant 0$,

$$\mathbb{P}\bigl(\xi_{K+1} \leqslant n\bigr) = \mathbb{P}\bigl(Y_K^{**} \leqslant n\bigr) = \mathbb{P}\bigl(h_{\mathsf{e}}(\mathcal{T}_n) > K\bigr) = \mathbb{P}\bigl(h_{\mathsf{e}}(\mathcal{T}_n) \geqslant K-1\bigr). \tag{4.17}$$

The result follows by replacing K by $K-1$. □

5. Proofs of Theorems 1.2 and 1.4

We next prove Theorem 1.2, using Drmota, Fuchs, Hwang and Neininger [5, Theorem 4 and its proof in Section 6.1].

Proof of Theorem 1.2. Let $n = n_K$, $K \geqslant 1$, be such that

$$\log_2 n = K - \sqrt{2K} + \frac{1}{2}\log_2 K - \frac{1}{\log 2} + \frac{\log_2 K}{4\sqrt{2K}} + \frac{a_K}{\sqrt{K}} \tag{5.1}$$

for some sequence a_K. Later in the proof we will choose a_k such that $a_K \to \pm\infty$, arbitrarily slowly, and we may assume $a_K = o\bigl(\sqrt{K}\bigr)$ as $K \to \infty$. Define

$$\tilde{k} := \log_2 n + \sqrt{2\log_2 n} - \frac{1}{2}\log_2\log_2 n + \frac{1}{\log 2}, \tag{5.2}$$

$$\tilde{\theta} := \frac{3\log_2\log_2 n}{4\sqrt{2\log_2 n}}, \tag{5.3}$$

and, as in [5],

$$k_H := \lfloor \tilde{k} \rfloor, \tag{5.4}$$

$$k_\ell := k_H + \ell, \text{ for } \ell \in \mathbb{Z}, \tag{5.5}$$

$$\theta := \tilde{k} - k_H \in [0,1), \tag{5.6}$$

Elementary calculations show that

$$\sqrt{\log_2 n} = \sqrt{K} - \frac{1}{\sqrt{2}} + \frac{\log_2 K}{4\sqrt{K}} + \mathcal{O}\Bigl(\frac{1}{\sqrt{K}}\Bigr), \tag{5.7}$$

$$\log_2\log_2 n = \log_2 K + \mathcal{O}\Bigl(\frac{1}{\sqrt{K}}\Bigr), \tag{5.8}$$

$$\tilde{k} - \tilde{\theta} = K - 1 + \frac{a_K}{\sqrt{K}} + \mathcal{O}\Bigl(\frac{1}{\sqrt{K}}\Bigr), \tag{5.9}$$

$$k_1 = \tilde{k} - \theta + 1 = K + \tilde{\theta} - \theta + \frac{a_K + \mathcal{O}(1)}{\sqrt{K}}, \tag{5.10}$$

$$\theta - \tilde{\theta} = K - k_1 + \frac{a_K + \mathcal{O}(1)}{\sqrt{K}}. \tag{5.11}$$

In particular, since $a_K = o\big(\sqrt{K}\big)$ and $\tilde{\theta} = o(1)$, (5.9) implies

$$\tilde{k} = K - 1 + o(1), \tag{5.12}$$

and thus, for all large K,

$$k_H = \lfloor \tilde{k} \rfloor \in \{K-1, K-2\}. \tag{5.13}$$

In other words, for large K, either $K = k_H + 1 = k_1$ or $K = k_H + 2 = k_2$.

Suppose now that $a_K \to -\infty$. On the subsequence where $K = k_2$ (if there are any such K), we have by (1.1) and [5, Lemma 14], writing $H_n := h_{\mathsf{e}}(\mathcal{T}_n)$ as in [5],

$$\mathbb{P}(\xi_k \leqslant n) = \mathbb{P}\big(H_n \geqslant K\big) = \mathbb{P}\big(H_n > k_H + 1\big) \to 0. \tag{5.14}$$

On the subsequence where $K = k_1$ (if there are any such K), (5.11) yields

$$\theta - \tilde{\theta} = -\frac{\omega(1)}{\sqrt{K}} = -\frac{\omega(1)}{\sqrt{\log_2 n}}, \tag{5.15}$$

and thus, using also [5, Remark 5], $\mathbb{P}(H_n = k_1) \to 0$, and thus

$$\mathbb{P}(\xi_k \leqslant n) = \mathbb{P}\big(H_n \geqslant K\big) = \mathbb{P}\big(H_n = k_H + 1\big) + \mathbb{P}\big(H_n > k_H + 1\big) \to 0. \tag{5.16}$$

Together, (5.14) and (5.16) show that if $a_K \to -\infty$, then $\mathbb{P}(\xi_k \leqslant n) \to 0$ as $K \to \infty$, regardless of whether $K = k_1$ or k_2.

On the other hand, suppose that $a_K \to +\infty$. Since $\tilde{\theta} > 0$ (for large K at least), (5.9) implies that for large K, $\tilde{k} \geqslant K - 1$, and thus, by (5.13), $k_H = K - 1$ and $K = k_1$. Furthermore, (5.11) implies

$$\theta - \tilde{\theta} = \frac{\omega(1)}{\sqrt{K}} = \frac{\omega(1)}{\sqrt{\log_2 n}}. \tag{5.17}$$

Hence, [5, Remark 5 and Lemma 13] imply that $\mathbb{P}(H_n \leqslant k_H) \to 0$, and thus (1.1) yields

$$\mathbb{P}(\xi_k \leqslant n) = \mathbb{P}\big(H_n \geqslant K\big) = \mathbb{P}\big(H_n > k_H\big) \to 1. \tag{5.18}$$

Finally, define

$$Z_K := \sqrt{K}\Big(\log_2 \xi_K - \Big(K - \sqrt{2K} + \frac{1}{2}\log_2 K - \frac{1}{\log 2} + \frac{\log_2 K}{4\sqrt{2K}}\Big)\Big). \tag{5.19}$$

Then, (5.14), (5.16) and (5.18) show, together with (5.1), that if $a_K \to -\infty$, then $\mathbb{P}(Z_K \leqslant a_K) \to 0$, while if $a_K \to +\infty$, then $\mathbb{P}(Z_K \leqslant a_K) \to 1$. This is equivalent to $Z_K = \mathcal{O}_{\mathrm{p}}(1)$, and thus to (1.3). □

Finally, we use Theorem 1.2 to prove Theorem 1.4 on the mean.

Proof of Theorem 1.4. In this proof, all limits are as $K \to \infty$. First, (1.5) implies,

$$\xi_K / m_K \xrightarrow{\mathrm{p}} 1, \tag{5.20}$$

and thus, by (2.5) and the law of large numbers,

$$\Xi_K/m_K \xrightarrow{\mathrm{p}} 1. \tag{5.21}$$

Note that this immediately implies, by Fatou's lemma [7, Theorem 5.5.3],

$$\liminf_{K\to\infty} \frac{\mathbb{E}\,\Xi_K}{m_K} \geqslant 1. \tag{5.22}$$

To obtain also an upper bound, we use tail estimates by Drmota [3]. Note that Drmota uses the internal height, thus his $H_n = h_{\mathsf{e}}(\mathcal{T}_n)-1$. Furthermore, $P_k(x)$ in [3] is the distribution function of the Poissonized version of H_n, and thus in our notation

$$P_k(x) = \mathbb{P}\big(h_{\mathsf{e}}(\mathfrak{T}_x) - 1 \leqslant k\big). \tag{5.23}$$

Hence, by (1.2), for $K \geqslant 2$ and $x \geqslant 0$,

$$\mathbb{P}(\Xi_K > x) = \mathbb{P}\big(h_{\mathsf{e}}(\mathfrak{T}_x) \leqslant K-1\big) = P_{K-2}(x). \tag{5.24}$$

We use [3, Lemma 4], for convenience denoting n_{K-2} there by $\bar n_K$ and noting that $\frac12 < c_k < 1$ for large k; this yields together with (5.24), for large K,

$$\mathbb{P}(\Xi_K > x) \geqslant \Big(1 - \frac{1}{\bar n_K}\Big)e^{-x/\bar n_K}, \qquad 0 \leqslant x \leqslant \bar n_K, \tag{5.25}$$

$$\mathbb{P}(\Xi_K > x) \leqslant e^{-x/(2\bar n_K)}, \qquad x \geqslant \bar n_K. \tag{5.26}$$

Let $\varepsilon > 0$. Then (5.21) says that $\mathbb{P}\big((1-\varepsilon)m_K < \Xi_K < (1+\varepsilon)m_K\big) \to 1$, which combined with (5.25)–(5.26) (taking $x = \bar n_K$) implies that for large K we must have $(1-\varepsilon)m_K < \bar n_K < (1+\varepsilon)m_K$. In other words,

$$\bar n_K/m_K \to 1. \tag{5.27}$$

Hence, (5.21) is equivalent to $\Xi_K/\bar n_K \xrightarrow{\mathrm{p}} 1$, which means

$$\mathbb{P}\big(\Xi_K/\bar n_K > x\big) \to \mathbf{1}\{x<1\} \tag{5.28}$$

for every $x \neq 1$. Furthermore, (5.26) implies that, for large K,

$$\mathbb{P}\big(\Xi_K/\bar n_K > x\big) \leqslant \mathbf{1}\{x<1\} + e^{-x/2} \tag{5.29}$$

for every $x \geqslant 0$. Consequently, dominated convergence yields

$$\mathbb{E}\,\frac{\Xi_K}{\bar n_K} = \int_0^\infty \mathbb{P}\Big(\frac{\Xi_K}{\bar n_K} > x\Big)\,\mathrm{d}x \to \int_0^\infty \mathbf{1}\{x<1\}\,\mathrm{d}x = 1 \tag{5.30}$$

as $K \to \infty$. The result follows by (5.27) and (2.6). □

Remark 5.1. To prove Conjecture 1.3 by similar arguments, one would need much stronger tail estimates than (5.25)–(5.26). It seems that the method of proof of [5, Lemma 14] might give the required estimates; however, we have not verified the (non-trivial) details and leave the conjecture as an open problem.

6. b-ary trees

We have in this paper only considered binary trees. A random b-ary digital search tree can be constructed in the same way for any given $b \geqslant 2$, using strings W_i with letters from an alphabet $\mathcal{A}$ of size b, for example $\mathcal{A} = \{0, 1, \cdots, b-1\}$; we still assume that the letters are independent and that all letters have the same probability (viz. $1/b$).

Similarly, the border aggregation model can be defined on b-ary trees as in Definition 2.5, where now the random walk at each step selects a child with probability $1/b$ each.

Most of the results above hold with only trivial changes. The harmonic measure (2.2) becomes $b^{-d(v)}$. In Definitions 3.1 and 3.3, the rate should be $b^{-d(v)}$. In particular, Theorems 1.1 and 4.1 still hold (by the same arguments).

However, Theorem 1.2 uses results for the binary case proved in [5]; the results and methods there ought to generalize to arbitrary b, but that has not yet been done, so we cannot extend this result to larger b. Nevertheless, we conjecture that for the border aggregation model on regular b-ary trees, for a suitable constant $c_b > 0$,

$$\log_b \xi_K = K - \sqrt{2K} + c_b \log_b K + \mathcal{O}_{\mathrm{p}}(1). \tag{6.1}$$

References

[1] David Aldous and Paul Shields. A diffusion limit for a class of randomly-growing binary trees. *Probab. Theory Related Fields* **79** (1988), no. 4, 509–542.

[2] Martin T Barlow, Robin Pemantle and Edwin A. Perkins. Diffusion-limited aggregation on a tree. *Probab. Theory Related Fields* **107** (1997), no. 1, 1–60.

[3] Michael Drmota. The variance of the height of digital search trees. *Acta Inform.* **38** (2002), no. 4, 261–276.

[4] Michael Drmota, *Random Trees*, Springer, Vienna, 2009.

[5] Michael Drmota, Michael Fuchs, Hsien-Kuei Hwang and Ralph Neininger, Node profiles of symmetric digital search trees: Concentration properties. *Random Struct Alg.*, Early View (2020).

[6] Michael Drmota, Svante Janson and Ralph Neininger. A functional limit theorem for the profile of search trees. *Ann. Appl. Probab.* **18** (2008), no. 1, 288–333.

[7] Allan Gut. *Probability: A Graduate Course*. Springer, New York, 2005.

[8] Philippe Jacquet and Wojciech Szpankowski: *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, Cambridge, 2015.

[9] Harry Kesten. How long are the arms in DLA? *J. Phys. A* **20** (1987), no. 1, L29–L33.

[10] Charles Knessl and Wojciech Szpankowski. Asymptotic behavior of the height in a digital search tree and the longest phrase of the Lempel-Ziv scheme. *SIAM J. Comput.* **30** (2000), no. 3, 923–964.

[11] Donald E. Knuth: *The Art of Computer Programming. Vol. 3: Sorting and Searching*. 2nd ed., Addison-Wesley, Reading, MA, 1998.

[12] Gregory F. Lawler, Maury Bramson and David Griffeath: Internal diffusion limited aggregation. *Ann. Probab.* **20** (1992), no. 4, 2117–2140.
[13] Lionel Levine and Yuval Peres. Internal erosion and the exponent 3/4. Preprint, 2007.
`http://www.math.cornell.edu/~levine/erosion.pdf`
[14] Lionel Levine and Vittoria Silvestri. How long does it take for internal DLA to forget its initial profile? *Probab. Theory Related Fields* **174** (2019), no. 3-4, 1219–1271.
[15] Hosam M. Mahmoud: *Evolution of Random Search Trees*, Wiley, New York, 1992.
[16] Vittoria Silvestri. Internal DLA on cylinder graphs: fluctuations and mixing. Preprint, 2019. `arXiv`:1909.09893.
[17] Debleena Thacker and Stanislav Volkov. Border aggregation model. *Ann. Appl. Probab.* **28** (2018), no. 3, 1604–1633.
[18] T. A. Witten and L. M. Sander. Diffusion-limited aggregation. *Phys. Rev. B (3)* **27** (1983), no. 9, 5686–5697.

(Svante Janson) Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden
Email address: `svante.janson@math.uu.se`
URL: `http://www.math.uu.se/svante-janson`

(Debleena Thacker) Department of Mathematics, NYU, Shanghai, 1555 Century Avenue, Pudong New District, Shanghai, China 200122
Email address: `thackerdebleena@gmail.com`