ENS Rennes 2023-2024

INTERNSHIP REPORT

Root-finding algorithms in random trees.

Arthur Maritch-Roy

McGill

Foreword

This document is the report of the internship that marks the end of my second year at ENS Rennes. I did it under the supervision of Louigi Addario-Berry (McGill University, Montreal). I had to do it remotedly from France for personal reasons, and I sincerely want to thank Louigi Addario-Berry and François Bolley for giving me this possibility.

Despite this unforeseen event, this internship definitely taught me a lot on how to understand math, to learn math, and especially on how to write and explain math.

Contents

Introduction

A lot of situations such as a social network, a biological network, the World Wide Web can be modeled by a set of nodes connected by links. For example, if a node is a website, an oriented edge from a node to another one can model a link that directs to another website. Inside these situations, we may want to study the spread of some information (the spread of a rumor on Facebook, the spread of a genetic mutation), and more specifically, find the origin of such a phenomenon. Here, we focus on some phenomena that have a tree structure. The general principle is that we know the distribution related to the spread of the information, and we want to know how to find the node where the information began to spread. Such an algorithm is called a root-finding algorithm.

The tree generation models we study here are called preferential-attachment models. They are so called because a new node is more likely to connect with an other node if it has already some links connected to it. We do not use such models randomly. Indeed, we observed in some real-life cases (citations and references in research papers, see [1]) that the numbers of connections of

nodes are power-law distributed, and it is known [2] that such power-law distributions are likely to appear in such cases [1].

We want the algorithms to output a set of vertices that contain the root with an important probability, without the number of vertices depending on the size of the tree, and that is where it is challenging. This documents begins with a bunch of important results about trees that I needed to get familiar with this structure and most importantly to understand what follows. We can thus precisely define a root-finding algorithm. The main purpose of the document is to introduce the first root-finding algorithms that have ever been found for two important models (uniform and preferential attachment models, they have been found in 2015 by Lugosi, Devroye and Bubeck in [4]) and to introduce an algorithm that has been discovered later for the preferential attachment in 2023, and which is better than the previous one (Contat, Curien, Lacroix, Lasalle, Rivoirard [5]).

1 Trees and random trees

In this section, we introduce trees as a type of graph and then we introduce the different types of random trees on which we will apply root-finding algorithms.

1.1 Graphs and trees

1.1.1 General information about trees

A *graph* is a network of nodes called *vertices* that are connected by links we call *edges*. More precisely, graphs can be described as follows.

Definition 1. A couple $G = (V, E) = (V(G), E(G))$, with V a finite set and E a subset of $\mathcal{P}_2(E)$, where $\mathcal{P}_2(A)$ denotes all the subsets of A with two elements, is said to be a (undirected) *graph*. The size of the graph G is $|G| := |V|$.

One important notion we will use below is the *degree*.

Definition 2. For $G = (V, E)$ a graph, the *degree* of a vertex $v \in V$, noted $\mathrm{d}(v)$, is the number of edges $e \in E$ such that $v \in e$. It is the number of edges connected to v.

Remark 1. For any graph, we have the following:

$$\sum_{v \in V} \mathrm{d}(v) = 2|E|$$

We can represent a graph with a sagittal diagram like in Figure 1.

Now, a tree can be defined as a particular case of graph.

Definition 3. A graph $G = (E, V)$ is a *tree* if it is:

- connected: $\forall x \neq y \in V, \exists x_1, \ldots, x_k \in V, \{x, x_1\}, \ldots, \{x_k, y\} \in E$
- acyclic: $\forall x \in V, \forall x_1, \ldots, x_k \in V, \{x, x_1\}, \ldots, \{x_k, x\} \in E \Rightarrow k = 0$.

The vertices that are linked to a single other vertex are called *leaves*, and the subset of $V(T)$ containing the leaves of T will be denoted by $\mathcal{L}(T)$.

In other words, a tree is a graph in which there is a unique path between each couple of vertices. Such trees are called *unrooted trees*. When we introduce the notion of a root, we can talk about children and parents as we would do for a genealogical tree.

[1] For example, Zipf observed power-law distributions when he modeled the appearances of words in a text in a way such that a new word was added with a probability proportional to the number of appearances of this word in the text until then, see [3].

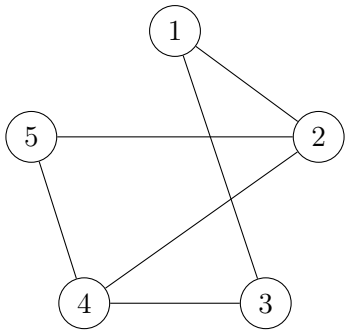

Figure 1: Sagittal diagram of the graph $G = (\llbracket 1, 5 \rrbracket, \{\{1,2\},\{1,3\},\{2,4\},\{2,5\},\{3,4\},\{4,5\}\})$. We have, for example, $\mathrm{d}(1) = 3, \mathrm{d}(5) = 2$.

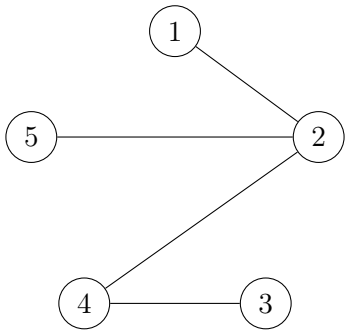

Figure 2: If we erase some edges from the previous graph, we obtain a tree.

Definition 4. A *rooted tree* is a couple (T, u) where T is a tree and $u \in V(T)$. The vertex u is called the *root*. Once we have a root, we can partially order the set of vertices as follows. If $\{u = x_0, x_1\}, \ldots, \{x_{k-1}, x_k\} \in E(T)$, we define:

$$x_i \preccurlyeq x_j \Leftrightarrow i \leq j$$

If $j = i + 1$, x_j is a *child* of x_i and x_i is the *parent* of x_j.

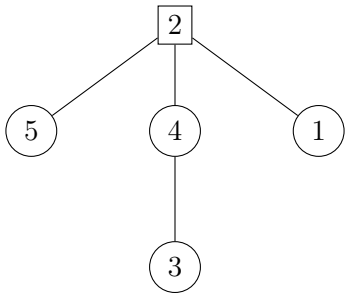

Figure 3: If we define a root, we can represent the tree like that, so we can see the tree structure. Here the tree is the same as in the previous figure, and the root is 2.

Definition 5. The *height* of a rooted tree (T, u) is the length longest path you can build starting from the root, noted $\mathrm{h}(T)$. The height of the tree of figure 3 is 3.

Definition 6. A *subgraph* of a graph G is a graph such that:

$$G' = (V', E' \subset \mathcal{P}_2(V'))$$

with $V' \subset V(G)$. Similarly, we can define a *subtree* of a tree T as being a subgraph of T which is a tree itself. Furthermore, if (T, u) is a rooted tree, we denote by $(T, u)_{v\downarrow}$ the subtree of T with all the vertices that are descendants of v, rooted in v.

Remark 2. Generally, a subgraph of a tree is not a tree and is called a *forest*. The connected components of a forest are always trees.

1.1.2 Some important families of trees

Here, we make a list of several groups of trees that will be useful later. Let us begin with the most general class of trees.

Definition 7. The trees as they have been defined above are called *labeled* trees. If there are n vertices, the set of vertices will be $[\![1, n]\!]$ by default.

Theorem 1. (Cayley's formula) There are n^{n-2} labeled trees with n vertices.

Proof. Let's denote by c_n the number of such trees. It corresponds to the number of ways we have to connect n points with $(n-1)$ edges. A way to calculate c_n is to count the number of sequences of edges that can be added to an empty graph (with no edge) to make a tree in two different ways.

- First, to get such a sequence, we can choose one of the c_n trees, and order the edges, which gives us $n!c_n$ sequences.
- Also, if we added $n-k$ edges already, we must have a forest of k trees. Then, for the next edges, we must choose one vertex among all the vertices (n choices) and another vertex among the roots of the trees we already have, excluding the tree containing the first vertex. By doing this, at each step, we keep having a forest and at the end, we have a tree. To do all this, we have a total of:
$$\prod_{k=2}^{n} n(k-1) = n^{n-2}n!.$$

All in all, we obtain that $c_n = n^{n-2}$. □

Definition 8. A *recursive tree* is a rooted tree with $V(T) = [\![1, n]\!]$ in which we have:

$$\forall i, j \in [\![1, n]\!], i \preccurlyeq j \Rightarrow i \leq j.$$

In particular, the root of a recursive tree must be 1.

Proposition 2. There are $(n-1)!$ recursive trees with n vertices.

Proof. For each $j \leq 2$, the vertex j has a unique parent in $[\![1, j-1]\!]$, so we can see a recursive tree as an element of:

$$A = [\![1]\!] \times [\![1, 2]\!] \times \ldots \times [\![1, n-1]\!],$$

so that there are $|A| = (n-1)!$ recursive trees. □

Definition 9. A *plane-oriented recursive tree* is a recursive tree given with an ordering of the children of each vertex. If we denote by $\mathrm{c}_T(j)$ the number of children of the vertex j, a plane-oriented recursive tree is given by a recursive tree T and an element of:

$$\mathfrak{S}_{\mathrm{c}(1)} \times \ldots \times \mathfrak{S}_{\mathrm{c}(n)}.$$

Proposition 3. There are $\prod_{k=1}^{n-1}(2k-1)$ plane-oriented recursive trees with n vertices.

Proof. If we have a plane-oriented recursive tree with $n-1$ vertices, there are $(2n-3)$ ways to build a plane-oriented recursive tree with n vertices from this one. For each vertex v of the tree, we have exactly $\mathrm{d}(v)$ ways to add the new vertex in the sequence of children of v, except for the root where we have $\mathrm{d}(1)-1$ choices. Altogether, we can build

$$\sum_{v\in V} \mathrm{d}(v) - 1 = 2|E| - 1 = 2(n-1) - 1 = 2n-3$$

new trees, which gives us the result by induction. □

1.1.3 Isomorphism between trees

If we have a labeled tree, there are a lot of quantities that do not depend on the labeling of it (height,size,...). Sometimes, we only focus on the structure of the tree, which corresponds to its isomorphism class.

Definition 10. Two trees T and T' are said to be *isomorphic* if there exists a bijection $\varphi : V(T) \to V(T')$ such that:

$$\forall \{v,w\} \in E(T), \{\varphi(v), \varphi(w)\} \in E(T').$$

The set containing all the trees that are isomorphic to T is called the *isomorphism class* of T, noted T°. An isomorpishm class of a labeled tree is called an *unlabeled tree*. If T is a rooted tree, we may want the root to be preserved by any isomorphism, and then define the height of an unlabeled tree. Often, we will denote the root of an unlabeled tree by $\emptyset$.

Proposition 4. There are $C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}$ unlabeled rooted trees with n vertices.

Proof. A Dyck word with $2n$ characters is a word composed with only two symbols, which are (and), in equal parts, and which is well-parenthesized. For example, (()()) is a Dyck word with 6 characters while ())(() is not. There are exactly C_n Dyck words with $2n$ characters. Indeed, if we take a word that is not a Dyck word, and if we swap every parenthesis starting from the first wrong), we obtain a word with $n+1$ closing parentheses and $n-1$ opening ones. By doing this, we get a bijection between words that are not Dyck words and words with $n+1$ closing parentheses, so we know the number of non Dyck words is equal to $\binom{2n}{n-1}$. We thus have $\binom{2n}{n} - \binom{2n}{n-1} = C_n$ Dyck words.

Now, let's see that we can build each and every unlabeled rooted tree with n vertices with a Dyck word with $2(n-1)$ characters. To do so, we start from a vertex which is the root, and for each character of the word, either we create a new vertex (when we encounter a (), either we go backwards (when we encounter a)). This is the bijection we need, which concludes the proof. □

Inside of a tree, we can find some similarity. One can quantify this similarity with the following functions.

Definition 11. Let T be a rooted tree, and $v \in V(T)$. Let $T_1, \dots, T_k$ be the subtrees of T rooted at the children of T. Let $S_1, \dots, S_\ell$ be the different classes of isomorphism of the previous trees. For j in $[\![1,\ell]\!]$, let ℓ_j be the number of subtrees T_i such that $T_i^\circ = S_j$. With this notation we can define:

$$\mathrm{Aut}(v,T) = \prod_{j=1}^{\ell} \ell_j!.$$

One must understand this term this way: the more $\mathrm{Aut}(T,v)$ is important, the more the descendants of the children of v look like each other. With this definition, we can determine the number of recursive trees with a fixed structure.

Proposition 5. Let T be an unlabeled rooted tree. Then the number of recursive trees S such that $S^\circ = T$ is equal to

$$\frac{|T|!}{\prod_{v\in V(T)\setminus\mathcal{L}(T)}(|T|_{v\downarrow}\operatorname{Aut}(v,T))}.$$

Proof. Let us prove this result by induction on the number n of vertices of the tree. The formula is true for $n = 2$. Let $T_1, \dots, T_{k=\mathrm{d}(\emptyset)}$ the subtrees rooted at the children of the root. In order to have an increasing labeling of the tree T, we have to choose a partition of $[\![2, n]\!]$ into k subsets of sizes $|T_1|, \dots, |T_k|$, and then, for each subtree, label it with the corresponding elements. The number of partitions is given by the following multinomial coefficient:

$$\binom{n-1}{|T_1|,\dots,|T_k|} = \frac{(n-1)!}{|T_1|!\cdots|T_k|!}.$$

Once we have chosen the partition, we must label the different subtrees, but as the order of the subtrees do not matter, we have to pay attention to not count separately the cases where we label two isomorphic trees with different subsets. To do so, we have to divide the product of all the number of labeling for each subtree, and divide it by $\operatorname{Aut}(v, T)$.

We thus obtained the following induction formula:

$$\begin{aligned}|\{\text{rec. trees isomorphic to } T\}| &= \frac{(n-1)!}{\operatorname{Aut}(\emptyset,T)\prod_{i=1}^k |T_i|!}\prod_{i=1}^k |\{\text{rec. trees isomorphic to } T_i\}| \\ &= \frac{|T|!}{(|T|\operatorname{Aut}(\emptyset,T))\prod_{i=1}^k |T_i|!}\prod_{i=1}^k |\{\text{rec. trees isomorphic to } T_i\}|.\end{aligned}$$

This concludes the proof by induction. □

We can do the same for plane-oriented recursive trees.

Proposition 6. Let T be an unlabeled rooted tree. Then the number of plane-oriented recursive trees S such that $S^\circ = T$ is equal to

$$\frac{|T|!\,\mathrm{d}_T(\emptyset)!\prod_{v\neq\emptyset}(\mathrm{d}(v)-1)!}{\prod_{v\in V(T)\setminus\mathcal{L}(T)}(|T|_{v\downarrow}\operatorname{Aut}(v,T))}.$$

Proof. We just have to choose an ordering for the children of each vertex. For the root we have $\mathrm{d}(\emptyset)!$ possibilities and for any other vertex v we have $(\mathrm{d}(v)-1)!$ possibilities. With the previous result we have what we want. □

We conclude this paragraph with a last definition that will be useful later.

Definition 12. Let T be an unlabeled unrooted tree, we define $\overline{\operatorname{Aut}}(u, T)$ as the number of vertices v such that (T, u) is isomorphic to (T, v).

1.2 Random trees

In this paragraph, we will see what we mean when we talk about random trees. There will be several ways to building random trees. The first one is by giving a law on a set of trees, like for example the uniform law over all recursive trees. The other one consists on having built the law for trees with $(n-1)$ vertices, and then giving the probability of attaching the new vertex to each of the $(n-1)$ others. A last way of building random trees we will encounter is by having a mapping from any set to a set of trees, knowing the law over the starting set and then have a law over the set of trees. All this schemes are obviously linked to each other, and using one or another can be useful depending on the context. In the following paragraph, we introduce a common way to build random trees by induction.

1.3 Uniform and preferential attachments

Here, the model we introduce attaches a new vertex to an existing one depending on the degree of it, to a certain power. It generates recursive trees (rooted in 1).

Definition 13. Let $n \in \mathbb{N}$ and $\alpha \in \mathbb{R}_+$. We define the law $\mathcal{T}_\alpha(n)$ as follows:

- $\mathcal{T}_\alpha(2) = (\{1,2\}, \{\{1,2\}\})$ the only recursive tree with two vertices.
- for $i \in [\![1, n-1]\!]$, the vertex n is attached to $i \in V(\mathcal{T}_\alpha(n-1))$ with a probability $\frac{\mathrm{d}(i)^\alpha}{\sum_{j=1}^{n-1} \mathrm{d}(j)^\alpha}$.

The more α grows, the more deterministic it becomes. When $\alpha = 0$, the choice of the vertex is completely uniform, and as α goes to infinity, we tend to choose the vertex with the highest degree.

Definition 14. For $\alpha = 0$, we call this model the *uniform attachment model*, noted $\mathrm{UA}(n)$ and for $\alpha = 1$, it is called *preferential attachment model*, noted $\mathrm{PA}(n)$.

Proposition 7. The uniform attachment model corresponds to the uniform law over the set of recursive trees, while the preferential attachment model corresponds to the uniform law over the set of plane-oriented recursive trees.

Proof. We obtain simply this result by looking at the way we determined the number of such trees previously. □

1.4 The union-find algorithm

The union-find algorithm is a process we use to build random trees, and we will use it to prove a result about the height of a tree obtained by the uniform attachment model. Here is how to build a tree with this algorithm:

- First, we begin with the n sets: $\{1\}, \ldots, \{n\}$, which are labeled from 1 to n.
- Then we chose two different sets uniformly among those n sets. Then we merge the two sets and the label of the new set is the label of the first one we picked.
- We keep doing this, at each step we chose uniformly two sets and we merge them.
- At the end, by reading the merged sets, we obtain a random labeled tree: if we merged the sets labeled i and j, we add the tree j as a subtree of i.

Remark 3. We can see the union-find algorithm as a mapping F from the set

$$\mathcal{C}_n = \{(a_1, a_2, \ldots, a_{2n-3}, a_{2n-2}) \in [\![1,n]\!]^{2n-2}, \forall p \in [\![1, n-2]\!], \forall k > 2p, a_k \notin \{a_2, a_4, \ldots, a_{2p}\}\}$$

to the set of labeled tress with n vertices. This mapping is not injective (for example $(1,3,1,2)$ and $(1,2,1,3)$ give the same tree with 3 vertices), but is clearly surjective. As an example, two sequences of integers possible to generate the tree of Figure 3 are $(4,3,2,4,2,5,2,1)$ and $(4,3,2,1,2,5,2,4)$.

We denote by $\mathrm{UF}(n)$ the law obtained by this process. L. Devroye showed [6] that such trees have a height that, re-scaled by a factor $\log(n)$, converges in probability and in expectancy to e. In order to have the same result for the uniform attachment model, we need the following theorem.

Theorem 8. Let $n \in \mathbb{N}^*$. Then we have the following equality of laws:

$$\mathrm{UF}(n)^\circ = \mathrm{UA}(n)^\circ$$

Once we have this result, we immediately have the following result, because the height of a tree only depends on its shape.

Corollary 9. The height of uniform attachment-random tree, re-scaled by a factor $\log(n)$, converges in probability and in expectancy:

$$\frac{\mathrm{h}(\mathrm{UA}(n))}{\log(n)} \xrightarrow[n\to\infty]{} \mathrm{e}.$$

Let us move on to the proof of the above theorem.

Proof. We already know that we have $(n-1)$ recursive trees with n vertices. Similarly, we want to know the number of different ways we have to generate trees by the union-find algorithm. In other words, we have to determine the cardinality of $\mathcal{C}_n$. To get an element of $\mathcal{C}_n$, we must choose a first element (n choices), then a second one ($n-1$ choices). The third one we choose can not be the second one we chose previously, so we have $(n-1)$ choices. For the fourth one, we can not choose neither the second one nor the third one, and so on. All in all:

$$|\mathcal{C}_n| = n(n-1) \times (n-1)(n-2) \times \cdots \times 2 = n(n-1)!^2.$$

Therefore, as the laws over $\mathcal{C}_n$ and the set of recursive trees with n vertices are both uniform, one has the following equivalences for any unlabeled rooted tree T:

$$\begin{aligned}\mathbb{P}(\mathrm{UF}(n)^\circ = T) = \mathbb{P}(\mathrm{UA}(n)^\circ = T) &\Leftrightarrow \frac{|\{c \in \mathcal{C}_n, F(c)^\circ = T\}|}{n(n-1)!^2} = \frac{|\{\text{rec. trees } S, S^\circ = T\}|}{(n-1)!} \\ &\Leftrightarrow \frac{|\{c \in \mathcal{C}_n, F(c)^\circ = T\}|}{|\{\text{rec. trees } S, S^\circ = T\}|} = n!.\end{aligned}$$

In order to show this last equality, we can show that for each recursive tree with n vertices which has a fixed structure, there are $n!$ elements of $\mathcal{C}_n$ (*i.e.* $n!$ ways to execute the union-find algorithm) that lead to a tree with the same structure. To do so, we use the bijection used before to count the number of recursive trees with n vertices: each recursive tree with n vertices can be seen as an element $(\omega_1, \ldots, \omega_{n-1}) \in [\![1]\!] \times \ldots \times [\![1, n-1]\!]$. For $T = (\omega_1, \ldots, \omega_{n-1})$ a recursive tree with n vertices, we introduce:

$$\varphi_T(\sigma) = (\sigma(\omega_{n-1}), \sigma(n), \sigma(\omega_{n-2}), \sigma(n-1), \ldots, \sigma(\omega_1 = 1), \sigma(2)),$$

where $\sigma \in \mathfrak{S}_n$. One can know verify that for two recursive trees T and S, $\varphi_T(\mathfrak{S}_n) = \{c \in \mathcal{C}_n, F(c)^\circ = T^\circ\}$ and that $\varphi_T(\mathfrak{S}_n) = \varphi_S(\mathfrak{S}_n)$ if and only if $T^\circ = S^\circ$. This concludes the proof because $|\mathfrak{S}_n| = n!$. □

2 Root-finding algorithms

2.1 General principle

Now, we have all the elements necessary to introduce root-finding algorithms. The absolute goal is the following: given an unlabeled and unrooted tree, and knowing the way it has been generated, we want a subset of the vertices of the tree (the smallest possible), called a *packet*, which contains the root of the tree with high probability. An important condition we want on the number of vertices given by the algorithm is that is has to be independent of the number of vertices in the tree, and depend only on the precision we want to reach at the end. Notice that we are not sure in general if such an algorithm even exists. In the two following paragraphs, we will formalize the notion of a root-finding algorithm: first with a fixed packet size and then with a random one.

2.2 Deterministic or random packet size

First, we give the definition of root-finding algorithms which output a set of K vertices such that K only depends on the probability bound fixed *a priori*.

Definition 15. Let $\alpha \in \{0,1\}$ and K a function of $\varepsilon \in]0,1[$. A family of mappings (H_ε) from unlabeled trees to subsets of $K(\varepsilon)$ elements of $\mathbb{N}$ is a *root-finding algorithm* (with deterministic packet size) if

$$\lim_{\varepsilon \to 0^+} \liminf_{n \to \infty} \mathbb{P}(1 \in H_\varepsilon(\mathcal{T}_\alpha(n)^\circ)) = 1.$$

If we think about it, it is not very important if the packet is sometimes a little bit too large, as long as we know it does not occur often. This leads to the following definition.

Definition 16. Let $\alpha \in \{0,1\}$. A family of mappings (H_ε) from unlabeled trees to subsets of $\mathbb{N}$ is a *root-finding algorithm* (with random packet size) if there exists a function K of ε only such that

$$\lim_{\varepsilon \to 0^+} \mathbb{P}\left(\sup_{n \geq 1} |H_\varepsilon(\mathcal{T}_\alpha(n)^\circ)| \leq K(\varepsilon) \text{ and } \forall n \geq 1, 1 \in H_\varepsilon(\mathcal{T}_\alpha(n)^\circ)\right) = 1.$$

In the next section, we focus on root-finding algorithm with a deterministic packet size for the uniform attachment model.

3 Upper and lower bounds for uniform attachment

The main purpose of this section is to give to bounds on the number of vertices given by a root-finding algorithm:

$$K^-(\varepsilon) \leq K(\varepsilon) \leq K^+(\varepsilon),$$

where

- $K^+(\varepsilon)$ is such that there exists a root-finding algorithm for which $K = K^+$,
- $K^-(\varepsilon)$ is such that any root-finding algorithm must have $K \geq K^-$.

The two bounds introduced in this section are from [4]. In order to compare the bounds, we will always try to have the lim inf term like $1 - O(\varepsilon)$.

3.1 A simple upper bound

The algorithm we will exhibit in this paragraph emerges from the following idea: the root of the tree must be a *central* vertex, that is, its subtrees must have a homogeneous size, and thus not be too big. Conversely, the leaves of the tree are completely not central, as they have a unique subtree starting from them, which has the size of the entire tree. To quantify these intuitions, we introduce the following function.

Definition 17. For a tree T, let $\psi : V(T) \to \mathbb{N}$ be defined by:

$$\psi_T(u) = \max_{v \neq u} |(T,u)_{v\downarrow}|.$$

It corresponds to the size of the largest subtree starting at a child of u.

As we said earlier, the root should have a small ψ value. This leads to the following theorem.

Theorem 10. Let $K(\varepsilon) \geq K_\psi(\varepsilon) = \frac{5}{2}\frac{\log(1/\varepsilon)}{\varepsilon}$. Then the mapping H_ψ, which consists in taking the $K(\varepsilon)$ vertices with the smallest ψ values, is a root-finding algorithm.

Proof. The goal of the proof is to give an upper bound on $\mathbb{P}(1 \notin H_\psi)$. If the root is not in the packet, it means that there are K values among $[\![2, n]\!]$ for which ψ is smaller than $\psi(1)$. In particular there must be one such value greater $i > K$. By interposing a $(1-\varepsilon)n$, one obtains the inequality:

$$\mathbb{P}(1 \notin H_\psi) \leq \mathbb{P}(\exists i > K, \psi(i) \leq \psi(1)) \leq \mathbb{P}(\psi(1) \geq (1-\varepsilon)n) + \mathbb{P}(\exists i > K, \psi(i) \leq (1-\varepsilon)n).$$

Now, to take the limit in n, we would like to have interesting convergences in law. Let $k \in \mathbb{N}^*$, for $i \leq k$, we denote by $T_{i,k}$ the tree containing vertex i in the forest obtained by removing in $T \sim \mathrm{UA}(n)^\circ$ all edges between $\{1, k\}$. The vector $(|T_{1,k}|, \ldots, |T_{k,k}|)$ follows a standard Pólya urn with k colors, and we can use the famous following result [7].

Fact 1. We consider a urn containing balls of k different colors. Initially, the urn contains α_1 balls of color 1, α_2 balls of color 2 and so on. We randomly choose randomly on ball, look at its color, replace it in the urn with another ball of the same color, and so on. After n draws, we denote be $X_1, \ldots, X_k$ the proportions of different colors. When n goes to infinity, the distribution of the vector $(X_1, \ldots, X_k)$ converges to the Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$. This distribution has the following density:

$$f_\alpha(x_1, \ldots, x_k) = \frac{1}{\mathrm{B}(\alpha)} \sum_{i=1}^{k} x_i^{\alpha_i - 1} \mathbf{1}_\Sigma,$$

Where B is the multinomial beta function and Σ the simplex $\{x_1, \ldots, x_k > 0, \sum_{i=1}^k x_i = 1\}$.

Then, we can have a limit bound for the term with $\psi(1)$, because:

$$\psi(1) \leq \max(|T_{1,2}|, |T_{2,2}|),$$

and we know the limit law of $|T_{1,2}|$ and $|T_{2,2}|$:

$$\frac{|T_{1,2}|}{n}, \frac{|T_{2,2}|}{n} \xrightarrow[n\to\infty]{\text{law}} \mathcal{U}(0,1).$$

All in all:

$$\limsup_{n\to\infty} \mathbb{P}(\psi(1) \geq (1-\varepsilon)n) \leq 2\,\mathbb{P}(\mathcal{U}(0,1) \geq (1-\varepsilon)) = 2\varepsilon.$$

For the other term, we have to notice that $\psi(i)$, for $i > K$, is the biggest subtree starting at a child of i. In particular, it is bigger than the subtree S starting at the child of i that leads to the root. And S contains all the trees $|T_{\cdot,K}|$, except the one which leads to i. In other words, the following inequality stands for all $i > K$:

$$\psi(i) \geq \min_{1 \leq k \leq K} \sum_{\substack{j=1 \\ j \neq k}}^{K} |T_{j,K}|.$$

To have a limit law, we only need to know that partial sums of Dirichlet distribution are beta laws:

$$\forall k \in [\![1, K]\!], \qquad \frac{1}{n} \sum_{\substack{j=1 \\ j \neq k}}^{K} |T_{j,K}| \xrightarrow[n\to\infty]{\text{law}} \beta(K-1, 1),$$

so that we have:

$$\limsup_{n\to\infty} \mathbb{P}(\exists i > K, \psi(i) \leq (1-\varepsilon)n) \leq \mathbb{P}(\beta(K-1,1) \leq 1-\varepsilon) = K(1-\varepsilon)^{K-1} \leq 2\varepsilon,$$

if we take $K > K_\psi(\varepsilon)$. In short,

$$\liminf_{n\to\infty} \mathbb{P}(1 \in H_\psi(\mathrm{UA}(n)^\circ)) \geq 1 - 4\varepsilon,$$

which concludes the proof. □

To analyze this result, we can use the following result.

Proposition 11. For any integer K:

$$\mathbb{P}(1 \notin H_\psi(\mathrm{UA}(n)^\circ)) \geq \frac{C}{K},$$

where $C > 0$ is a constant that is independent of n.

Proof. If 1 is a leaf in $\mathrm{UA}(K)$, then it will maximize the value of ψ, so it will surely not be in the packet:

$$\mathbb{P}(1 \notin H_\psi) \geq \mathbb{P}(1 \in \mathcal{L}(\mathrm{UA}(K))) = \prod_{i=2}^{K-1} \left(1 - \frac{1}{i}\right) \geq \frac{1}{K-1},$$

where the last inequality can easily be shown by induction. □

Combining the last proposition and the theorem, we obtain two constants C_1, C_2 such that:

$$\frac{C_1 \varepsilon}{\log(1/\varepsilon)} \leq \mathbb{P}(1 \notin H_\psi) \leq C_2 \varepsilon.$$

The theorem is optimal up to a logarithmic factor, we will try to do better later on. Before that, we will investigate the other bound for K.

3.2 Maximum likelihood estimator

The goal of this paragraph is to have an impossibility result, *i.e.* a minimum number of vertices we have to output in order to be sure that our algorithm will be satisfying. The general theory of the maximum likelihood estimator and its optimality give an optimal set of K vertices. In the main theorem of this paragraph, we give a lower bound on the number of vertices output by the maximum likelihood estimator procedure. Before doing this, let us determine this estimator.

Proposition 12. Given an unlabeled rooted tree T, the maximum likelihood estimator for the root in the uniform attachment model is the vertex minimizing the function:

$$\zeta_T(u) = \overline{\mathrm{Aut}}(u, T) \prod_{v \in V(T) \setminus \mathcal{L}((T,u))} (|(T,u)_{v\downarrow}| \mathrm{Aut}(v, (T,u))).$$

Proof. Among all the recursive trees which have a given shape T, and for any vertex u, the $\overline{\mathrm{Aut}}(u, T)$ issues that begin with all the roots isomorphic to u have the same probability (we recall that $\mathrm{UA}(n)$ corresponds to the uniform law over the set of all recursive trees with n vertices). Therefore, if $t \sim \mathrm{UA}(n)$:

$$\mathbb{P}(u = 1 \cap t^\circ = T) = \frac{1}{\overline{\mathrm{Aut}}(u, T)},$$

which implies that:

$$\mathbb{P}(u = 1 | t^\circ = T) = \frac{|\{\text{rec. trees isomorphic to } T\}|}{(n-1)!\, \overline{\mathrm{Aut}}(u, T)}.$$

With a previous proposition, we can conclude that the desired likelihood is equal to:

$$L(u; T) = \frac{|T|!}{(n-1)!\, \overline{\mathrm{Aut}}(u, T) \prod_{v \in V(T) \setminus \mathcal{L}((T,u))} (|(T,u)_{v\downarrow}| \mathrm{Aut}(v, (T,u)))}.$$

So if we want to maximize it, it is equivalent to minimizing its inverse, and after having got rid of the constant terms, we get to the desired result. □

Theorem 13. There exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$, any root-finding algorithm which satisfies:

$$\liminf_{n\to\infty} \mathbb{P}(1 \notin H(\mathrm{UA}(n)^\circ)) \geq 1 - \varepsilon$$

must have $K(\varepsilon) \geq \exp\left(\sqrt{c\log(\frac{1}{2\varepsilon})}\right)$, where c is a positive constant.

Proof. Let $K = \exp\left(\sqrt{c\log(\frac{1}{2\varepsilon})}\right)$, with a constant $c \geq \frac{1}{30}$ such that K is an integer. As any procedure for a tree with $(n+1)$ vertices can be implemented on a tree with n vertices, the optimality of the maximum likelihood estimator procedure implies that $\mathbb{P}(1 \notin H_\zeta(\mathrm{UA}(n)^\circ))$ is increasing with n (H_ζ outputs the set with K smallest ζ values). If we succeed to show that $\mathbb{P}(1 \notin H_\zeta(\mathrm{UA}(K+1)^\circ)) > \varepsilon$, we will be done by contrapositive. To do so, we will exhibit a subset of recursive trees with $K+1$ vertices $\mathcal{T}$ for which $\mathbb{P}(\mathrm{UA}(K+1) \in \mathcal{T}) > \varepsilon$ and such that:

$$\forall t \in \mathcal{T}, \zeta_T(1) > \zeta_T(i), i \in \{2, \dots, K+1\}.$$

We thus denote by $\mathcal{T}$ all the recursive trees with $K+1$ vertices which satisfy the following three properties:

- (i) The first $10\log(K)^2$ vertices form a straight line $1 - 2 - \cdots - 10\log(K)$.
- (ii) All other vertcies are descendants of $10\log(K)$
- (iii) The height of the subtree rooted in $10\log(K)$ is smaller than $4\log(K)$.

First, let us find a lower bound for the probability of this set.

- $\mathbb{P}((i)) = \frac{1}{(10\log(K)-1)!} \geq \exp(-10\log(K)^2)$ by using the inequality $\Gamma(x) \leq \mathrm{e}^{(x-1)^2} \quad \forall x > 1$.
- $\mathbb{P}((iii)|(ii)) = \mathbb{P}(\mathrm{h}(\mathrm{UA}(K-10\log(K)+1)) \leq 4\log(K))$. But with our previous convergence result on the height of random recursive trees, we get that $\mathrm{h}(\mathrm{UA}(K-10\log(K)+1)) \simeq \mathrm{e}\log(K-10\log(K)+1) \leq 4\log(K)$. So for K large enough, $\mathbb{P}((iii)|(ii)) \leq \frac{1}{2}$.
- $\mathbb{P}((ii)) = \prod_{i=10\log(K)}^{K}(1 - \frac{10\log(K)-1}{i}) = \frac{1}{\binom{K}{10\log(K)-1}} \geq K^{-10(\log(K)-1)} \geq \exp(-20\log(K)^2)$, where we have used the following inequality: $\forall a, b \geq 1, \binom{b}{a} \leq b^a$.

All in all,

$$\mathbb{P}(\mathcal{T}) = \mathbb{P}((i))\,\mathbb{P}((ii))\,\mathbb{P}((iii)|(ii)) \geq \frac{1}{2}\exp(-30\log(K)^2) \geq \varepsilon.$$

We have $1 \notin H_\zeta(T)\,\forall T \in \mathcal{T}$ left to show to conclude the proof. As trees of $\mathcal{T}$ have $K+1$ vertices, we only have to show that 1 has the biggest ζ value. The maximum likelihood estimator has a quite complicated expression, but the comparison of ζ values can be simplified because we only need to compare the factors corresponding to vertices belonging to the path between the two vertices we want to compare. Mathematically speaking, for any tree T and any vertices u and v, with $(u = u_1, \dots, u_k = v)$ being the (unique) path between u and v:

$$\begin{aligned}\zeta_T(u) > \zeta_T(v) \Leftrightarrow \overline{\mathrm{Aut}}(u,T)\prod_{i=1}^{k}(|(T,u)_{u_i\downarrow}|\mathrm{Aut}(u_i,(T,u))) \\ > \overline{\mathrm{Aut}}(v,T)\prod_{i=1}^{k}(|(T,v)_{u_i\downarrow}|\mathrm{Aut}(u_i,(T,v))).\end{aligned}$$

Indeed, if $w \notin \{u_1, \dots, u_k\}$, one can easily see that $(T,u)_{w\downarrow} = (T,v)_{w\downarrow}$, so that we also have $\mathrm{Aut}(w,(T,u)) = \mathrm{Aut}(w,(T,v))$. Now, we would like to get rid of the Aut terms. To do so, we will use the following fact, which links the Aut terms with (T,v) with those with (T,u).

[2]To avoid boring notations, we will make as if a lot of quantities are actually integers.

Fact 2. For all $i \in [\![1, k]\!]$,

$$\mathrm{Aut}(u_i, (T, v)) \leq |(T, v)_{u_i\downarrow}| \, \mathrm{Aut}(u_i, (T, u)).$$

To show this fact, one has to remind that the Aut term (for v) is based on a list of k_v subtrees, and only one subtree is modified or added when we compute $\mathrm{Aut}(u_i, (T, u))$, which results on a multiplicative change of at most $k_v + 1$, which can be simply dominated by $|(T, v)_{u_i\downarrow}|$.

Combining this fact with the previous equivalence and the bounds $1 \leq \overline{\mathrm{Aut}}(u, T) \leq |T|$, we obtain the following sufficient condition for having a greater ζ_T value:

$$\prod_{i=1}^{k} |(T, u)_{u_i\downarrow}| > |T| \prod_{i=1}^{k} |(T, v)_{u_i\downarrow}|^2 \Rightarrow \zeta_T(u) > \zeta_T(v).$$

We will now use this condition to show that 1 has the biggest ζ value in trees of $\mathcal{T}$. Let $T \in \mathcal{T}$ and $v \in [\![10\log(K) + 1, K + 1]\!]$, with $(1 = u_1, \ldots, u_k = v)$ the unique path between the root and v. Let $S = (T, 1)_{10\log(K)\downarrow}$. One the one hand, we have:

$$\prod_{i=1}^{k} |(T, u)_{u_i\downarrow}| \geq |S|^{10\log(K)} = (K + 1 - 10\log(K))^{10\log(K)}.$$

And on the other hand:

$$\begin{aligned}
\prod_{i=1}^{k} |(T, v)_{u_i\downarrow}| &= \prod_{\substack{i=1 \\ u_i \in S}}^{k} |(T, v)_{u_i\downarrow}| \prod_{\substack{i=1 \\ u_i \notin S}}^{k} |(T, v)_{u_i\downarrow}| \\
&\leq (K + 1)^{\mathrm{h}(S)} \, (10\log(K))! \\
&\leq (K + 1)^{4\log(K)} \, (10\log(K))^{10\log(K)}.
\end{aligned}$$

Where the last inequality comes from the definition of trees of $\mathcal{T}$ and the inequality $a! \leq a^a$. We can finally conclude:

$$\begin{aligned}
\prod_{i=1}^{k} |(T, 1)_{u_i\downarrow}| &\geq (K + 1 - 10\log(K))^{10\log(K)} \\
&\geq (K + 1)^{8\log(K)+1} \, (10\log(K))^{20\log(K)} \qquad \text{for } K \text{ large enough} \\
&\geq |T| \prod_{i=1}^{k} |(T, v)_{u_i\downarrow}|^2 \\
&\Rightarrow \zeta_T(1) > \zeta_T(v).
\end{aligned}$$

□

3.3 A tighter upper bound

In this paragraph, we aim to have a tighter upper bound for K. A first idea could be to assess the performance of the maximum likelihood estimator, but its expression would make it too difficult (due to the presence of the automorphism numbers). A simple solution for this is to slightly modify it to make it simpler, but not too simple to still have good estimations. We denote by φ this modification, defined by:

$$\varphi_T(u) = \prod_{v \neq u} |(T, u)_{\downarrow}|.$$

It has a simpler expression than the maximum likelihood estimator, but catches more information than ψ did in the first place. As before, we denote by H_φ the associated mapping. Like for ψ,

we see that if u is a leaf, $\varphi(u)$ will be high. However, it will not necessarily be maximal, and that is why it H_φ will be a better algorithm. Even if 1 is a leaf, $\varphi(1)$ can have small values, as long as 2 has a significant number of chlidren (which is likely to happen conditionned to 1 being a leaf). Before stating the main theorem, we will show a little result about φ that will be useful later.

Proposition 14. Let u and v be to vertices of an unlabeled tree T, and let γ be the path from u to v excluding u and v. Then for all $w \in \gamma, \varphi(w) \leq \max(\varphi(u), \varphi(v))$.

Proof. Let T_u be the subtree of (T, w) rooted at a child of w containing u, and let T_v be defined similarly. We suppose, without loss of generality, that $|T_u| \geq |T_v|$. Notice that because $w \in \gamma$, T_u and T_v have no vertex in common. We will show that $\varphi(w) \leq \varphi(v)$, which automatically leads to the desired result. We distinguish three cases:

- if $x \notin \gamma$, then we have $(T, w)_{x\downarrow} = (T, u)_{x\downarrow} = (T, v)_{x\downarrow}$ so $|(T, w)_{x\downarrow}| = |(T, v)_{x\downarrow}|$,
- if $x \in \gamma, x \neq w$, x is either between u and w, where we still have $(T, w)_{x\downarrow} = (T, v)_{x\downarrow}$, or between w and v, where we have $(T, w)_{x\downarrow} \subset T_v$ and $T_u \subset (T, v)_{x\downarrow}$ so that $|(T, w)_{x\downarrow}| \leq (T, v)_{x\downarrow}$ because $|T_u| \leq |T_v|$,
- the last factors we have to compare are the ones that are only in one of the two products: $(T, w)_{v\downarrow} \subset T_v$ and $T_u \subset (T, v)_{w\downarrow}$ so $|(T, w)_{v\downarrow}| \leq |(T, v)_{w\downarrow}|$.

Therefore, by doing the product of all this (in)equalities, we obtain the result. □

Theorem 15. There exists $a, b > 0$ two universal constants such that if $K \geq a \exp\left(b \frac{\log(1/\varepsilon)}{\log(\log(1/\varepsilon))}\right)$, then

$$\liminf_{n\to\infty} \mathbb{P}(1 \in H_\varphi(\mathrm{UA}(n)^\circ)) \geq 1 - \varepsilon.$$

Proof. Here, the vertices we will focus on are the one which are at the bottom right of the tree (if we order births from left to right). To quantify it, we will change the labeling of the vertices. The root is labeled with $\emptyset$, and the node $(j_1, \ldots, j_\ell) \in \mathbb{N}^\ell$ is the ℓ-th child (in birth order) of $(j_1, \ldots, j_{\ell-1})$. We define $\ell(v)$ as the depth of v, that is, the length of the tuple that represents it.

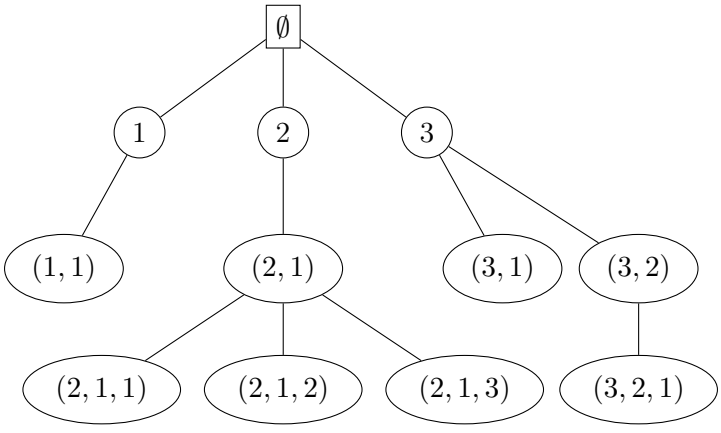

Figure 4: Here is what the new labeling looks like. The node with the greatest s value is $(3, 2, 1)$. Notice that this new labeling only depends on the geometry of the tree (if we order the children), we do not know for example if $(1, 1)$ appeared after or before $(3, 2, 1)$.

Now we can define:

$$s(v) = \sum_{k=1}^{\ell(v)} (\ell(v) + 1 - k) j_k,$$

which is what quantifies the fact to be on the right of the tree. We denote by $\mathcal{S}$ the set of vertices of $T \sim \mathrm{UA}(n)$ which satisfy $s(v) > 3S$, where S will be chosen later. We want to calculate the limit probability that elements of $\mathcal{S}$ have greater φ values than 1. The proof will then be separated in four steps. In the first place, we characterize the elements of $\mathcal{S}$ to give a first bound on the probability that an element of $\mathcal{S}$ has a greater φ value than the root. For any $v \in \mathcal{S}$, one has either:

- there exists u such that $s(u) \in]S, 3S]$ and $v \in (T, \emptyset)_{u\downarrow}$, or,
- there exists u such that $s(u) \leq S$ and $v \in (T, \emptyset)_{(u,j)\downarrow}$ for a $j > s$.

We show that by induction on the depth of v. For $\ell(v) = 1$, the second assertion is true with $u = \emptyset$. For $\ell(v) > 1$, we simply choose u the parent of v, and then we have three cases:

- If $u \in \mathcal{S}$, we can apply the induction hypothesis on u.
- If $s(u) \in]S, 3S]$, then the first assertion is true.
- If $s(u) \leq S$, let $v = (u, j)$ and then

$$\begin{aligned} s((u,j)) &= \sum_{k=1}^{\ell(u)+1} (\ell(u) + 2 - k) j_k \\ &= \sum_{k=1}^{\ell(u)} (\ell(u) + 2 - k) j_k + j \\ &= \sum_{k=1}^{\ell(u)} (\ell(u) + 1 - k) j_k + \sum_{k=1}^{\ell(u)} j_k + j \\ &= s(u) + \sum_{k=1}^{\ell(u)} j_k + j \leq j + 2s(u), \end{aligned}$$

  so that j is necessarily strictly greater than S, and so the second assertion is true, which means we proved the property.

In terms of probability, the previous property corresponds to:

$$\begin{aligned} \mathbb{P}(\exists v \in \mathcal{S} \text{ and } \varphi(v) \leq \varphi(1)) \leq\, & \mathbb{P}(\exists v, s(v) \in]S, 3S] \text{ and } \varphi(v) \leq \varphi(1)) \\ & + \mathbb{P}(\exists v, j, s(v) \leq S, j > S \text{ and } \varphi(v) \leq \varphi(1)). \end{aligned}$$

However, if $\varphi(v) \leq \varphi(1)$, we have also $\varphi((v,j)) \leq \varphi(1) = \max(\varphi(v), \varphi(1))$. Finally, we have the following bound:

$$\begin{aligned} \mathbb{P}(\exists v \in \mathcal{S} \text{ and } \varphi(v) \leq \varphi(1)) \leq\, & \mathbb{P}(\exists v, s(v) \in]S, 3S] \text{ and } \varphi(v) \leq \varphi(1)) \\ & + \mathbb{P}(\exists v, j, s(v) \leq S, j > S \text{ and } \varphi((v,j)) \leq \varphi(1)). \end{aligned}$$

Now the second step of the proof consists in simplifying this bound. To do so, we focus on the comparison of two φ values, as we did previously for the maximum likelihood estimator. As before, we only have to compare the values on the path between the two vertices.

$$\varphi(1) \geq \varphi(v) \Leftrightarrow \prod_{i=1}^{\ell(v)} |(T, \emptyset)_{(j_1,\dots,j_i)\downarrow}| \geq \prod_{i=0}^{\ell(v)-1} |(T, v)_{(j_1,\dots,j_i)\downarrow}|.$$

One can remark that each single tree in the first product corresponds to the complementary of a unique tree in the other product, so we have:

$$\varphi(1) \geq \varphi(v) \Leftrightarrow \prod_{i=1}^{\ell(v)} |(T, \emptyset)_{(j_1,\dots,j_i)\downarrow}| \geq \prod_{i=1}^{\ell(v)} (n - |(T, \emptyset)_{(j_1,\dots,j_i)\downarrow}|).$$

Let us denote by $A(v)$ the first product and by $B(v)$ the second one. Then for $j > S$, one has:

$$A((v,j)) = A(v)|(T,\emptyset)_{(v,j)\downarrow}| \leq \sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|$$

$$B((v,j)) = B(v)(n - |(T,\emptyset)_{(v,j)\downarrow}|) \geq \left(n - \sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|\right).$$

In particular,

$$\begin{aligned}\exists j > S, \varphi((v,j)) \leq \varphi(1) \Rightarrow & \prod_{i=1}^{\ell(v)} |(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}| \left(\sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|\right) \\ & \geq \prod_{i=1}^{\ell(v)} (n - |(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}|) \left(n - \sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|\right).\end{aligned}$$

The sum $\sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|$ represents the amount of vertices there are below v after the $(S+1)$-th child. To get more nodes there, we must have had the node $(v, S+1)$. Then, having a certain amount of vertices there is less likely than having the same amount of vertices below (v,S). Then, we have the following stochastic dominance:

$$\sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}| \preccurlyeq |(T,\emptyset)_{(v,S)\downarrow}|.$$

All in all,

$$\begin{aligned}&\mathbb{P}(\exists j > S, \varphi((v,j)) \leq \varphi(1)) \\ &\leq \mathbb{P}\left(\prod_{i=1}^{\ell(v)} |(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}| \left(\sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|\right) \geq \prod_{i=1}^{\ell(v)} (n - |(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}|) \left(n - \sum_{j=S+1}^{\infty} |(T,\emptyset)_{(v,j)\downarrow}|\right)\right) \\ &\leq \mathbb{P}\left(\prod_{i=1}^{\ell(v)} |(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}||(T,\emptyset)_{(v,S)\downarrow}| \geq \prod_{i=1}^{\ell(v)} (n - |(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}|) \left(n - |(T,\emptyset)_{(v,S)\downarrow}|\right)\right) \\ &= \mathbb{P}(A(v)|(T,\emptyset)_{(v,S)\downarrow}| \geq B(v)(n - |(T,\emptyset)_{(v,S)\downarrow}|)) \\ &= \mathbb{P}(A((v,S)) \geq B((v,S))) \\ &= \mathbb{P}(\varphi((v,S)) \leq \varphi(1)).\end{aligned}$$

We can bound the second term of the bound of the first step:

$$\begin{aligned}\mathbb{P}(\exists v, j, s(v) \leq S, j > S \text{ and } \varphi((v,j)) \leq \varphi(1)) &\leq \sum_{s(v)\leq S} \mathbb{P}(\exists j > S, \varphi((v,j)) \leq \varphi(1)) \\ &\leq \sum_{s(v)\leq S} \mathbb{P}(\varphi((v,S)) \leq \varphi(1)).\end{aligned}$$

We can reindex the above sum if we recall that

$$s((v,S)) \leq S + 2s(v) \text{ so that } s(v) \leq S \Rightarrow s((v,S)) \in]S, 3S],$$

to get:

$$\mathbb{P}(\exists v, j, s(v) \leq S, j > S \text{ and } \varphi((v,j)) \leq \varphi(1)) \leq \sum_{s(v)\in]S,3S]} \mathbb{P}(\varphi(v) \leq \varphi(1)).$$

Finally, after another union bound, we obtain:

$$\mathbb{P}(\exists v \in \mathcal{S} \text{ and } \varphi(v) \leq \varphi(1)) \leq \sum_{S<s(v)\leq 3S} \mathbb{P}(\varphi(v) \leq \varphi(1)).$$

The next step of the proof is to control the limit value of $\mathbb{P}(\varphi(v) \leq \varphi(1))$. To do so, we have to use a general limit theorem to have something that does not depend on the problem anymore like we did for ψ previously. Let us get interested in the vector:

$$V_n(v) = \left(\frac{1}{n}|(T,\emptyset)_{(j_1,\dots,j_i)\downarrow}|\right)_{i=1,\dots,\ell(v)}.$$

The sequence in n of each first component of this vector corresponds to the following experience: at the beginning, we have an urn with one white ball. At each step, we randomly choose a ball in the urn, if it is not white we add one more ball of the same color, and if it is white, we add a new color. The limit of this first component corresponds to the limit proportion of the j_1-th color when we pursue a large number of draws. According to the general theory of Pólya urns, this limit proportion follows the law of the product of j_1 independent uniform $\mathcal{U}(0,1)$ laws. Then, we can go down a notch in the tree, and notice that among the nodes under j_1, the nodes that are below (j_1, j_2) correspond to the same experience. In short, we know the limit law of $V_n(v)$:

$$V_n(v) \xrightarrow[\text{law}]{n\to\infty} (U_{j_1,1}, U_{j_1,1} \times U_{j_2,2}, \dots, U_{j_1,1} \times \dots \times U_{j_\ell,\ell}),$$

where $(U_{j,m})_m$ are independent copies of products of j uniform $\mathcal{U}(0,1)$ laws. Once we have the limit law, we just have to rewrite things:

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(\varphi(v) \leq \varphi(1)) &= \mathbb{P}\left(\prod_{i=1}^{\ell(v)} \prod_{k=1}^{i} U_{j_k,k} \geq \prod_{i=1}^{\ell(v)} \left(1 - \prod_{k=1}^{i} U_{j_k,k}\right)\right) \\ &= \mathbb{P}\left(\prod_{i=1}^{\ell(v)} U_{j_i,i}^{\ell(v)+1-i} \geq \prod_{i=1}^{\ell(v)} \left(1 - \prod_{k=1}^{i} U_{j_k,k}\right)\right).\end{aligned}$$

The thing we have left to calculate is just a probability that does not depend of the problem strictly speaking, and it can be dominated thanks to probability lemmas that can be found in [4] (we will not dwell on it because it would be too long and not really interesting). Finally, we can show that for any vertex with $s(v) \geq 10^{10}$:

$$\limsup_{n\to\infty} \mathbb{P}(\varphi(v) \leq \varphi(1)) \leq 7\exp(-0.21\sqrt{s(v)}\log(s(v))).$$

The last step of the proof consists in bringing all the pieces together and check that $\mathcal{S}$ contains enough vertices. Let $S \geq 10^{10}$, with the results of the two previous steps, we get that:

$$\limsup_{n\to\infty} \mathbb{P}(\varphi(v) \leq \varphi(1)) \leq 14\,|\{v, s(v) \leq 3S\}|\mathrm{e}^{-0.21\sqrt{S}\log(S)} = 14\,(n - |\mathcal{S}|)\mathrm{e}^{-0.21\sqrt{S}\log(S)}.$$

Now we need Erdös' non-asymptotic version of the Hardy-Ramanujan formula on the number of partitions of an integer [8]:

$$\left|\left\{(j_1,\dots,j_\ell) \in \mathbb{N}^\ell, \sum_{k=1}^{\ell} kj_k = s\right\}\right| \leq \exp\left(\pi\sqrt{\frac{2s}{3}}\right),$$

to show that $n - |\mathcal{S}| \leq 3S\exp(\pi\sqrt{2S})$. All in all, one obtains:

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(\varphi(v) \leq \varphi(1)) &\leq 42S\exp(-0.21\sqrt{S}\log(S) + \pi\sqrt{2S}) \\ &\leq \exp\left(-\frac{1}{100}\sqrt{S}\log(S)\right).\end{aligned}$$

We now need to check that with $\exp\left(-\frac{1}{100}\sqrt{S}\log(S)\right) \leq \varepsilon$, we can have $K(\varepsilon) \geq 3S\exp(\pi\sqrt{2S})$ with the bound written in the statement of the theorem. We thus assume that:

$$K(\varepsilon) \geq a\exp\left(b\frac{\log(1/\varepsilon)}{\log(\log(1/\varepsilon))}\right),$$

so that we have:

$$K(\varepsilon) \geq a\exp\left(b\frac{\sqrt{S}\log(S)}{100\log(\frac{\sqrt{S}\log(S)}{100})}\right) \geq a\exp(\sqrt[4]{S}\log(S))$$

for a certain $b > 0$. Finally,

$$K(\varepsilon) \geq aS^{\sqrt[4]{S}} \geq 3S\exp(\pi\sqrt{2S})$$

for a certain $a > 0$, which concludes the proof. □

Now, we can compare the two upper bounds we obtained. Let $x = 1/\varepsilon$, to compare the bounds, we get rid of the constants. The first bound we obtained behaves as:

$$K_1(x) = x^{1+\frac{\log(\log(x))}{\log(x)}},$$

and the second one as

$$K_2(x) = x^{\frac{1}{\log(\log(x))}}.$$

In the first one, the exponent goes to 1 as x goes to infinity, while in the second one it goes to 0. We can say that the second one is exponentially better than the first one. Also, we can look if our second bound is precise by checking how far it is from the lower bound. The lower bound behaves as:

$$K_3(x) = x^{\frac{1}{\sqrt{\log(x)}}}.$$

The exponent goes to 0 faster than it does for K_2. Our two bounds are not tighten. We will see in the next section that we can actually tighten the two bounds for the preferential attachment model.

Finally, let us say a word on the computation of these algorithms. The maximum likelihood estimator can be computed in polynomial time (polynomial in the number of vertices), and so are the two relaxations of it we studied, which is satisfying. The two relaxations offer a slight gain in terms of complexity, but we mainly studied them to obtain the bound in a simpler way.

4 Upper and lower bound for preferential attachment

4.1 A simple upper bound

In this paragraph, we investigate the use of the ψ algorithm on the preferential attachment model.

Theorem 16. Let $K(\varepsilon) \geq C\frac{\log(1/\varepsilon)^2}{\varepsilon^4}$ for a certain $C > 0$. Then we have:

$$\liminf_{n\to\infty} \mathbb{P}(1 \in H_\psi(\mathrm{PA}(n)^\circ)) \geq 1 - \varepsilon.$$

In particular, we have a root-finding algorithm.

Proof. We consider, as above the random vector $(|T_{1,k}|, \ldots, |T_{k,k}|)$. We do not have a simple Pólya urn as before. Indeed, if we recall what is a plane-oriented recursive tree, each time we add a vertex, we have two possibilities more to add a new one (to the left or to the right of the new vertex). This corresponds to a Pólya urn with replacement factors.

Fact 3. We consider a urn containing balls of k different colors. Initially, the urn contains α_1 balls of color 1, α_2 balls of color 2 and so on. We randomly choose randomly on ball, look at its color, replace it in the urn with r_j other balls of the same color if it is color j, and so on. After n draws, we denote be $X_1, \dots, X_k$ the proportions of different colors. When n goes to infinity, the distribution of the vector $(X_1, \dots, X_k)$ converges to the Dirichlet distribution $\mathrm{Dir}(\frac{\alpha_1}{r_1}, \dots, \frac{\alpha_k}{r_k})$.

Then, with this fact and what we said above, the vector $2(|T_{1,k}|, \dots, |T_{k,k}|)$ is a Pólya with replacement factors $(2, \dots, 2)$. Therefore, conditioned on $\mathrm{PA}(k)$:

$$\frac{1}{n}(|T_{1,k}|, \dots, |T_{k,k}|) \xrightarrow[n\to\infty]{\text{law}} \mathrm{Dir}\left(\frac{\mathrm{d}_k(1)}{2}, \dots, \frac{\mathrm{d}_k(k)}{2}\right),$$

with d_k the degree in $\mathrm{PA}(k)$. What we have to do now is to adapt the proof for the uniform attachment model with the new limit laws. We start once again by this inequality:

$$\mathbb{P}(1 \notin H_\psi) \leq \mathbb{P}(\exists i > K, \psi(i) \leq \psi(1)) \leq \mathbb{P}(\psi(1) \geq (1-\eta)n) + \mathbb{P}(\exists i > K, \psi(i) \leq (1-\eta)n).$$

For the first term, we use:

$$\psi(1) \leq \max(|T_{1,2}|, |T_{2,2}|) \text{ with } \frac{|T_{1,2}|}{n}, \frac{|T_{2,2}|}{n} \xrightarrow[n\to\infty]{\text{law}} \beta\left(\frac{1}{2}, \frac{1}{2}\right),$$

in order to get:

$$\limsup_{n\to\infty} \mathbb{P}(\psi(1) \geq (1-\eta)n \leq 2 \lim_{n\to\infty} \mathbb{P}(|T_{1,2}| \geq (1-\eta)n) = \frac{2}{\pi}\arcsin(\sqrt{\eta}) \leq \sqrt{\eta}.$$

For the other term, we still have:

$$\forall i > K, \psi(i) \geq \min_{1\leq k\leq K} \sum_{\substack{j=1 \\ j\neq k}}^{K} |T_{j,K}|.$$

However, the sums in the minimum do not have all the same law anymore, but fortunately we can stochastically lower-bound them all by one of them:

$$\frac{1}{n}\sum_{j=2}^{K} |T_{j,K}| \preccurlyeq \frac{1}{n}\sum_{\substack{j=1 \\ j\neq k}}^{K} |T_{j,K}|.$$

And this lower bound converges in distribution to $\beta(K - 1 - \frac{\mathrm{d}_K(1)}{2}, \frac{\mathrm{d}_K(1)}{2})$. Then we get:

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(\exists i > K, \psi(i) \leq (1-\eta)n) &\leq \lim_{n\to\infty} \mathbb{P}\left(\exists k \in [\![1, K]\!], \sum_{\substack{j=1 \\ j\neq k}}^{K} |T_{j,K}| \leq (1-\eta)n\right) \\ &\leq K\,\mathbb{P}\left(\beta\left(K - 1 - \frac{\mathrm{d}_K(1)}{2}, \frac{\mathrm{d}_K(1)}{2}\right) \leq 1 - \eta\right).\end{aligned}$$

It has been shown in [9] that this probability could be upper-bounded by η for $K \geq C\frac{\log(1/\eta)}{\eta^2}$. The proof is ended by taking $\eta = \varepsilon^2$. □

Like we did for the uniform attachment model, we can estimate the performance of this algorithm.

Proposition 17. For any integer K:

$$\mathbb{P}(1 \notin H_\psi(\mathrm{UA}(n)^\circ)) \geq \frac{C}{\sqrt{K}},$$

where $C > 0$ is a constant that is independent of n.

Proof. We do as we did previously, if 1 is a leaf in $\mathrm{PA}(K)$, then it will maximize the value of ψ, so it will surely not be in the packet:

$$\mathbb{P}(1 \notin H_\psi) \geq \mathbb{P}(1 \in \mathcal{L}(\mathrm{UA}(K))) = \prod_{i=2}^{K-1} \left(1 - \frac{1}{2(i-1)}\right) \geq \frac{1}{\sqrt{K}},$$

where the last inequality can easily be shown by induction. □

And then we obtain the following bounds:

$$\frac{C_1 \varepsilon^2}{\log(1/\varepsilon)} \leq \mathbb{P}(1 \notin H_\psi) \leq C_2 \varepsilon.$$

And if we wanted to have bounding as before, we would have to choose $K \geq \frac{C}{\varepsilon^2}$, which means we have a quadratic gap between this lower bound and the upper bound given by the above theorem. It is less precise than it was for the uniform attachment model. Indeed, for preferential attachment trees, we expect to have more compact trees, so ψ values are less spread out. In the last paragraph, we study an algorithm which takes this into account to be more precise. But before, let us focus on an impossibility result for this model.

4.2 Maximum likelihood estimator

One idea to have an impossibility result for the preferential attachment model would be to study the maximum likelihood estimator.

Proposition 18. Given an unlabeled rooted tree T, the maximum likelihood estimator for the root in the preferential attachment model is the vertex minimizing the function:

$$\xi_T(u) = \frac{\overline{\mathrm{Aut}}(u,T)}{\mathrm{d}(u)} \prod_{v \in V(T) \setminus \mathcal{L}((T,u))} (|(T,u)_{v\downarrow}| \mathrm{Aut}(v,(T,u))) = \frac{\zeta_T(u)}{\mathrm{d}(u)}.$$

Proof. As our trees are distributed uniformly over the set of all plane-oriented recursive trees, all the issues leading to the possible roots with a given shape T have the same probability. Then, for $t \sim \mathrm{PA(n)}$:

$$\mathbb{P}(u = 1 \cap t^\circ = T) = \frac{1}{\overline{\mathrm{Aut}}(u,T)},$$

so we have:

$$\mathbb{P}(u = 1 | t^\circ = T) = \frac{|\{\text{plane-oriented rec. trees isomorphic to } T\}|}{\prod_{k=1}^{n-1}(2k-1)\,\overline{\mathrm{Aut}}(u,T)}.$$

With a previous proposition, that gives an expression of the above cardinality, we conclude that the desired likelihood equals:

$$L(u;T) = \frac{|T|!\,\mathrm{d}_T(\emptyset)! \prod_{v \neq \emptyset} (\mathrm{d}(v)-1)!}{(\prod_{k=1}^{n-1}(2k-1)\,\overline{\mathrm{Aut}}(u,T)) \prod_{v \in V(T)\setminus\mathcal{L}(T)} (|T|_{v\downarrow}\,\mathrm{Aut}(v,T))}.$$

We can rewrite this likelihood by putting aside all the constants:

$$L(u;T) = \frac{\mathrm{d}_T(\emptyset)}{\overline{\mathrm{Aut}}(u,T) \prod_{v \in V(T)\setminus\mathcal{L}(T)} (|T|_{v\downarrow}\,\mathrm{Aut}(v,T))} \frac{|T|! \prod_{v \in V(T)} (\mathrm{d}(v)-1)!}{\prod_{k=1}^{n-1}(2k-1)}.$$

Indeed, the second fraction is a constant on the subset of plane-oriented recursive trees with a given shape. Minimizing the inverse of this likelihood gives the expected result. □

This estimator gives a higher importance to vertices with a high degree, and it is what one would expect from it, as the degree of the root must be more important in this model. Unfortunately, this degree term in the estimator makes it difficult to use to show an impossibility result. However, for the preferential attachment model, there exists a simpler way to show such a result.

4.3 A simple lower bound

We begin this paragraph with a very important yet easy to show result.

Proposition 19. Let $\alpha_n = \prod_{k=1}^{n-2}(1+\frac{1}{2k})$ for $n \in \mathbb{N}^*$. The sequence $(\frac{\mathrm{d}_n(i)}{\alpha_n} = \frac{\mathrm{d}_{\mathrm{PA}(n)}(i)}{\alpha_n})_n$ is a positive martingale for the canonical filtration $(\mathcal{F}_n)_n$ so it converges almost surely.

Proof. Let $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left(\frac{\mathrm{d}_{n+1}(i)}{\alpha_{n+1}}|\mathcal{F}_n\right) &= \frac{1}{\alpha_{n+1}}\mathbb{E}(\mathrm{d}_{n+1}(i)|\mathcal{F}_n)\\ &= \frac{1}{\alpha_{n+1}}\left[(\mathrm{d}_n(i)+1)\frac{\mathrm{d}_n(i)}{2(n-1)} + \mathrm{d}_n(i)\left(1-\frac{\mathrm{d}_n(i)}{2(n-1)}\right)\right]\\ &= \frac{1}{\alpha_{n+1}}\left(1+\frac{1}{2(n-1)}\right)\mathrm{d}_n(i) = \frac{\mathrm{d}_n(i)}{\alpha_n}.\end{aligned}$$

□

Moreover,

$$\alpha_n = \frac{\prod_{k=1}^{n-1}(2k+1)}{\prod_{k=1}^{n-1}(2k)} = \frac{(2n-3)!}{2^{2(n-2)}(n-2)!^2} \sim \frac{2\sqrt{n}}{\sqrt{\pi}}$$

by Stirling's approximation. Then by Slutsky's theorem, $(\mathrm{d}_n(i)/\sqrt{n})_n$ converges in distribution to a random variable $\mathbf{d}_i$.

Theorem 20. There exists $c > 0, \varepsilon_0$ such that for all $\varepsilon \in]0, \varepsilon_0[$, any root-finding algorithm which satisfies:

$$\liminf_{n\to\infty} \mathbb{P}(1 \notin H(\mathrm{UA}(n)^\circ)) \geq 1-\varepsilon$$

must have $K(\varepsilon) \geq \frac{c}{\varepsilon}$.

Proof. Like for the previous model, we only have to show that the optimal procedure must have a probability of error at least ε for some finite n. We will show that:

$$\mathbb{P}\left(\overline{\mathrm{Aut}}(1,T) \geq \frac{2c}{\varepsilon}\right) \geq 2\varepsilon.$$

This implies that the probability of error of any procedure that outputs more than $\frac{c}{\varepsilon}$ vertices must be larger than ε. To show that, we recall that we know that the probability that 1 is a leaf is at least $\frac{1}{\sqrt{n}}$. Now, notice that conditioned to 1 being a leaf, all the leaves connected to 2 are isomorphic to 1. Condition to 1 being a leaf is the same as considering that the tree begins at 2. Considering the dominance $\mathrm{d}_{\frac{n}{2}}(1) \preccurlyeq \mathrm{d}_n(1) - \mathrm{d}_{\frac{n}{2}}(1)$, we get:

$$\mathbb{P}(\mathrm{d}_n(1) - \mathrm{d}_{\frac{n}{2}} > c_0\sqrt{n}) \geq \mathbb{P}(\mathrm{d}_{\frac{n}{2}}(1) > c_0\sqrt{n}) \geq 1-2\varepsilon,$$

for n large enough so $\mathbb{P}(\mathrm{d}_{\frac{n}{2}}(1) > c_0\sqrt{n})$ is close enough to $\mathbb{P}(\mathbf{d}_1 > c_0) \geq 1-\varepsilon$. Then we have to be sure that in this vertices, we have enough leaves when n goes huge. Let $k \in [\![\frac{n}{2}+1, n]\!]$, we have:

$$\mathbb{P}(k \in \mathcal{L}(T)) = \left(1-\frac{1}{2(k-1)}\right)\cdots\left(1-\frac{1}{2(n-1)}\right).$$

Taking the logarithm, we get:

$$\log(\mathbb{P}(k \in \mathcal{L}(T))) \geq \sum_{i=\frac{n}{2}}^{n-1} \log\left(1 - \frac{1}{2i}\right) \sim \log(1/2).$$

So, independently of k, for n large enough, one has $\mathbb{P}(k \in \mathcal{L}(T)) \geq \rho > 0$. Then when n goes to infinity:

$$\mathbb{P}(|\{n/2, \ldots, n\} \cap \mathcal{L}(T)| \geq \rho c_0 \sqrt{n}) \geq 1 - 2\varepsilon.$$

Then, for ε_0 small enough and $n \sim \frac{1}{\varepsilon^2}$, we obtain the desired result. □

Comparing this bound with the one for the uniform attachment model confirms the idea that it is more difficult to find the root for the preferential attachment model. In particular, it is exponentially more difficult.

4.4 An upper bound which joins the lower one

In this paragraph, we study an optimal root-finding algorithm for the preferential attachment, in the sense that it joins the lower bound we just exposed. The difference with all the algorithms we studied before is that it has a random packet size. The proof is way more complicated, and uses a lot of difficult results. We will not prove these important lemmas, as their proofs would make this document way too long (and also because I would not have had the time to study them all).

Theorem 21. Let $\eta \in]0, \frac{1}{8}[$. For $\varepsilon \in]0, 1[$, there exists a sequence of packets $(\mathcal{P}_\varepsilon(n))_n$ that only depends on the structure of each tree such that:

$$\mathbb{P}\left(\sup_{n \geq 1} |\mathcal{P}_\varepsilon(n)| \leq \varepsilon^{-1-\eta} \text{ and } 1 \in \mathcal{P}_\varepsilon(n) \text{ for all } n\right) \geq 1 - 2\varepsilon^{1-\eta}.$$

As η goes to zero, we get a root-finding algorithm that joins the lower bound we just studied. One can notice that it can be computed in polynomial time (in the number of vertices).

Proof. For this theorem, one of the main results we need is the convergence of the renormalized degrees. We define:

$$D_i(n) = \frac{\mathrm{d}_n(i)}{\alpha_n \sqrt{\pi}}.$$

We already know that for each i, D_i converges almost surely towards a random variable we call $\mathbf{D}_i$. It appears that we have a stronger convergence for these variables, that is, a ℓ^∞ convergence (shown in [10]):

$$\|D_i(n) - \mathbf{D}_i\|_\infty \xrightarrow[n \to \infty]{\text{a.s.}} 0.$$

Now let us see how we will define the packet. We know that in general, the root has an important degree. But if we define the packet by all the vertices with an important degree, we get a bound that is not better than what we already did. We have to go to the next order, considering the neighbors of vertices with a high degree. Indeed, if the root has exceptionally few children, then one of its children must have a lot. We can show that $\mathbf{D}_1 \mathbf{D}_2^2 \geq \varepsilon$ with probability of order $1 - \varepsilon^{-1-\eta}$, which motivates the following definition:

$$\mathcal{P}_\varepsilon(n) = \left\{ i \in [\![1, n]\!], \exists j \sim i, D_j(n) D_i(n)^2 > \varepsilon \text{ or } D_i(n) D_j(n)^2 > \varepsilon \right\}.$$

The $\sim$ symbol is a commonly used one to denote the connection between two vertices in a graph.

First, we will prove that 1 is in the packet with a $1-\varepsilon^{1-\eta}$ probability:

$$\mathbb{P}(1 \in \mathcal{P}_\varepsilon(n) \text{ for all } n) \geq 1-\varepsilon^{1-\eta}$$

for $\eta \in]0,1[$ fixed and ε small enough. The key idea is to express what we want in terms of limit distribution that we are able to compute.

$$\begin{aligned}
1-\mathbb{P}(1 \in \mathcal{P}_\varepsilon(n)\, \forall n \text{ large enough}) &= \mathbb{P}(|\{n, 1 \in \mathcal{P}_\varepsilon(n)\}| = +\infty) \\
&= \mathbb{P}(|n, D_1(n)D_2(n)^2 \leq \varepsilon \text{ and } D_2(n)D_1(n)^2 \leq \varepsilon| = +\infty) \\
&\leq \mathbb{P}(\mathbf{D}_1\mathbf{D}_2^2 \leq \varepsilon \text{ and } \mathbf{D}_1^2\mathbf{D}_2 \leq \varepsilon),
\end{aligned}$$

where the last inequality follows from the definition of almost sure convergence. It turns out we can express the distributions of $\mathbf{D}_1$ and $\mathbf{D}_2$ in terms of usual laws, and with some calculus (special case of theorem 1.1 of [10] and then [5]), get to:

$$\forall \eta > 0, \forall \varepsilon > 0 \text{ small enough}, \mathbb{P}(\mathbf{D}_1\mathbf{D}_2^2 \leq \varepsilon \text{ and } \mathbf{D}_1^2\mathbf{D}_2 \leq \varepsilon | 3 \sim 2) \leq \varepsilon^{1-\eta}.$$

However, by a symmetry argument, we know that:

$$\mathbb{P}(\mathbf{D}_1\mathbf{D}_2^2 \leq \varepsilon \text{ and } \mathbf{D}_1^2\mathbf{D}_2 \leq \varepsilon) = \mathbb{P}(\mathbf{D}_1\mathbf{D}_2^2 \leq \varepsilon \text{ and } \mathbf{D}_1^2\mathbf{D}_2 \leq \varepsilon | 3 \sim 2),$$

so that:

$$\mathbb{P}(1 \in \mathcal{P}_\varepsilon(n)\, \forall n \text{ large enough}) \geq 1-\varepsilon^{1-\eta}.$$

We know want this to be uniform in n. The solution for that is Markov property with Doob's inequality. Let us define:

$$\theta = \inf\{n \geq 1, D_1(n)D_2(n)^2 \leq \varepsilon \text{ and } D_2(n)D_1(n)^2 \leq \varepsilon\},$$

so that:

$$\theta < +\infty \Leftrightarrow \exists n \geq 1, 1 \notin \mathcal{P}_\varepsilon(n).$$

We see that θ is a stopping time, so by Markov property, $((D_1(n))_{n\geq\theta}|\mathcal{F}_\theta)$ is a martingale, to which we can apply Doob's inequality:

$$\mathbb{P}(\sup_{n\geq\theta} D_1(n) \geq 3D_1(\theta) | \theta < +\infty) \leq \frac{1}{3}.$$

By doing the same for D_2, we deduce that there is a probability at least $1-\frac{1}{3}-\frac{1}{3}=\frac{1}{3}$ that $D_1(n) \leq 3D_1(\theta)$ and $D_2(n) \leq 3D_2(\theta)$. In particular:

$$\mathbb{P}(\mathbf{D}_1 \leq 3D_1(\theta) \text{ and } \mathbf{D}_2 \leq 3D_2(\theta) | \theta < +\infty) \geq \frac{1}{3}.$$

So on the one hand, we have:

$$\begin{aligned}
\mathbb{P}(\mathbf{D}_1 \leq 3D_1(\theta) \text{ and } \mathbf{D}_2 \leq 3D_2(\theta)) &= \mathbb{P}(\mathbf{D}_1 \leq 3D_1(\theta) \text{ and } \mathbf{D}_2 \leq 3D_2(\theta) | \theta < +\infty)\, \mathbb{P}(\theta < +\infty) \\
&\geq \frac{\mathbb{P}(\theta < +\infty)}{3},
\end{aligned}$$

and on the other hand:

$$\mathbb{P}(\mathbf{D}_1 \leq 3D_1(\theta) \text{ and } \mathbf{D}_2 \leq 3D_2(\theta)) \leq \mathbb{P}(\mathbf{D}_1\mathbf{D}_2^2 \leq 3^3\varepsilon \text{ and } \mathbf{D}_2\mathbf{D}_1^2 \leq 3^3\varepsilon) \leq \varepsilon^{1-\eta}.$$

By joining the two inequalities, one gets:

$$\mathbb{P}(\theta < +\infty) \leq 3^{3(1-\eta)}\varepsilon^{1-\eta},$$

so we obtain what we want by readjusting ε.

The second and last step of the proof consists in showing that we can control the size of the packets: for $\eta \in]0, \frac{1}{8}[$, we want

$$\mathbb{P}(\sup_{n\geq 1} |\mathcal{P}_\varepsilon(n)| \leq \varepsilon^{1-\eta}) \geq 1 - \varepsilon^{1-\eta}.$$

We will use methods that are similar to the previous step of the proof. We will obtain estimations on degrees thanks to the limit values and their known distribution and Markov property to have uniformity in n. After that, we will use the definition of the packet and our estimations to conclude.

A first concentration inequality we can have (thanks to the distribution of $\mathbf{D}_1$, [10],[5]) is:

$$\mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{\sqrt{i}}\right) \leq C_0 \exp(-A^{-\frac{2}{3}}),$$

where $A \geq 8$ and $C_0 > 0$. However, if we prove that for all $i \geq 1$:

$$\mathbb{P}\left(\sup_{n\geq i} D_i(n) \geq \frac{A}{\sqrt{i}}\right) \leq C_1 \, \mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{\sqrt{i}}\right),$$

it gives, all in all,

$$\mathbb{P}\left(\sup_{n\geq i} D_i(n) \geq \frac{A}{\sqrt{i}}\right) \leq C \exp\left(-\frac{A^{\frac{2}{3}}}{4^{\frac{2}{3}}}\right).$$

To prove the statement above, we will use Markov property, let:

$$\theta_i = \inf\left\{n \geq i, D_i(n) \geq \frac{A}{\sqrt{i}}\right\},$$

which is a stopping time such that:

$$\mathbb{P}(\theta_i < +\infty) = \mathbb{P}\left(\sup_{n\geq i} D_i(n) \geq \frac{A}{\sqrt{i}})\right).$$

We get, with the law of total probabilities first and Markov property:

$$\begin{aligned}\mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{4\sqrt{i}}\right) &\geq \sum_{k=i}^{\infty} \mathbb{P}\left(D_i(k) \geq \frac{A}{\sqrt{i}}, \theta_i = k\right) \mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{4\sqrt{i}} | D_i(k) \geq \frac{A}{\sqrt{i}}, \theta_i = k\right) \\ &= \sum_{k=i}^{+\infty} \mathbb{P}(\theta_i = k) \, \mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{4\sqrt{i}} | D_i(k) \geq \frac{A}{\sqrt{i}}\right).\end{aligned}$$

Now we can lower bound the sum of θ probabilities by $\mathbb{P}(\theta_i < \infty)$ and the other term by the infimum of all these values, and check that this infimum is strictly positive.

$$\begin{aligned}&\mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{4\sqrt{i}}\right) \geq \\ &\mathbb{P}(\theta_i < \infty) \inf\left\{\mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{4\sqrt{i}} | \mathrm{d}_k(i) = m\right), k \geq i, m \geq 1, 2(k-1) > m, m \geq \frac{A}{\sqrt{i}}(\sqrt{\pi}\alpha_k)\right\}.\end{aligned}$$

It turns out (once again Theorem 1.1 in [10]) we know the distribution do $\mathbf{D}_i$ conditioned on the degree $\mathrm{d}_k(i)$ being equal to m. We call this variable $X_{m,k}$. We know the first two moments of this distribution. First:

$$\mathbb{E}(X_{m,k}) = \frac{m}{2(k-1)} \frac{\Gamma(k)}{\Gamma(\frac{2k-1}{2})} \leq \sqrt{k}$$

so that:

$$\mathbb{P}\left(X_{m,k} \geq \frac{A}{4\sqrt{i}}\right) \geq \mathbb{P}\left(X_{m,k} \geq \frac{A}{4\sqrt{ik}}\,\mathbb{E}(X_{m,k})\right).$$

Now we want to use Paley-Zygmund inequality. To do so, we have to check that $\vartheta = \frac{A}{4\sqrt{ik}} \in]0,1[$. But:

$$\vartheta = \frac{A}{4\sqrt{ik}} < \frac{k-1}{2\sqrt{\pi}\alpha_k\sqrt{k}} < \frac{2}{3}.$$

So, by Paley-Zygmund inequality:

$$\mathbb{P}\left(X_{m,k} \geq \frac{A}{4\sqrt{i}}\right) \geq (1-\vartheta)^2 \frac{\mathbb{E}(X_{m,k})^2}{\mathbb{E}(X_{m,k}^2)}.$$

With

$$\mathbb{E}(X_{m,k}^2) = \frac{m(m+1)}{2(k-1)(2k-1)}\,\frac{\Gamma(\frac{2k+1}{2})}{\Gamma(\frac{2k-1}{2})},$$

we obtain:

$$\begin{aligned}
\mathbb{P}\left(X_{m,k} \geq \frac{A}{4\sqrt{i}}\right) &\geq \left(1-\frac{A}{4\sqrt{ik}}\right)^2 \frac{m}{m+1}\,\frac{(2k-1)\Gamma(k)^2}{2(k-1)\Gamma(\frac{2k-1}{2})\Gamma(\frac{2k+1}{2})} \\
&= \left(1-\frac{A}{4\sqrt{ik}}\right)^2 \frac{m}{m+1}\,\frac{1}{k-1}\frac{\Gamma(k)^2}{\Gamma(k-\frac{1}{2})^2} \\
&= \left(1-\frac{A}{4\sqrt{ik}}\right)^2 \frac{1}{2}\,\frac{1}{k-1}\frac{\Gamma(k)^2}{\Gamma(k-\frac{1}{2})^2} \\
&\geq C_2\left(1-\frac{A}{4\sqrt{ik}}\right)^2,
\end{aligned}$$

where $C_2 > 0$ does not depend on k. Also, $2k-2 > \frac{A}{\sqrt{i}}\sqrt{\pi}\alpha_k$, so $(1-\frac{A}{4\sqrt{ik}})^2$ is lower-bounded by a constant that does not depend on k, so our infimum, which we set being equal to C_1^{-1}, is strictly positive, so for all $i \geq 1$:

$$\mathbb{P}(\theta_i < +\infty) \leq \mathbb{P}\left(\sup_{n\geq i} D_i(n) \geq \frac{A}{\sqrt{i}}\right) \leq C_1\,\mathbb{P}\left(\mathbf{D}_i \geq \frac{A}{\sqrt{i}}\right),$$

which is what we wanted to prove.

Now that we have our estimation on the degrees, the idea is the following, we are going to control the values of the degrees with high probability, and then, knowing the values of the degrees, we will see that we can control the cardinality of the packet. Let $\eta \in]0, \frac{1}{8}[$, and ε small enough so that $\varepsilon^{-\eta} \geq -\kappa\log(\varepsilon)$ and $\eta\varepsilon^{-\eta} \geq 2\kappa$ for $\kappa \in \{C, 4^{-\frac{2}{3}}\}$. Therefore, the estimation on the concentration of the degrees we just obtained and a union bound give:

$$\mathbb{P}\left(\exists i, \sup_{n\geq i} D_i(n) \geq \frac{(i/\varepsilon)^{\frac{3\eta}{2}}}{\sqrt{i}}\right) \leq C\sum_{i\geq 1}\exp\left(-\frac{\varepsilon^{-\eta}i^{\eta}}{4^{\frac{2}{3}}}\right).$$

With $i^{\eta} = \mathrm{e}^{\eta\log(i)} \leq 1+\eta\log(i)$ we get:

$$\mathbb{P}\left(\exists i, \sup_{n\geq i} D_i(n) \geq \frac{(i/\varepsilon)^{\frac{3\eta}{2}}}{\sqrt{i}}\right) \leq C\sum_{i\geq 1}\exp\left(-\frac{\varepsilon^{-\eta}}{4^{\frac{2}{3}}}(1+\eta\log(i))\right)$$

And finally, with our hypotheses on ε:

$$\mathbb{P}\left(\exists i, \sup_{n\geq i} D_i(n) \geq \frac{(i/\varepsilon)^{\frac{3\eta}{2}}}{\sqrt{i}}\right) \leq C\sum_{i\geq 1}\exp(\log(\varepsilon) - 2\log(i)) = C\zeta(2)\,\varepsilon = O(\varepsilon).$$

Then, if we introduce the event:

$$G = \left\{\forall i \geq 1, \sup_{n\geq i} D_i(n) < \frac{(i/\varepsilon)^{\frac{3\eta}{2}}}{\sqrt{i}}\right\},$$

we just proved that $\mathbb{P}(G) \geq 1 - O(\varepsilon)$. Now we can estimate the number of trees in the packet that satisfy the upper-bounding of the degrees.

$$\sup_{n\geq 1}|\mathcal{P}_\varepsilon(n)|\mathbf{1}_G \leq 2\left|\left\{(i,j), 1\leq i<j, i\sim j, \frac{(i/\varepsilon)^{3\eta}(j/\varepsilon)^{\frac{3\eta}{2}}}{i\sqrt{j}} > \varepsilon \text{ or } \frac{(j/\varepsilon)^{3\eta}(i/\varepsilon)^{\frac{3\eta}{2}}}{j\sqrt{i}} > \varepsilon\right\}\right|\mathbf{1}_G.$$

Now we are going to simplify this disjunction of two inequalities. First, notice that the first inequality is equivalent to $i\sqrt{j} < \varepsilon^{-\frac{1+9\eta/2}{1-3\eta}}$. Then, the function $i \mapsto i^{-(\frac{1}{2}-\frac{3\eta}{2})}$ is decreasing due to the value of η so that the second inequality implies $j^{-\frac{3\eta}{2}-\frac{3}{2}} > \varepsilon^{\frac{9\eta}{2}}$, which is equivalent to $\varepsilon < j^{\frac{1}{3\eta}-\frac{1}{3}}$ which is always true because j is an integer and $\eta < 1$. To lighten the notation, we set $-\frac{1+9\eta/2}{1-3\eta} = -1-\eta'$ so that, all in all:

$$\sup_{n\geq 1}|\mathcal{P}_\varepsilon(n)|\mathbf{1}_G \leq 2\left|\left\{(i,j), 1\leq i<j, i\sim j, i\sqrt{j} \leq \varepsilon^{-1-\eta'}\right\}\right|\mathbf{1}_G.$$

We can simplify this even more. Notice that if $j > \varepsilon^{-2-2\eta'}$, the inequality cannot be satisfied. Also, because we are studying trees, there is for each $j \geq 2$ a unique $i < j$ such that $i \sim j$, so:

$$\sup_{n\geq 1}|\mathcal{P}_\varepsilon(n)|\mathbf{1}_G \leq 2\sum_{1\leq j\leq \varepsilon^{-2-\eta'}} \mathbf{1}_{\{\exists 1\leq i<j, i\leq \frac{\varepsilon^{-1-\eta'}}{\sqrt{j}}\}}\,\mathbf{1}_G.$$

Let $X_j = \mathbf{1}_{\{\exists 1\leq i<j, i\leq \frac{\varepsilon^{-1-\eta'}}{\sqrt{j}}\}}$. If $(\mathcal{G}_j)$ is the natural filtration generated by $(\mathrm{d}_k(i))_{1\leq i\leq k\leq j}$, then:

$$X_j \,|\, \mathcal{G}_{j-1} \sim \mathcal{B}\left(\sum_{1\leq i\leq (\frac{\varepsilon^{-1-\eta'}}{\sqrt{j}}\wedge j)} \frac{\mathrm{d}_{j-1}(i)}{2(j-2)}\right),$$

where $\mathcal{B}$ denotes Bernoulli distribution. Therefore, conditionally on $\mathcal{G}_{j-1}$, $\mathbf{1}_G X_j$ is stochastically dominated by $Y_j \sim \mathcal{B}(p_j)$, independent of $\mathcal{G}_{j-1}$, as long as p_j is greater than the above parameter. We can then use the fact that we are on the event G to use the bound on the degrees and get:

$$p_j = 1 \wedge \sum_{1\leq i\leq \frac{\varepsilon^{-1-\eta'}}{\sqrt{j}}} \frac{\sqrt{\pi}\alpha_j(i/\varepsilon)^{\frac{3\eta}{2}}}{2(j-1)\sqrt{i}}.$$

Using the equivalent for α_j, we get, for some constant $M > 0$ independent of j:

$$\begin{aligned} p_j &\leq 1 \wedge \frac{M}{\sqrt{j}}\varepsilon^{-\frac{3\eta}{2}} \sum_{i=1}^{\varepsilon^{-1-\eta'}/\sqrt{j}} i^{-\frac{1}{2}+\frac{3\eta}{2}} \\ &\leq 1 \wedge M\left(\frac{\varepsilon^{-\frac{1}{2}}}{j^{\frac{3}{4}}}\right)^{1+\eta''}, \end{aligned}$$

by calculating the geometrical sum, and with $\eta'' \to 0$ when $\eta \to 0$. To get what we want, this bounds turns out not be sufficient. Therefore, if $j \geq \varepsilon^{-\frac{5}{6}}$, we have $p_j < \frac{1}{2}$, so we can dominate Y_j by the Poisson distribution $\mathcal{P}(2p_j)$. It is interesting to dominate by independent Poisson variables because they are additive in the parameter. For other values of j, we set $Y_j \preccurlyeq 1$, in order to get:

$$\sup_{n\geq 1} |\mathcal{P}_\varepsilon(n)|\mathbf{1}_G \preccurlyeq 2\left(\varepsilon^{-\frac{5}{6}} + \mathcal{P}\left(M \sum_{\varepsilon^{-\frac{5}{6}} \leq j \leq \varepsilon^{-2-2\eta'}} \left(\frac{\varepsilon^{-\frac{1}{2}}}{j^{\frac{3}{4}}}\right)^{1+\eta''}\right)\right).$$

All this huge parameter of the Poisson distribution can be seen as smaller than $\varepsilon^{-1-\eta'''}$, where η''' goes to zero when η does. Finally, a concentration inequality (Benett's inequality, see *An improvement to Bennett's inequality for the Poisson distribution* on Terence Tao's website) gives:

$$\mathbb{P}(\mathcal{P}(\varepsilon^{-1-\eta'''}) \geq 2\varepsilon^{-1-\eta'''}) \leq \varepsilon,$$

for $\varepsilon > 0$ small enough. Gathering up all the pieces we obtain:

$$\begin{aligned}
\mathbb{P}\left(\sup_{n\geq 1} \mathcal{P}_\varepsilon(n) \geq 6\varepsilon^{-1-\eta'''}\right) &= \mathbb{P}\left(\sup_{n\geq 1} \mathcal{P}_\varepsilon(n)\mathbf{1}_G \geq 6\varepsilon^{-1-\eta'''}\right) + \mathbb{P}\left(\sup_{n\geq 1} \mathcal{P}_\varepsilon(n)\mathbf{1}_{\overline{G}} \geq 6\varepsilon^{-1-\eta'''}\right) \\
&\leq \mathbb{P}\left(2(\varepsilon^{-\frac{5}{6}} + \mathcal{P}(\varepsilon^{-1-\eta'''})) \geq 6\varepsilon^{-1-\eta'''}\right) + \mathbb{P}(\overline{G}) \\
&\leq O(\varepsilon) + \mathbb{P}\left(\mathcal{P}(\varepsilon^{-1-\eta'''}) \geq 3\varepsilon^{-1-\eta'''} - \varepsilon^{-\frac{5}{6}}\right) \\
&\leq O(\varepsilon) + \mathbb{P}\left(\mathcal{P}(\varepsilon^{-1-\eta'''}) \geq 2\varepsilon^{-1-\eta'''}\right) \text{ for } \varepsilon \text{ small enough} \\
&\leq O(\varepsilon) \leq \varepsilon^{-1-\eta'''} \text{ for } \varepsilon \text{ small enough.}
\end{aligned}$$

This concludes the proof of the theorem. □

Conclusion

Even if we can get root-finding algorithms with simple Pólya urn models, it becomes way more difficult when we want to be more tight on the number of vertices we output. The proofs that I exposed which give more precise algorithms involve very important and tough probability results, even though the models studied can be explicated simply. For preferential attachment model, we managed to join the upper bound to the impossibility result. It has not been done yet for the uniform attachment model. It seems more difficult because the trees are more spread. Furthermore, the technique used for the preferential attachment can not be implemented for this model because the convergence result for the degrees is specific to preferential attachment trees.

Other possible extensions to this are the study of other degree exponents ($\alpha \neq 0, 1$), the study of other models of random growing trees, like the *affine preferential attachment trees*[3] or even more generally the study of random growing graphs.

* *
*

References

[1] Hofstad, R. v. d. *Preferential Attachment Models*, 256–300. Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, 2016).

[3] Where the probability is now proportional to the degree to the power α, plus a constant β.

[2] Bollobás, B., Riordan, O., Spencer, J. & Tusnády, G. The degree sequence of a scale-free random graph process. *Random Struct. Algorithms* **18**, 279–290 (2001).

[3] Kent, R. G. & Zipf, G. K. Relative frequency as a determinant of phonetic change. *Language* **6**, 86 (1930). URL https://api.semanticscholar.org/CorpusID:149728126.

[4] Bubeck, S., Devroye, L. & Lugosi, G. Finding adam in random growing trees (2015). URL https://arxiv.org/abs/1411.3317. 1411.3317.

[5] Contat, A., Curien, N., Lacroix, P., Lasalle, E. & Rivoirard, V. Eve, adam and the preferential attachment tree (2023). URL https://arxiv.org/abs/2303.04752. 2303.04752.

[6] Devroye, L. Branching processes in the analysis of the heights of trees. *Acta Informatica* **24**, 277–298 (1987). URL https://api.semanticscholar.org/CorpusID:11540185.

[7] Mahmoud, H. *Polya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science (CRC Press, 2008). URL https://books.google.fr/books?id=7Bizo28c2LQC.

[8] Hardy, G. H. & Ramanujan, S. Asymptotic Formulaæ in Combinatory Analysis. *Proceedings of the London Mathematical Society* **s2-17**, 75–115 (1918). URL https://doi.org/10.1112/plms/s2-17.1.75. https://academic.oup.com/plms/article-pdf/s2-17/1/75/4394266/s2-17-1-75.pdf.

[9] Janson, S. Limit theorems for triangular urn schemes. *Probability Theory and Related Fields* **134**, 417–452 (2006).

[10] Peköz, E. A., Röllin, A. & Ross, N. Joint degree distributions of preferential attachment random graphs (2016). URL https://arxiv.org/abs/1402.4686. 1402.4686.