Electron. J. Probab. **24** (2019), no. 18, 1–15.
ISSN: 1083-6489 https://doi.org/10.1214/19-EJP268

Finding the seed of uniform attachment trees *†

Gábor Lugosi[‡§¶] Alan S. Pereira[‖]

Abstract

A uniform attachment tree is a random tree that is generated dynamically. Starting from a fixed "seed" tree, vertices are added sequentially by attaching each vertex to an existing vertex chosen uniformly at random. Upon observing a large (unlabeled) tree, one wishes to find the initial seed. We investigate to what extent seed trees can be recovered, at least partially. We consider three types of seeds: a path, a star, and a random uniform attachment tree. We propose and analyze seed-finding algorithms for all three types of seed trees.

Keywords: random trees; uniform attachment; discrete probability; seed.
AMS MSC 2010: NA.
Submitted to EJP on January 5, 2018, final version accepted on January 22, 2019.

Dynamically growing networks represent complex relationships in numerous areas of science. In a rapidly increasing number of applications, one does not observe the entire dynamical growth procedure but merely a present-day snapshot of the network is available for observation. Based on this snapshot, one wishes to infer various properties of the *past* of the network. Such problems belong to an area now loosely being called *network archeology*, see Navlakha and Kingsford [17].

The simplest dynamically grown networks are trees that are grown by attaching vertices sequentially to the existing tree at random, according to a certain rule. In the *uniform attachment* model, at each step, an existing vertex is selected uniformly at random, and a new vertex is attached to it by an edge. When the process is initialized from a single vertex, this procedure gives rise to the well-studied *uniform random recursive tree*, see Drmota [10]. In *preferential attachment* models (such as plane-oriented recursive trees) existing vertices with higher degrees are more likely to be chosen to

*Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU.
†Alan Pereira was supported by Brazilian National Council for Scientific and Technological Development.
‡Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain.
E-mail: gabor.lugosi@upf.edu
§ICREA, Pg. LluÃs Companys 23, 08010 Barcelona, Spain
¶Barcelona Graduate School of Economics
‖Instituto Nacional de Matemática Pura e Aplicada (IMPA), Estrada Dona Castorina, 110 - Jardim Botânico, Rio de Janeiro - RJ, Brazil. E-mail: alanand@impa.br

be attached to. In this paper we consider randomly growing uniform attachment trees that are grown from a fixed *seed*. Thus, initially, the tree is a given fixed (small) tree and further vertices are attached according to the uniform attachment process.

"Archeology" of randomly growing trees has received increasing attention recently, see Brautbar and Kearns [3], Borgs, Brautbar, Chayes, Khanna, and Lucier [1], Bubeck, Devroye, and Lugosi [4], Bubeck, Mossel, and Rácz [6], Bubeck, Eldan Mossel, and Rácz [5], Curien, Duquesne, Kortchemski, and Manolescu [7], Frieze and Pegden [11], Jog and Loh [15, 14], Shah and Zaman [20, 19] for a sample of the growing literature.

Several papers consider the problem of finding the initial vertex (or root) in a randomly growing tree started from a single vertex, see Brautbar and Kearns [3], Borgs, Brautbar, Chayes, Khanna, and Lucier [1], Frieze and Pegden [11], Shah and Zaman [20, 19], Bubeck, Devroye, and Lugosi [4], Jog and Loh [15, 14] for various models. Randomly growing trees started from an initial seed tree were considered by Bubeck, Mossel, and Rácz [6], Bubeck, Eldan Mossel, and Rácz [5], and Curien, Duquesne, Kortchemski, and Manolescu [7]. These papers prove that in uniform and preferential attachment models, for any pair of possible seed trees, one may construct a hypothesis test that decides which of the two seeds generated the observed tree, with a probability of error strictly smaller than $1/2$, regardless of the size of the observed tree.

In this paper we consider the problem of *finding* the seed tree (of known structure) in a large observed tree. The questions we seek to answer are: (1) to what extent is it possible to identify the seed tree? (2) what is the role of the structure of the seed in the difficulty of the reconstruction problem? While we are far from completely answering these questions, this paper contributes to the understanding of these problems. In particular, we consider three types of possible seed trees, namely paths, stars, and random uniform recursive trees. For each of these examples, we present algorithms to recover, at least partially, the seed tree. In all cases, partial recovery is possible, with any prescribed probability of error, regardless of the size of the observed tree. However, the difficulty of the recovery depends heavily on the structure of the tree. Paths and stars are considerably easier to find than uniform random recursive trees. Finally, we mention the recent paper of Devroye and Reddad [9] that considers the same kind of seed-finding algorithms in uniform attachment trees and obtain several interesting results that complement the findings of this paper.

In Section 1 we introduce the mathematical model and state the main results. The proofs of all results are presented in Section 2.

1 Setup and results

Let $\ell \geq 1$ be a positive integer and let S_ℓ be a tree (i.e., a connected acyclic graph) on the vertex set $\{1, \ldots, \ell\}$. Let $n > \ell$ be another positive integer. We say that a random tree T_n on the vertex set $\{1, \ldots, n\}$ is a *uniform attachment tree with seed* S_ℓ if it is generated as follows:

1. $T_\ell = S_\ell$;
2. For $\ell < i \leq n$, T_i is obtained from T_{i-1} by joining vertex i to a vertex of T_{i-1} chosen uniformly at random, independently of all previous choices.

The problem we study in this paper is the following. Suppose one observes a tree T_n generated by the uniform attachment process with seed S_ℓ but with the vertex labels hidden. The goal is to find the seed tree S_ℓ in the observed unlabeled tree. More precisely, given a target accuracy $\epsilon \in (0, 1)$ a seed-finding algorithm of *first kind* outputs a set $H_1(T_n, \epsilon)$ of vertices of size $k_\ell \leq \ell$, such that, with probability at least $1-\epsilon$, $H_1(T_n, \epsilon) \subset S_\ell$, that is, all elements of $H_1(T_n, \epsilon)$ are vertices of the seed tree S_ℓ. (Here, with a slight abuse of notation, we identify the seed S_ℓ with its vertex set $\{1, \ldots, \ell\}$.)

Similarly, a seed-finding algorithm of *second kind* outputs a set $H_2(T_n, \epsilon)$ of vertices of size $k_\ell \geq \ell$, such that, with probability at least $1-\epsilon$, $S_\ell \subset H_2(T_n, \epsilon)$, that is, $H_2(T_n, \epsilon)$ contains all vertices of the seed tree S_ℓ.

In both cases, one would like to have k_ℓ as close to ℓ as possible, even for small values of ϵ.

Bubeck, Devroye, and Lugosi [4] considered the case $\ell = 1$, that is, when the seed tree is a single vertex and seed-finding algorithms of the second kind. Thus, the aim of the seed-finding algorithm is to find the root of the observed tree. Their main finding is that, for all ϵ, the optimal value of k_1 stays bounded as the size n of the observed tree goes to infinity. They also show that there exist seed-finding algorithms of the second kind such that $k_1 = o(\epsilon^{-a})$ for all $a > 0$.

In this paper we show that, if ℓ is sufficiently large (depending on ϵ), then k_ℓ may be made *proportional* to ℓ for seed-finding algorithms of second kind, and we make similar statements for k_ℓ for certain seed-finding algorithms of first kind. How the required value of ℓ depends on ϵ and what the achievable proportions are depend heavily on the structure of the seed. We consider three prototypical examples of seeds:

• A *path* P_ℓ on ℓ vertices is a tree that has exactly two vertices of degree one and $\ell - 2$ vertices of degree two.

• A *star* E_ℓ on ℓ vertices is a tree that has $\ell - 1$ vertices of degree one and one vertex of degree $\ell - 1$.

• The third example we consider is when the seed S_ℓ is a uniform random recursive tree on ℓ vertices. In this case the proposed seed finding algorithm does not need to know the structure of the tree. Thus, this example may be considered as a generalization of the root-finding problem studied in [4]. Here, instead of trying to locate the root of the tree, the goal is to find the first ℓ generations of the observed uniform random recursive tree T_n.

In what follows we present the main findings of the paper that establish the existence of seed-finding algorithms that are able to recover a constant fraction of the seed if it is a uniform random recursive tree. If the seed is either a path or a star, then the situation is even better as one can recover almost the entire seed.

Importantly, all bounds established below are independent of the size n of the observed tree, meaning that (partial) reconstruction of the seed is possible regardless of how large the observed tree T_n is.

1.1 Finding the seed when it is a path

We begin with the case when the seed is a path:

Theorem 1.1. *Let $\epsilon \in (0,1)$ and $\gamma \in (0,1)$ and let $\ell \geq \max\left\{\frac{2e^2}{\gamma}\log\frac{1}{\epsilon}, \frac{2e^2}{\gamma}\log(4e^2)\right\}$ be a positive integer. Then for all $n \geq \ell$ sufficiently large, if T_n is a uniform attachment tree with seed $S_\ell = P_\ell$ (a path of ℓ vertices), then there exists a seed-finding algorithm that outputs a vertex set $H_n \subset \{1, \ldots, n\}$ with $|H_n| \geq (1-\gamma)\ell$ such that*

$$\mathbb{P}\{H_n \subset P_\ell\} \geq 1 - \epsilon\,.$$

The theorem states that, for any fixed $\gamma > 0$, if the size of the seed path ℓ is at least of the order of $\log(1/\epsilon)$, then there exists an algorithm that finds all but a γ-fraction of the seed path, regardless of how large the observed tree T_n is. Note that the required length of the path is merely logarithmic in $1/\epsilon$. In fact, this dependence is essentially best possible. The following result shows that if the seed path has less than $\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$ vertices, then *any* seed finding algorithm must miss at least half of the seed, with probability greater than ϵ.

Theorem 1.2. *Let $\epsilon \in (0, e^{-e^2})$. Suppose that T_n is a uniform attachment tree with seed $S_\ell = P_\ell$ for $\ell \le \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$. Then, for all $n \ge 2\ell$, any seed-finding algorithm that outputs a vertex set H_n of size ℓ has*

$$\mathbb{P}\left\{|H_n \cap P_\ell| \le \frac{\ell}{2}\right\} \ge \epsilon \ .$$

1.2 Finding the seed when it is a star

Next we state our results for the case when the seed tree is a star E_ℓ on ℓ vertices.

Theorem 1.3. *There exists a numerical positive constant C such that the following holds. Let $\epsilon \in (0,1)$ and $\gamma \in (0,1)$ and let $\ell \ge \max(C, 8/\gamma)\log(1/\epsilon)$ be a positive integer. Then for all $n \ge \ell$ sufficiently large, if T_n is a uniform attachment tree with seed $S_\ell = E_\ell$ (a star of ℓ vertices), then there exists a seed-finding algorithm that outputs a vertex set $H_n \subset \{1, \dots, n\}$ with $|H_n| \le (1+\gamma)\ell$ such that*

$$\mathbb{P}\left\{E_\ell \subset H_n\right\} \ge 1 - \epsilon \ .$$

Once again, the order of magnitude for the required size of the seed star is essentially optimal as a function of ϵ. The proof of the next theorem is similar to that of Theorem 1.2 and thus it is omitted.

Theorem 1.4. *Let $\epsilon \in (0, e^{-e^2})$. Suppose that T_n is a uniform attachment tree with seed $S_\ell = E_\ell$ for $\ell \le \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$. Then, for all $n \ge 2\ell$, any seed-finding algorithm that outputs a vertex set H_n of size ℓ has*

$$\mathbb{P}\left\{|H_n \cap E_\ell| \le \frac{\ell}{2}\right\} \ge \epsilon \ .$$

1.3 Finding the first generations

Finally, we consider the case when the seed tree is a uniform random recursive tree in ℓ vertices. Unlike in the previous two examples, here the seed finding algorithm does now "know" the exact structure of the seed. This model may be equivalently formulated as follows: starting from a single vertex, one grows a uniform random recursive tree T_n of n vertices. Upon observing T_n (without vertex labels), one's aim is to recover as much of the tree T_ℓ (containing vertices attached in the first ℓ generations) as possible. The next theorem establishes the existence of a seed-finding algorithm of the first kind that identifies an $\Omega(1/\log(1/\epsilon))$ fraction of the vertices of the seed T_ℓ with probability at least $1-\epsilon$, whenever ℓ is at least proportional to $\log^3(1/\epsilon)$. One should note that this result is weaker than the one obtained for seed paths and seed stars above in various ways. First, unlike in the cases of Theorems 1.1 and 1.3, here we cannot guarantee that almost all of the seed tree is identified, but only a fraction of it whose size depends on ϵ–although in a mild manner. Second, the size of the seed tree needs to be somewhat larger as a function of ϵ as before. While in the previous cases ℓ needed to be logarithmic in $1/\epsilon$, now it needs to scale as $\log^3(1/\epsilon)$. Below we show that to some extent these weaker results are inevitable and that finding the seed tree T_ℓ is inherently harder than finding more structured seed trees such as stars and paths.

Our main positive result is as follows.

Theorem 1.5. *Let T_n be a uniform random recursive tree on n vertices and let $\epsilon > 0$ and $\ell \ge 1$. Let $a = 2\log(4\ell^2/\epsilon) + 1$. If ℓ is so large that*

$$\ell \ge 64a^2 \log(22a\ell^2/\epsilon) \ ,$$

then there exists a seed-finding algorithm that outputs a vertex set $H_n \subset \{1, \dots, n\}$ with $|H_n| \ge \ell/(3a)$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left\{H_n \subset T_\ell\right\} \ge 1 - \epsilon \ .$$

Note that the condition for ℓ is satisfied for $\ell \geq C \log^2(1/\epsilon) \log\log(1/\epsilon)$ for a constant C.

Note that Theorems 1.1 and 1.5 establish the existence of seed-finding algorithms of the first kind (i.e., outputting a subset of the seed), whereas Theorem 1.3 is on an algorithm of the second kind (that outputs a superset of the seed). Naturally, in all these cases, one may consider seed-finding algorithms of the other kind as well. These often lead to quite different considerations and our aim was to provide a sample of ideas, as opposed to give a complete characterization. The recent paper of Devroye and Reddad [9] studies seed finding algorithms of the second kind for general seeds. They also introduce a third type of goal where the objective is to output a set of vertices that contains *at least one* vertex of the seed tree.

Next we show that, regardless how large ℓ is, for n sufficiently large any seed-finding algorithm of the first kind needs to output a set of vertices whose size is at most $c\ell$ where c is strictly smaller than 1. Similarly, any seed-finding algorithm of second kind needs to output a set of vertices whose size is at least $C\ell$ where $C > 1$.

In other words, when the seed tree is a uniform random recursive tree, the problem of finding it is strictly harder than finding a seed path or a seed star in the sense that no algorithm can have a performance as the one established in Theorem 1.1 or Theorem 1.3. Note however, that there remains a gap between the performance bound of Theorem 1.5 and the impossibility bound of Theorem 1.6 below, as the size of the vertex set in the seed found by the algorithm of Theorem 1.5 is only guaranteed to be of the order of $\ell/\log(1/\epsilon)$, a linear fraction but depending on ϵ.

The impossibility results mentioned above follow from the fact that, at time 2ℓ, a linear fraction of the vertices of the seed T_ℓ become indistinguishable from vertices that arrive between time $\ell + 1$ and 2ℓ. To make the statement precise, we need a few definitions.

In a uniform random recursive tree T_ℓ, we call a vertex a *singleton* if it is a leaf and it is the only descendant of its parent vertex.

Now consider a vertex v in T_ℓ and its position in the tree $T_{2\ell}$. We say that v is a *camouflaging* vertex if

1. In T_ℓ, v is a parent of a singleton d;
2. Between time $\ell + 1$ and 2ℓ a vertex w is attached to v such that w is a leaf of $T_{2\ell}$
3. d is a leaf of $T_{2\ell}$.

Clearly, at time 2ℓ, and therefore at any time $n \geq 2\ell$, the two descendants d and w of any camouflaging vertex v are indistinguishable. Let G_ℓ denote the number of camouflaging vertices. Then if a seed-finding algorithm of the first kind outputs a vertex set that contains an $(1-\gamma)\ell$ vertices of the seed, then one must have $G_\ell < \gamma\ell$. Similarly, if a seed-finding algorithm of the second kind outputs a vertex set of size less than $(1-\gamma)\ell$ that contains the seed, then, necessarily, $G_\ell < \gamma\ell$.

The next proposition shows that $\gamma \geq 1/384$ with high probability.

Theorem 1.6. *For any $\ell \geq 2$,*

$$\mathbb{E}G_\ell \geq \frac{\ell}{384}$$

and for any $t \geq 0$,

$$\mathbb{P}\left\{G_\ell \leq \frac{\ell}{384} - t\right\} \leq e^{\frac{-t^2}{2\ell}} .$$

2 Proofs

In this section we present the proofs of all theorems. The construction of all seed-finding algorithms uses a simple notion of centrality that we recall first.

2.1 Centrality

Let T be a tree with vertex set $V(T)$. A *rooted tree* (T, v) is the tree T with a distinguished vertex $v \in V(T)$. For a vertex $u \in V(T)$, denote by $(T, v)_{u\downarrow}$ the rooted subtree of T whose root is u and whose vertex set contains all vertices w of $V(T)$ such that the (unique) path connecting w and v in T contains u.

Given tree T, the *anti-centrality* of a vertex $v \in V(T)$ is defined by

$$\psi(v) = \max_{u \in V(T)\setminus\{v\}} |(T, v)_{u\downarrow}| \ .$$

Thus, $\psi(v)$ is the size of the largest subtree of the tree T rooted at v. Note that leaves of a tree T have the largest anti-centrality with $\psi(v) = |V(T)| - 1$. We say that v is *at least as central as* w if $\psi(v) \le \psi(w)$.

For a positive integer k, we denote by $H_\psi(k)$ the set of k vertices with smallest anti-centrality, where ties may be broken arbitrarily.

This notion of centrality played a crucial role in some of the root-finding algorithms of [4]. We refer to Jog and Loh [15, 14] for a study of this notion in various random tree models, including uniform random recursive trees.

2.2 Proof of Theorem 1.1

Let ϵ, γ, and ℓ be as in the assumptions of the theorem. We may assume, without loss of generality, that $\gamma\ell/2$ is an integer. We analyze a simple seed-finding algorithm that achieves the performance stated in the theorem. The proposed algorithm simply takes the $(1-\gamma)\ell$ most central vertices, as measured by the function ψ defined in Section 2.1.

Formally, let $k_\ell = (1-\gamma)\ell$ and define $H_n = H_\psi(k_\ell)$ be the set of k_ℓ most central vertices of the observed tree T_n.

It suffices to prove that, for all sufficiently large n, with probability at least $1-\epsilon$, all vertices of T_n not in the seed P_ℓ are less central than any vertex in P_ℓ whose distance to the leaves of P_ℓ is at least $\gamma\ell/2$, that is,

$$\mathbb{P}\left\{ \min_{\ell < i \le n} \psi(i) > \max_{\ell\gamma/2 \le j \le \ell(1-\gamma/2)} \psi(j) \right\} \ge 1 - \epsilon \ . \tag{2.1}$$

(Recall that the vertex set of the seed P_ℓ is $\{1, \dots, \ell\}$.)

Let $C_1, \dots, C_\ell$ denote the components of the forest obtained by removing the edges of P_ℓ from T_n such that $k \in C_k$ for $k = 1, \dots, \ell$. Then

$$\begin{aligned} \mathbb{P}\left\{ \min_{\ell < i \le n} \psi(i) \le \max_{\ell\gamma/2 \le j \le \ell(1-\gamma/2)} \psi(j) \right\} &\le \sum_{j=\gamma\ell/2}^{(1-\gamma/2)\ell} \mathbb{P}\left\{ \min_{\ell < i \le n} \psi(i) \le \psi(j) \right\} \\ &\le \sum_{j=\gamma\ell/2}^{(1-\gamma/2)\ell} \sum_{k=1}^{\ell} \mathbb{P}\left\{ \exists v \in C_k \setminus \{k\} : \psi(v) \le \psi(j) \right\} \ . \end{aligned}$$

To bound the probabilities on the right-hand side, suppose, without loss of generality, that $k \le j$. (The case $k > j$ is analogous.) Let $v \in C_k \setminus \{k\}$ be such that $\psi(v) \le \psi(j)$. Let u be a vertex connected to v such that $|(T, v)_{u\downarrow}|$ is maximal (i.e., $\psi(v) = |(T, v)_{u\downarrow}|$). Then there are two possibilities:

(a) $(T, v)_{u\downarrow}$ is contained in C_k. In this case $|C_k| \ge \sum_{i \ne k} |C_i|$;

(b) $(T, v)_{u\downarrow} = \left(\bigcup_{i=1, i\ne k}^{\ell} C_i \right) \cup C'_k$ for some $C'_k \subset C_k$. In this case

$$\left| \bigcup_{i \ne k} C_i \right| \le \psi(v) \le \psi(j) \le \left| \bigcup_{i=1}^{j} C_i \right|$$

which implies $\sum_{i=j+1}^{\ell} |C_i| \leq |C_k|$.

By this observation, we have

$$\begin{aligned}\mathbb{P}\left\{\exists v \in C_k \setminus \{k\} : \psi(v) \leq \psi(j)\right\} &\leq \mathbb{P}\left\{|C_k| \geq \sum_{i\neq k} |C_i|\right\} + \mathbb{P}\left\{\sum_{i=j+1}^{\ell} |C_i| \leq |C_k|\right\} \\ &\leq \mathbb{P}\left\{|C_k| \geq \sum_{i\neq k} |C_i|\right\} + \mathbb{P}\left\{\sum_{i=(1-\gamma/2)\ell}^{\ell} |C_i| \leq |C_k|\right\}\end{aligned}$$

Now let $t = \gamma/e^2$. Then the right-hand side of the inequality above may be bounded further by

$$\mathbb{P}\left\{\sum_{i=1, i\neq k}^{\ell} |C_i| \leq nt\right\} + \mathbb{P}\left\{\sum_{i=1}^{\gamma\ell} |C_i| \leq nt\right\} + 2\mathbb{P}\left\{|C_k| \geq nt\right\}$$

Thus, we have

$$\begin{aligned}&\mathbb{P}\left\{\min_{\ell<i\leq n} \psi(i) \leq \max_{\ell\gamma/2\leq j\leq \ell(1-\gamma/2)} \psi(j)\right\} \\ &\quad\leq (1-\gamma)\ell^2 \left(\mathbb{P}\left\{\sum_{i=1, i\neq k}^{\ell} |C_i| \leq nt\right\} + \mathbb{P}\left\{\sum_{i=1}^{\gamma\ell} |C_i| \leq nt\right\} + 2\mathbb{P}\left\{|C_k| \geq nt\right\}\right)\end{aligned}$$

To understand the behavior of the probabilities on the right-hand side, note that, for any $k = 1, \ldots, \ell-1$, $\sum_{i=1}^{k} |C_i|$ is just the number of red balls after taking n samples in a standard Pólya urn initialized with k red and $\ell - k$ blue balls. This implies that $\sum_{i=1}^{k} |C_i|/n$ converges, in distribution, to a Beta$(k, \ell - k)$ random variable. Hence,

$$\lim_{n\to\infty} \mathbb{P}\left\{|C_k|/n \geq t\right\} = (1-t)^{\ell-1} \leq e^{-t(\ell-1)}$$

and

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left\{\sum_{i=1, i\neq k}^{\ell} |C_i|/n \leq t\right\} &\leq \lim_{n\to\infty} \mathbb{P}\left\{\sum_{i=1}^{\gamma\ell} |C_i|/n \leq t\right\} \\ &= (\ell-1)\binom{\ell-1}{\gamma\ell-1} \int_0^t x^{\gamma\ell-1}(1-x)^{\ell-\gamma\ell-1} dx\,.\end{aligned}$$

We may bound the expression on the right-hand side by

$$\frac{\ell^{\gamma\ell}}{(\gamma\ell-1)!} \int_0^t x^{\gamma\ell-1} dx = \frac{(t\ell)^{\gamma\ell}}{(\gamma\ell)!} \leq \left(\frac{e\ell t}{\gamma\ell}\right)^{\gamma\ell} \leq e^{-\gamma\ell}\,,$$

where we used Stirling's formula and the choice $t = \gamma/e^2$. Putting everything together, we have that

$$\limsup_{n\to\infty} \mathbb{P}\left\{\min_{\ell<i\leq n} \psi(i) \leq \max_{\ell\gamma/2\leq j\leq \ell(1-\gamma/2)} \psi(j)\right\} \leq 2\ell^2 \left(e^{-\gamma\ell} + e^{-\gamma(\ell-1)/e^2}\right) \leq \epsilon$$

under our conditions for ℓ, as desired. □

2.3 Proof of Theorem 1.2

Let E be the event that either (1) vertex i attaches to vertex $i-1$ for all $i = \ell+1, \dots, 2\ell$ or (2) vertex $\ell+1$ attaches to vertex 1 and for all $i = \ell+2, \dots, 2\ell$, vertex i attaches to vertex $i-1$. On this event, $T_{2\ell}$ is a path of 2ℓ vertices such that the seed P_ℓ is on one of the two extremes of $T_{2\ell}$. The probability of this event is

$$\frac{2}{\ell} \cdot \frac{1}{\ell+1} \cdot \dots \cdot \frac{1}{2\ell-1} \geq 2\frac{\ell!}{(2\ell)!} \geq 2(2\ell)^{-\ell} .$$

On this event, for $n \geq 2\ell$, for any seed-finding algorithm, the first and second halves of the path $T_{2\ell}$ are indistinguishable. At least one of the two halves of $T_{2\ell}$ is such that H_n intersects that half in at most $\ell/2$ vertices. Thus, (conditionally on E), the algorithm misses at least half of the seed path, with probability $1/2$. Hence

$$\mathbb{P}\left\{|H_n \cap P_\ell| \leq \frac{\ell}{2}\right\} \geq \frac{\mathbb{P}\{E\}}{2} \geq (2\ell)^{-\ell} \geq \epsilon$$

whenever $\ell \leq \frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}$ and $\epsilon \leq e^{-e^2}$.

2.4 Proof of Theorem 1.3

Let $k_\ell = (1+\gamma)\ell$. Again, we may assume that k_ℓ is an integer. The seed finding algorithm we propose is slightly different. It is specifically tailored to the case when the seed tree to be found is a star. Let $v_n^* = \operatorname{argmin}_{i=1,\dots,n} \psi(i)$ be the most central vertex of T_n. We define H_n as the set of vertices that includes v_n^* and $k_\ell - 1$ other vertices j with largest value of $|(T_n, v_n^*)_{j\downarrow}|$ among the neighbors of v_n^* in T_n. In other words, the algorithm outputs the most central vertex v_n^* and those neighbors whose subtree away from v_n^* is largest.

First we recall that by Jog and Loh [15, Theorem 4], there exists a numerical constant C such that, if $\ell \geq C\log(1/\epsilon)$ and the uniform attachment tree is initialized with a star E_ℓ as seed of ℓ vertices and central vertex 1, then

$$\mathbb{P}\left\{v_n^* = 1 \text{ for all } n = \ell+1, \ell+2, \dots\right\} \geq 1 - \frac{\epsilon}{2} ,$$

that is, with probability at least $1-\epsilon/2$, the center of the seed star remains the most central vertex of T_n for all n.

Let $v_1 \leq v_2 \leq \cdots$ be the vertices that are attached to vertex 1 (i.e., to the center of the seed star E_ℓ) in the uniform attachment process. (Thus, $v_1 > \ell$.) In view of the above-mentioned result of Jog and Loh, it suffices to show that for all n sufficiently large, all vertices v_j with $j > \gamma\ell$ have $\left|(T_n, 1)_{v_j\downarrow}\right|$ smaller than $|(T_n, 1)_{i\downarrow}|$ for all vertices i in the seed star E_ℓ, with probability at least $1-\epsilon/2$. Thus, writing $g(i) = |(T_n, 1)_{i\downarrow}|$, we need to prove that

$$\limsup_{n\to\infty} \mathbb{P}\left\{\max_{j>\gamma\ell} g(v_j) < \min_{i=2,\dots,\ell} g(i)\right\} > 1 - \frac{\epsilon}{2} . \tag{2.2}$$

To prove (2.2), first we write

$$\mathbb{P}\left\{\max_{j>\gamma\ell} g(v_j) \geq \min_{i=2,\dots,\ell} g(i)\right\} \leq \mathbb{P}\left\{v_{\gamma\ell+1} \leq m\right\} + \mathbb{P}\left\{\max_{v_j>m} g(v_j) \geq \min_{i=2,\dots,\ell} g(i)\right\} , \tag{2.3}$$

where we take $m = \lfloor e^{\gamma\ell/4} \rfloor$. The first term on the right-hand side is the probability that more than $\gamma\ell$ vertices are attached to vertex 1 up to time m. In order to bound this probability, denote by X_t, for $t \geq \ell$, the number of vertices attached to vertex 1 between time $\ell+1$ and t. Thus, $X_\ell = 0$ and

$$\mathbb{P}\left\{v_{\gamma\ell+1} \leq m\right\} = \mathbb{P}\left\{X_m > \gamma\ell\right\} .$$

Since

$$\mathbb{E}[X_t|X_{t-1}] = X_{t-1} + \frac{1}{t}\,,$$

$$Y_t = X_t - \sum_{k=\ell+1}^{t} \frac{1}{k},\ \ t \geq \ell+1$$

is a martingale with respect to the filtration generated by $X_\ell, X_{\ell+1}, \ldots$. Denote the corresponding martingale difference sequence by $Z_t = Y_t - Y_{t-1} = X_t - X_{t-1} - 1/t$. By Markov's inequality,

$$\mathbb{P}\left\{X_m > \gamma\ell\right\} = \mathbb{P}\left\{\sum_{j=\ell+1}^{m} Z_j + \sum_{j=\ell+1}^{m} \frac{1}{j} > \gamma\ell\right\} \leq \frac{e^{\sum_{j=\ell+1}^{m} \frac{1}{j}} \cdot \mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}\right]}{e^{\gamma\ell}}\,. \tag{2.4}$$

In order to bound the right-hand side, observe that

$$\begin{aligned}
\mathbb{E}\left[e^{Z_m}|X_\ell, \ldots, X_{m-1}\right] &= \mathbb{E}\left[e^{X_m - X_{m-1} - \frac{1}{m}}|\ X_\ell, \ldots, X_{m-1}\right] \\
&= e^{-X_{m-1}-\frac{1}{m}}\mathbb{E}\left[e^{X_m}|\ X_\ell, \ldots, X_{m-1}\right] \\
&= e^{-X_{m-1}-\frac{1}{m}}\left(\frac{1}{m}e^{X_{m-1}+1} + \frac{(m-1)}{m}e^{X_{m-1}}\right) \\
&= \frac{e^{-\frac{1}{m}}}{m}(e+m-1) \\
&\leq \frac{(m+2)e^{-\frac{1}{m}}}{m}\,,
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}|X_\ell, \ldots, X_{m-1}\right]\right] \\
&= \mathbb{E}\left[e^{\sum_{j=\ell+1}^{m-1} Z_j}\mathbb{E}\left[e^{Z_m}|X_\ell, \ldots, X_{m-1}\right]\right] \\
&\leq \frac{(m+2)e^{-\frac{1}{m}}}{m}\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m-1} Z_j}\right]\,.
\end{aligned}$$

Thus, by induction we obtain

$$\mathbb{E}\left[e^{\sum_{j=\ell+1}^{m} Z_j}\right] \leq \frac{(m+2)^2}{\ell^2}e^{-\sum_{j=\ell+1}^{m}\frac{1}{j}}.$$

Substituting into (2.4), we get

$$\mathbb{P}\left\{v_{\gamma\ell+1} \leq m\right\} = \mathbb{P}\left\{X_m > \gamma\ell\right\} \leq \frac{(m+2)^2}{\ell^2 e^{\gamma\ell}} \leq \frac{\epsilon}{4}$$

by our choice of m and by the condition on the value of ℓ. Hence, by (2.3), it suffices to show that

$$\mathbb{P}\left\{\max_{v_j > m} g(v_j) \geq \min_{i=2,\ldots,\ell} g(i)\right\} \leq \frac{\epsilon}{4}\,.$$

We proceed by writing

$$\mathbb{P}\left\{\max_{v_j > m} g(v_j) \geq \min_{i=2,\ldots,\ell} g(i)\right\} \leq \sum_{i=2}^{\ell} \mathbb{P}\left\{\max_{v_j > m} g(v_j) \geq g(i)\right\}\,.$$

Now fix $i \in \{2, \ldots, \ell\}$ and notice that $\max_{v_j > m} g(j)$ is bounded by the number of vertices A attached to the tree formed by vertex 1 and all vertices in the subtrees $(T_n, 1)_{j\downarrow}$ for $j > m$ such that vertex j is attached to vertex 1.

Denoting $B = g(i)$ and $C = n - A - B$, note that, conditioned on the tree T_m, the triple (A, B, C) behaves as the number of red, blue, and white balls in a Pólya urn in which initially (i.e., at time m) there is one red ball, $B_m = |(T_m, 1)_{i\downarrow}|$ blue balls, and $m - 1 - |(T_m, 1)_{i\downarrow}|$ white balls. Hence, for each $i = 2, \dots, \ell$, we have

$$\begin{aligned}\mathbb{P}\left\{\max_{v_j > m} g(v_j) \geq g(i)\right\} &\leq \mathbb{P}\{A > B\} \\ &\leq \mathbb{P}\left\{A > B | B_m \geq \frac{m\epsilon}{32\ell^2}\right\} + \mathbb{P}\left\{B_m < \frac{m\epsilon}{32\ell^2}\right\} .\end{aligned}$$

In order to bound the second term on the right-hand side, note that by the standard theory of Pólya urns, B_m has a beta-binomial distribution with parameters $(m, 1, \ell - 1)$. Thus, B_m is distributed as a binomial random variable $\mathrm{Bin}(m, \pi)$ where the parameter π is an independent $\mathrm{Beta}(1, \ell - 1)$ random variable. Thus,

$$\begin{aligned}&\mathbb{P}\left\{B_m < \frac{m\epsilon}{32\ell^2}\right\} \\ &\leq \mathbb{P}\left\{\mathrm{Bin}(m, \epsilon/16\ell^2) < \frac{m\epsilon}{32\ell^2}\right\} + \mathbb{P}\left\{\pi < \frac{\epsilon}{16\ell^2}\right\} \\ &\leq e^{-m\epsilon/(128\ell^2)} + 1 - \left(1 - \frac{m\epsilon}{16\ell^2}\right)^{\ell-1} \\ &\quad \text{(by a standard binomial estimate and expressing the beta distribution)} \\ &\leq e^{-m\epsilon/(128\ell^2)} + \frac{\epsilon}{16\ell} \\ &\quad \text{(by the Bernoulli inequality)} \\ &\leq \frac{\epsilon}{8\ell}\end{aligned}$$

whenever $\ell > (4\gamma)\left(\log(1/\epsilon) + \log\log(8\ell/\epsilon) + \log(128\ell^2)\right)$. To finish the proof it remains to show that

$$\limsup_{n\to\infty} \mathbb{P}\left\{A > B | B_m \geq \frac{m\epsilon}{32\ell^2}\right\} \leq \frac{\epsilon}{8\ell} .$$

But this follows from the fact that this limiting probability is bounded by the probability that a $\mathrm{Beta}(1, m\epsilon/32\ell^2)$ random variable is greater than $1/2$ which is at most $2^{-m\epsilon/32\ell^2}$. Since $m = \lfloor e^{\gamma\ell/4} \rfloor$, this is bounded by $\epsilon/(8\ell)$ for $\ell > (8/\gamma \vee C)\log(1/\epsilon)$, as desired. □

2.5 Proof of Theorem 1.5

Fix $\epsilon \in (0, 1)$ and define $a = 2\log(4/\epsilon) + 1$ and $k_\ell = \frac{\ell}{3a}$. A seed-finding algorithm with the desired property simply selects the k_ℓ most central vertices. (Again, for simplicity of the presentation, we assume that k_ℓ is an integer.) With the notation introduced at the beginning of this section, we define $H_n = H_\psi(k_\ell)$. We need to show that the k_ℓ most central vertices of T_n are in T_ℓ with probability at least $1 - \epsilon$ for all sufficiently large n.

The strategy of our proof is as follows. First we show that, with probability at least $1 - \epsilon/2$, the seed T_ℓ contains at least k_ℓ "deep" vertices. Then we prove that for all n sufficiently large, all deep vertices of T_ℓ are more central in T_n than any vertex outside of the seed T_ℓ.

We call a vertex $v \in T_\ell$ *deep* if it has at least a descendants, that is, if

$$|(T_\ell, 1)_{v\downarrow}| \geq a + 1 .$$

Denote by $\mathcal{A}_\ell$ the set of all deep vertices of T_ℓ. Noticing that

$$\mathbb{P}\{H_n \not\subset T_\ell\} \leq \mathbb{P}\{|\mathcal{A}_\ell| \leq k_\ell\} + \mathbb{P}\{\exists v \in V(T_n)\backslash V(T_\ell), \exists u \in \mathcal{A}_\ell : \psi_n(v) \leq \psi_n(u)\} ,$$

it suffices to show that

$$\mathbb{P}\{|\mathcal{A}_\ell| \leq k_\ell\} \leq \frac{\epsilon}{2} . \tag{2.5}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left\{\exists v \in V(T_n)\backslash V(T_\ell), \exists u \in \mathcal{A}_\ell : \psi_n(v) \le \psi_n(u)\right\} \le \frac{\epsilon}{2} \ . \tag{2.6}$$

(2.5) follows from inequality (3.1) in the Appendix under the condition $\ell \ge 64a^2 \log(22a/\epsilon)$.

It remains to prove (2.6). To this end, for $i \in \{1, \dots, \ell\}$, denote by C_i the component of vertex i in the forest obtained by removing the edges of T_ℓ from T_n. Then

$$\begin{aligned}&\mathbb{P}\left\{\exists v \in V(T_n)\backslash V(T_\ell), \exists u \in \mathcal{A}_\ell : \psi(v) \le \psi(u) | T_\ell\right\} \\ &\quad \le \sum_{u\in\mathcal{A}_\ell} \sum_{k=1}^{\ell} \mathbb{P}\left\{\exists v \in C_k\backslash\{k\} : \psi(v) \le \psi(u) | T_\ell\right\} \ .\end{aligned}$$

Now fix T_ℓ and vertices $k \in \{1, \dots, \ell\}$ and $u \in \mathcal{A}_\ell$. For any vertex $v \in C_k\backslash\{k\}$ such that $\psi(v) \le \psi(u)$, there are two possibilities:
(1) either the largest subtree of T_n rooted at v is inside C_k, in which case $|C_k| \ge \sum_{i\neq k} |C_i|$;
(2) or the largest subtree of T_n rooted at v is $\left(\bigcup_{i=1, i\neq k}^{\ell} C_i\right) \cup C'_k$ for some $C'_k \subset C_k$. In this case, $\psi(v) \le \psi(u)$ implies that

$$\sum_{i\in T_n\backslash(T_\ell, v)_{u\downarrow}} |C_i| \le |C_k| \ .$$

Since $u \in A_\ell$, this means that the left-hand side is dominated by the number of red balls in a standard Pólya urn with after $n - \ell$ draws initialized with at least a red, one blue, and $n - a - \ell - 1$ white balls; while $|C_k|$ behaves like the number of blue balls in the same urn.

By the same calculations as in the proof of Theorem 1.1, the probability of case (1) may be bounded by

$$\limsup_{n\to\infty} \mathbb{P}\left\{|C_k| \ge \sum_{i\neq k} |C_i| | T_\ell\right\} = \limsup_{n\to\infty} \mathbb{P}\left\{|C_k| \ge (n-\ell)/2 | T_\ell\right\} \le e^{-(\ell-1)/2} \le \frac{\epsilon}{4\ell^2} \ .$$

Similarly, the probability of case (2) satisfies

$$\limsup_{n\to\infty} \mathbb{P}\left\{\sum_{i\in T_n\backslash(T_\ell, v)_{u\downarrow}} |C_i| \le |C_k| | T_\ell\right\} \le e^{-(a-1)/2} \le \frac{\epsilon}{4\ell^2}$$

by our choice $a = 2\log(4\ell^2/\epsilon) + 1$. This concludes the proof of (2.6) and hence that of Theorem 1.5.

2.6 Proof of Theorem 1.6

We prove the lower bound for the expected number of camouflaging vertices by induction. To this end, fix a singleton d and its parent v in T_ℓ. For $j \ge \ell$, let

$$E_j^{(v)} = \left\{\exists d' \in V(T_j)\backslash\{d\} : d' \sim v \text{ and } d', d \text{ are leaves in } T_j\right\} \ .$$

Observe that $E_{2\ell}^{(v)}$ is the event that v is a camouflaging vertex. Consider the sequences

$$\begin{aligned}a_j &= \mathbb{P}\left\{E_j^{(v)} | T_\ell\right\} \\ c_j &= \mathbb{P}\left\{d \text{ is a singleton in } T_j | T_\ell\right\} \ .\end{aligned}$$

Now, observe that the event $E_{j+1}^{(v)}$ occurs if $E_j^{(v)}$ occurs and the vertex $j+1$ is neither attached to d nor to d', or if d is a singleton of T_j and the $j+1$ is attached to v. Thus

$$a_{j+1} = a_j \cdot \left(1 - \frac{2}{j}\right) + c_j \cdot \frac{1}{j} \; .$$

Multiplying both sides by $j(j-1)$, we get

$$j(j-1)a_{j+1} = (j-1)(j-2)a_j + (j-1)c_j \; .$$

Summing over $j = \ell+1, \ldots, 2\ell-1$,

$$(2\ell-1)(2\ell-2)a_{2\ell} = \ell(\ell-1)a_{\ell+1} + \sum_{j=\ell+1}^{2\ell-1} (j-1)c_j \; ,$$

which implies that

$$a_{2\ell} \geq \frac{1}{(2\ell-1)(2\ell-2)} \sum_{j=\ell+1}^{2\ell-1} (j-1)c_j \geq \frac{1}{4(\ell-1)} \sum_{j=\ell+1}^{2\ell-1} c_j \; .$$

Note that, for $j \in \{\ell+1, \ldots, 2\ell-1\}$,

$$\begin{aligned} c_j &= \prod_{k=\ell}^{j-1} \left(1 - \frac{2}{k}\right) \\ &\geq \exp\left(-4 \sum_{k=\ell}^{j-1} \frac{1}{k}\right) \quad \text{(since } 1 - x \geq e^{-2x} \text{ for } x < 3/4\text{)} \\ &\geq \exp\left(4\log \ell - 4 \log j\right) \\ &> \frac{\ell^4}{(2\ell)^4} = \frac{1}{16} \; , \end{aligned}$$

and therefore

$$a_{2\ell} \geq \frac{1}{4(\ell-1)} \sum_{j=\ell+1}^{2\ell-1} c_j \geq \frac{1}{64} \; .$$

Let P_ℓ be the set of vertices in T_ℓ that are parents of a singleton. Then

$$\begin{aligned} \mathbb{E}[G_\ell | T_\ell] &= \mathbb{E}\left[\sum_{v \in P_\ell} 1_{E_{2\ell}^{(v)}} | T_\ell\right] \\ &= \sum_{v \in P_\ell} \mathbb{P}\left\{E_{2\ell}^{(v)} | T_\ell\right\} \\ &\geq \frac{1}{64} |P_\ell| \; , \end{aligned}$$

which implies that $\mathbb{E}G_\ell \geq \frac{1}{64}\mathbb{E}|P_\ell|$.

It remains to bound the expected number of singletons $\mathbb{E}|P_\ell|$ in the uniform random recursive tree T_ℓ. Write $S_k = |P_k|$ and note that S_k equals the number of parents of singletons in T_k.

When a new vertex is attached to the tree T_k, we lose one singleton if the new vertex is attached to the parent of a singleton. This happens with probability S_k/k. If a the new vertex is attached to a singleton, then the number remains the same. If the new vertex

is attached to some vertex that is not a leaf nor a parent of a singleton, then, the number of singletons also remains unchanged. Finally, if the new vertex is attached to a leaf that is not a singleton, the number of singletons increases by 1. Thus, denoting the number of leaves of T_k by L_k,

$$\begin{aligned}\mathbb{E}[S_{k+1}|T_k] &= (S_k-1)\frac{S_k}{k} + S_k\left(\frac{S_k}{k} + 1 - \frac{S_k}{k} - \frac{L_k}{k}\right) + (S_k+1)\left(\frac{L_k}{k} - \frac{S_k}{k}\right)\\ &= \left(1-\frac{2}{k}\right)S_k + \frac{L_k}{k}\ .\end{aligned}$$

Taking expectations and using the fact that $\mathbb{E}L_k = k/2$, we have that $\mathbb{E}S_\ell = \ell/6$. Summarizing, the expected number of camouflaging vertices satisfies

$$\mathbb{E}G_\ell \geq \frac{1}{64}\cdot\frac{\ell}{6} = \frac{\ell}{384}\ .$$

We prove the second inequality of Theorem 1.6 using the *bounded differences inequality* of McDiarmid [16] (see also [2, Theorem 6.2]).

Observe that given T_ℓ, there is a bijection between the set of recursive trees of size 2ℓ containing T_ℓ as subgraph and the set $\mathcal{S} = [\ell]\times\cdots\times[2\ell-1]$. The bijection is simply given by associating the vector $\kappa = (a_{\ell+1},\cdots,a_{2\ell})$ to the recursive tree $T(\kappa)$ where the vertex $k\in[\ell+1,2\ell]$ is attached to the vertex a_k, starting by T_ℓ until obtaining $T_{2\ell}$. Then we may consider the set $\mathcal{S}$ as the set of recursive trees with 2ℓ vertices that contain T_ℓ as subtree.

Importantly, the components of κ that represent the uniform random recursive tree $T_{2\ell}$ are independent random variables.

Given T_ℓ, consider the function $g:\mathcal{S}\to\mathbb{R}$ such that $g(T_{2\ell})$ is the number of camouflaging vertices.

By the bounded differences inequality, it suffices to show that, given $T, T'\in\mathcal{S}$, if T and T' differ by exactly one coordinate, then $|g(T)-g(T')|\leq 2$.

To this end, let $v\in V(T_n)$ be a parent of a singleton d. v is a camouflaging vertex of a tree $T=(a_{\ell+1},\cdots,a_{2\ell})$ if and only if

1. $d\notin\{a_{\ell+1},\cdots,a_{2\ell}\}$;
2. $\exists k\in\{\ell+1,\cdots,2\ell\}\backslash\{a_{k+1},\cdots,a_{2\ell}\}$ such that $a_k = v$.

Now, consider $T=(a_{\ell+1},\cdots,a_{2\ell})$, $T'=(b_{\ell+1},\cdots,b_{2\ell})$ two trees with $a_r\neq b_r$ for some r and $a_j = b_j$ for $j\neq r$. For a camouflaging vertex v in T (with corresponding singleton d in T_ℓ) not to be a camouflaging vertex in T', it is necessary (but not sufficient) that either

1. b_r is a child of v,
2. or $a_r = v$.

Similarly, for a not camouflaging vertex v in T (with corresponding singleton d in T_ℓ), to be a camouflaging vertex in T' it is necessary that either

1. a_r is a descendant of v,
2. or $b_r = v$.

Thus, $|g(T)-g(T')|\leq 2$, and the bounded differences condition is satisfied, proving the second inequality of Theorem 1.6.

3 Appendix

Devroye [8] proved a central limit theorem for the number of vertices with k descendants in a uniform random recursive tree. In particular, if $L_{k,n}$ denotes the the number of vertices with k descendants in a uniform random recursive tree of $n > k+1$ vertices, then Devroye shows that

$$\mathbb{E}L_{k,n} = \frac{n-k-1}{(k+1)(k+2)} + \frac{1}{k+1} = \frac{n+1}{(k+1)(k+2)}$$

and, for any fixed k, as $n \to \infty$,

$$\frac{L_{k,n} - \frac{n}{(k+1)(k+2)}}{\sqrt{n\sigma_k^2}}$$

converges, in distribution, to a standard normal random variable, where

$$\sigma_k^2 = \frac{1}{(k+1)(k+2)}\left(1 - \frac{1}{(k+1)(k+2)}\right) - \frac{2}{(k+1)(k+2)^2} + \frac{1}{(k+1)^2(2k+3)} \,.$$

Devroye's proof is based on representing $L_{k,n}$ as a sum of $(k+1)$-dependent indicator random variables and on a central limit theorem of Hoeffding and Robbins [12] for such sums. In this paper we need a non-asymptotic version of Devroye's theorem. Quantitative, Berry-Esseen-type versions of the Hoeffding-Robbins limit theorem are available via Stein's method, see, for example, Rinott [18, Theorem 2.2]. On the other hand, a simple bound may be proved by combining Devroye's representation with a concentration inequality of Janson [13, Corollary 2.4] for sums of dependent random variables, to obtain the following:

Proposition 3.1. *If $L_{k,n}$ denotes the the number of vertices with k descendants in a uniform random recursive tree of $n > k+1$, then for all $t > 0$,*

$$\mathbb{P}\left\{L_{k,n} \geq \mathbb{E}L_{k,n} + t\right\} \leq \exp\left(\frac{-8t^2(k+2)}{25(n+(k+1)(k+2)t/3}\right)$$

and

$$\mathbb{P}\left\{L_{k,n} \leq \mathbb{E}L_{k,n} - t\right\} \leq \exp\left(\frac{-8t^2(k+2)}{25n}\right) \,.$$

Note that the number of vertices with *at least* k descendants $M_{k,n} = \sum_{i=k}^{n-1} L_{i,n} = n - \sum_{i=0}^{k-1} L_{i,n}$ has expected value

$$\mathbb{E}M_{k,n} = \mathbb{E}\sum_{i=k}^{n-1} L_{i,n} = n - \sum_{i=0}^{k-1} \mathbb{E}L_{i,n} = \frac{n+1}{k+1} - 1 \,,$$

and therefore

$$\begin{aligned}
\mathbb{P}\left\{M_{k,n} \leq \frac{n+1}{k+1} - 1 - t\right\} &= \mathbb{P}\left\{\sum_{i=0}^{k-1} L_{i,n} \geq \sum_{i=0}^{k-1} \mathbb{E}L_{i,n} + t\right\} \\
&\leq \sum_{i=0}^{k-1} \mathbb{P}\left\{L_{i,n} \geq \mathbb{E}L_{i,n} + \frac{t}{k}\right\} \\
&\leq k\exp\left(\frac{-8t^2}{25k(n+(k+1)t/3}\right) \,.
\end{aligned}$$

In particular, by generously bounding constants, we get

$$\mathbb{P}\left\{M_{k,n} \leq \frac{n}{3k}\right\} \leq k\exp\left(-\frac{1}{32}\frac{n}{k^2}\right) \,. \tag{3.1}$$

References

[1] Christian Borgs, Michael Brautbar, Jennifer Chayes, Sanjeev Khanna, and Brendan Lucier: The power of local information in social networks. *In Internet and Network Economics*, 406-419. Springer, 2012.

[2] S. Boucheron, G. Lugosi, and P. Massart: Concentration inequalities: A Nonasymptotic Theory of Independence. *Oxford University Press*, 2013. MR-3185193

[3] Michael Brautbar and Michael J. Kearns: Local algorithms for finding interesting individuals in large networks. *In Innovations in Theoretical Computer Science (ITCS)*, 2010.

[4] Sebastien Bubeck, Luc Devroye, and Gábor Lugosi: Finding Adam in random growing trees. *Random Structures & Algorithms*, 50(2): 158-172, 2017. MR-3607120

[5] Sebastien Bubeck, Ronen Eldan, Elchanan Mossel, and Miklós Rácz: From trees to seeds: on the inference of the seed from large trees in the uniform attachment model. *Bernoulli*, 23(4A):2887-2916, 2017. MR-3648049

[6] Sebastien Bubeck, Elchanan Mossel, and Miklós Z Rácz: On the influence of the seed graph in the preferential attachment model. *IEEE Transactions on Network Science and Engineering*, 2(1):30-39, 2015. MR-3361606

[7] Nicolas Curien, Thomas Duquesne, Igor Kortchemski, and Ioan Manolescu: Scaling limits and influence of the seed graph in preferential attachment trees. *Journal de l'Ecole Polytechnique-Mathématiques*, 2:1-34, 2015. MR-3326003

[8] Luc Devroye: Limit laws for local counters in random binary search trees. *Random Structures & Algorithms*, 2(3):303-315, 1991. MR-1109697

[9] Luc Devroye and Tommy Reddad: On the discovery of the seed in uniform attachment trees. arXiv:1810.00969, 2018.

[10] Michael Drmota: Random trees: an interplay between combinatorics and probability. *Springer Science & Business Media*, 2009. MR-2484382

[11] Alan Frieze and Wesley Pegden: Looking for vertex number one. *The Annals of Applied Probability*, 27(1):582-630, 2017 MR-3619796

[12] Wassily Hoeffding and Herbert Robbins: The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773-780, 1948. MR-0026771

[13] Svante Janson: Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234-248, 2004. MR-2068873

[14] Varun Jog and Po-Ling Loh: Analysis of centrality in sublinear preferential attachment trees via the CMJ branching process. *IEEE Transactions on Network Science and Engineering*, 2017. MR-3625951

[15] Varun Jog and Po-Ling Loh: Persistence of centrality in random growing trees. *Random Structures & Algorithms*, 2017 MR-3731614

[16] C. McDiarmid: On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148-188. Cambridge University Press, Cambridge, 1989. MR-1036755

[17] Saket Navlakha and Carl Kingsford: Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Computational Biology*, 7(4):e1001119, 2011. MR-2805381

[18] Yosef Rinott: On normal approximation rates for certain sums of dependent random variables. *Journal of Computational and Applied Mathematics*, 55(2):135-143, 1994. MR-1327369

[19] Devavrat Shah and Tauhid Zaman: Finding rumor sources on random trees. *Operations Research*, 64(3):736-755, 2016. MR-3515208

[20] Devavrat Shah and Tauhid R. Zaman: Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57(8):5163-5181, 2011 MR-2849111

Acknowledgments. We thank Luc Devroye, Miklós Rácz, and Tommy Reddad for interesting conversations on the topic of the paper.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System `http://www.vtex.lt/en/ejms.html`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`