



**HAL**  
open science

# Revisiting Tree Isomorphism: AHU Algorithm with Primes Numbers

Florian Ingels

► **To cite this version:**

Florian Ingels. Revisiting Tree Isomorphism: AHU Algorithm with Primes Numbers. 2023. hal-04232137v1

**HAL Id: hal-04232137**

**<https://hal.science/hal-04232137v1>**

Preprint submitted on 7 Oct 2023 (v1), last revised 13 Feb 2024 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Revisiting Tree Isomorphism: AHU Algorithm with Primes Numbers

Florian Ingels

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISAL, F-59000 Lille, France

## Abstract

The AHU algorithm has been the state of the art since the 1970s for determining in linear time whether two unordered rooted trees are isomorphic or not. However, it has been criticized (by Campbell and Radford) for the way it is written, which requires several (re)readings to be understood, and does not facilitate its analysis. In this paper, we propose an alternative version of the AHU algorithm, which addresses this issue by being designed to be clearer to understand and implement, with the same theoretical complexity and equally fast in practice.. Whereas the key to the linearity of the original algorithm lay on the careful sorting of lists of integers, we replace this step by the multiplication of lists of prime numbers, and prove that this substitution causes no loss in the final complexity of the new algorithm.

**Keywords:** tree isomorphism, AHU algorithm, prime numbers multiplication

## 1 Introduction

### 1.1 Context

The Aho, Hopcroft and Ullman (AHU) algorithm, introduced in the 1970s [1, Example 3.2], establishes that the tree isomorphism problem can be solved in linear time, whereas the more general graph isomorphism problem is still an open problem today, where no proof of NP-completeness nor polynomial algorithm is known [34], even though very efficient algorithms exist [27, 4].

As far as we know, AHU remains the only state-of-the-art algorithm for practically determining whether two trees are isomorphic. Recently, Liu [25] proposed to represent a tree by a polynomial of two variables, computable in linear time, and where two trees have the same polynomial if and only if they are isomorphic. Unfortunately, the existence of an algorithm to determine the equality of two polynomials in polynomial time is still an open question [32]. We should also mention [10], which proposes an alternating logarithmic time algorithm for tree isomorphism – under NC complexity class framework, that is, problems efficiently solvable on a parallel computer [5].

One criticism made of the AHU algorithm concerns the way the algorithm is presented in the original article, which is claimed to be

*utterly opaque. Even on second or third reading. When an algorithm is written it should be clear, it should persuade, and it should lend itself to analysis.*

— Douglas M. Campbell and David Radford [11]

To the best of our knowledge, this remark seems to have remained a dead letter in the community, and no alternative, clearer version of the algorithm seems ever to have been published – with the exception of Campbell and Radford themselves, but with quasilinear complexity instead of linear.

In this article, we propose to revisit the AHU algorithm by giving an alternative version that is intended to be easier to understand and implement, with the same theoretical complexity and equally fast in practice. This variant is based on replacing multiset hashing, originally carried out in the form of sorting lists of integers, by the elementary principle of multiplying lists of prime numbers.

The rest of this section is devoted to introducing the notations and definitions useful for the rest of the paper. Section 2 presents the AHU algorithm, while Section 3 presents our variant of the same algorithm and its complexity analysis; both versions are then compared numerically in Section 4.

## 1.2 Tree isomorphisms

A rooted tree  $T$  is a connected directed graph without any undirected cycle such that (i) there exists a special node called the root and (ii) any node but the root has exactly one parent. The parent of a node  $u$  is denoted by  $\mathcal{P}(u)$ , whereas its children are denoted by  $\mathcal{C}(u)$ . The leaves  $\mathcal{L}(T)$  of  $T$  are the nodes without any children. Rooted trees are said to be unordered if the order among siblings is not significant. In this paper, we use *trees* to refer to unordered rooted trees.

The degree of a node  $u$  is defined as  $\deg(u) = \#\mathcal{C}(u)$  and the degree of a tree  $T$  as  $\deg(T) = \max_{u \in T} \deg(u)$ . The depth  $\mathcal{D}(u)$  of a node  $u$  is the length of the path between  $u$  and the root. The depth  $\mathcal{D}(T)$  of  $T$  is the maximal depth among all nodes. The level of a node  $u$  is defined as  $\mathcal{D}(T) - \mathcal{D}(u)$ . The sets of nodes of level  $d$  in a tree  $T$  is denoted by  $T^d$ , and the mapping  $d \mapsto T^d$  can be constructed in linear time by a simple traversal of  $T$ .

**Definition 1.** Two trees  $T_1$  and  $T_2$  are said to be isomorphic if there exists a bijective mapping  $\varphi : T_1 \rightarrow T_2$  so that (i) the roots are mapped together and (ii) for any  $u, v \in T_1, v \in \mathcal{C}(u) \iff \varphi(v) \in \mathcal{C}(\varphi(u))$ .

Such a mapping  $\varphi$  is called a *tree isomorphism*. In other words, two trees are isomorphic if one can be obtained from the other by simply swapping the children of each node. An example of isomorphic trees is provided in Figure 1.

Whenever two trees  $T_1$  and  $T_2$  are isomorphic, we note  $T_1 \simeq T_2$ . It is well known that  $\simeq$  is an equivalence relation on the set of trees [37]. The *tree isomorphism problem* consists in deciding whether two trees are isomorphic or not.

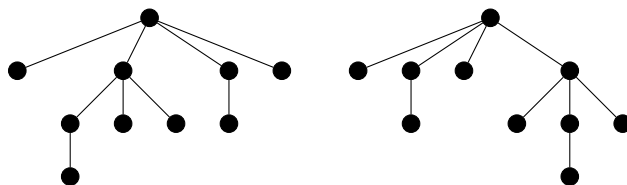


Figure 1: Two isomorphic trees.

For the broader graph isomorphism problem, it is not usual to explicitly construct the isomorphism  $\varphi$  – let us mention nonetheless [14, Section 3.3] and [19] – but rather to compute a certificate of non-isomorphism. For instance, Weisfeiler-Lehman algorithms, also known as colour refinement algorithms [18, 21], colour the nodes of each graph according to certain rules, and the histograms of the colour distributions are then compared: if they diverge, the graphs are not isomorphic. This test is not complete in the sense that there are non-isomorphic graphs with the same colour histogram – even though the distinguishing power of these algorithms is constantly being improved [16].

While the graph isomorphism problem is not solved in the general case, it is solved for trees by virtue of the AHU algorithm, which is built on a colouring principle similar to that of Weisfeiler-Lehman.

## 2 The Aho, Hopcroft and Ullman algorithm

In this section we introduce AHU algorithm, that solves the tree isomorphism problem. First, we present the general principle of the algorithm in Section 2.1, before reproducing and commenting in Section 2.2 the original presentation of the algorithm as it can be found in [1].

### 2.1 Principle

In [11], Campbell and Radford provide a very clear, step-by-step exposition of the intuitions that lead to the AHU algorithm, and we invite the interested reader to consult it. For the sake of self-containment, we offer here another intuition of how the AHU algorithm works, presented as a colouring process, thus making the connection with Weisfeiler-Lehman algorithms for graph isomorphism.

The core idea behind AHU algorithm is to provide each node in trees  $T_1$  and  $T_2$  a canonical representative of its equivalence class for  $\simeq$ , thus containing all the information about its descendants. The trees are isomorphic if and only if the canonical representatives of the roots are identical.

The nodes of both trees are simultaneously browsed in ascending levels. Suppose that each node  $u$  of level  $d - 1$  has been assigned a colour  $c(u)$ , supposed to represent its equivalence class for the relation  $\simeq$ . Each node  $u$  of level  $d$  is associated with a multiset  $\mathcal{C}_c(u) = \{\{c(v) : v \in \mathcal{C}(u)\}\}$  – if  $u$  is a leaf, this multiset is denoted  $\emptyset$ . Each distinct multiset is given a colour, which is assigned to the corresponding nodes. An illustration is provided in Figure 2.

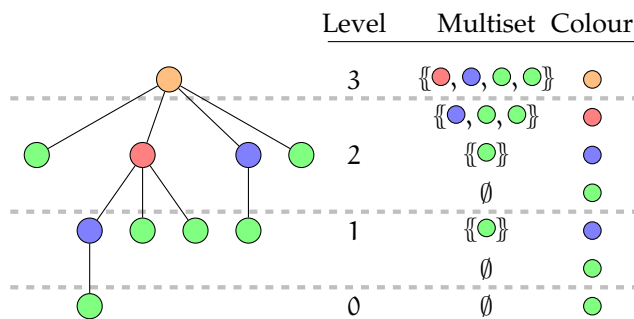


Figure 2: Assigning colours to nodes in AHU algorithm.

---

#### Algorithm 1: IDEALAHU

---

**Input:**  $T_1, T_2$   
**Output:**  $\top$  if and only if  $T_1 \simeq T_2$

```

1 if  $\mathcal{D}(T_1) \neq \mathcal{D}(T_2)$  then
2   return  $\perp$ 
3 else
4   for d from 0 to  $\mathcal{D}(T_1)$  do
5      $k \leftarrow 0$ 
6     Let  $f : \emptyset \mapsto 0$ 
7     for  $i \in \{1, 2\}$  and  $u \in T_i^d$  do
8        $\mathcal{C}_c(u) \leftarrow \{\{c(v) : v \in \mathcal{C}(u)\}\}$ 
9       if  $f(\mathcal{C}_c(u))$  is not defined then
10         $k \leftarrow k + 1$ 
11        Define  $f(\mathcal{C}_c(u)) = k$ 
12         $c(u) \leftarrow f(\mathcal{C}_c(u))$ 
13     if  $\{\{c(u) : u \in T_1^d\}\} \neq \{\{c(u) : u \in T_2^d\}\}$  then
14       return  $\perp$ 
15   return  $\top$ 

```


---

The latter can be addressed using pigeonhole sort [8] in linear time; whereas for the former, there are two main options: either by using hash functions specifically designed for multisets [12, 26], or by treating these

multisets as lists that are sorted before being hashed or compared. The second approach is commonly used, both by AHU and by Weisfeiler-Lehman algorithms.

Before examining the actual AHU algorithm in detail, we investigate the complexity of Algorithm 1. If we assume that determining whether  $f(\mathcal{C}_c(u))$  is defined or not, at line 9, can be accomplished in constant time (e.g. assuming a perfect hash function [23] working with multisets), then we have the following result.

**Proposition 1.** *Algorithm 1 runs in  $O(n)$ , where  $n = \#T_1 = \#T_2$ .*

**Proof.** Fix a level  $d$  and a node  $u \in T_i^d$ . Building  $\mathcal{C}_c(u)$  requires  $O(\deg(u))$ ; noticing that  $\sum_{u \in T_i^d} \deg(u) = \#T_i^{d-1}$  and that the comparison in line 13 can be done in  $O(\#T_i^d)$  – e.g. with pigeonhole sort [8]; summing over  $d$  leads to the result. 

Notably, this establishes that the tree isomorphism problem is solvable in linear time, provided that the assumption made above is valid. Note that there also exists a naive algorithm for tree isomorphism in  $O(n^2)$  [11] which makes extensive use of Knuth tuples [22].

## 2.2 Original algorithm

The description of the AHU algorithm in the original article [1, Example 3.2] is quite different from what has been presented previously. For the sake of self-containedness, we reproduce it here, where only minor changes have been made to fit the notations used in this paper:

1. First, assign to all leaves in  $T_1$  and  $T_2$  the integer 0.
2. Assume by induction that all nodes at level  $d - 1$  of  $T_1$  and  $T_2$  have been assigned an integer. Let  $L_1$  (respectively  $L_2$ ) be the list of nodes in  $T_1$  (respectively  $T_2$ ) at level  $d - 1$  sorted by non-decreasing value of the assigned integers.
3. Assign to the nonleaves of  $T_1$  at level  $d$  a tuple of integers by scanning the list  $L_1$  from left to right and performing the following actions:
  - For each vertex on list  $L_1$  take the integer assigned to  $u$  to be the next component of the tuple associated with  $\mathcal{P}(u)$ .
  - On completion of this step, each nonleaf  $w$  of  $T_1$  at level  $d$  will have a tuple  $(i_1, i_2, \dots, i_k)$  associated with it, where  $i_1, \dots, i_k$  are the integers, in non-decreasing order, associated with the children of  $w$ .
  - Let  $S_1$  be the sequence of tuples created for the vertices of  $T_1$  on level  $d$ .
4. Repeat Step 3 for  $T_2$  and let  $S_2$  be the sequence of tuples created for the vertices of  $T_2$  on level  $d$ .
5. Sort  $S_1$  and  $S_2$  lexicographically. Let  $S'_1$  and  $S'_2$ , respectively, be the sorted sequence of tuples.
6. If  $S'_1$  and  $S'_2$  are not identical, then halt: the trees are not isomorphic. Otherwise, assign the integer 1 to those vertices of  $T_1$  on level  $d$  represented by the first distinct tuple on  $S'_1$ , assign the integer 2 to the vertices represented by the second distinct tuple, and so on. As these integers are assigned to the vertices of  $T_1$  on level  $d$ , replace  $L_1$  by the list of the vertices so assigned. Append the leaves of  $T_1$  on level  $d$  to the front of  $L_1$ . Do the same for  $L_2$ .  $L_1$  and  $L_2$  can now be used for the assignment of tuples to nodes at level  $d + 1$  by returning to Step 3.
7. If the roots of  $T_1$  and  $T_2$  are assigned the same integer,  $T_1$  and  $T_2$  are isomorphic.

Note that, in Step 5, the authors resort to a variant of radix sort [1, Algorithm 3.2]. Actually, the tree isomorphism problem and AHU algorithm are only introduced in the book as an application example of this sorting algorithm.

To analyse the complexity of AHU algorithm, the authors make the assumption that trees are sufficiently small so that they can be described by a  $k$  bit word (i.e. with a 64-bit machine,  $\#T < 2^{64}$ ). For the purpose of this paper, we reframe this assumption as follows.

**Assumption 1.** For any considered tree  $T$ ,  $\log \#T = O(1)$ .

With this assumption, they show that tree isomorphism can be solved in linear time.

**Theorem 1.** AHU algorithm runs in  $O(n)$  where  $n = \#T_1 = \#T_2$ .

**Proof.** See the proofs in [1, Example 3.2] for the whole algorithm and especially [1, Algorithm 3.2] for sorting lists  $S_1$  and  $S_2$  in Step 5.

Assumption 1 is made to ensure that the largest integer manipulated in the various lists is not too large, and therefore that the (linear) sorting algorithm for these lists can effectively consider these numbers as integers and not as sequences of 0s and 1s. 

**Remark 1.** If Assumption 1 is relaxed, there are (large) trees for which the algorithm runs in  $O(n \log n)$ ; see [11].

As already mentioned in the introduction, Campbell and Radford describe this formulation of the algorithm as “utterly opaque. Even on second or third reading.” (sic) [11]. In their view, this is detrimental to understanding the algorithm and being able to analyse and implement it. Based on this observation, the natural question that arises is whether it is possible to find a version of AHU algorithm that is easier to understand and analyze, while remaining linear under Assumption 1, and manipulates only elementary concepts, just as original AHU does by sorting lists of integers.

Hash functions designed for multisets have already been mentioned [12, 26], but they involve advanced concepts, which would make implementation difficult for non-specialists. For this reason, they are beyond the scope of this paper. Instead, we propose an algorithm that uses only elementary concepts, replacing hash of multisets by multiplications of primes numbers.

### 3 Revisiting AHU algorithm

In Algorithm 1, we need to associate a unique integer  $f(\mathcal{C}_c(u))$  to each distinct multiset  $\mathcal{C}_c(u)$  of integers encountered. There is a particularly simple and fundamental example where integers are associated with multisets: prime factorization. Indeed, through the fundamental theorem of arithmetic, there is a bijection between integers and multisets of primes. For example,  $360 = 2^3 \cdot 3^2 \cdot 5$  is associated to the multiset  $\{2, 2, 2, 3, 3, 5\}$ . Note that this bijection is well known [9], and has already been successfully exploited in the literature for prime decomposition, but also usual operations such as product, division, gcd and lcd of numbers [36].

To the best of our knowledge, this link has never been exploited to replace multiset hashing, a fortiori in the context of graph isomorphism algorithms – such as Weisfeiler-Lehman, or AHU for trees. Note, however, that this approach has been used in the context of evaluating poker hands [35], where prime multiplication has been preferred to sorting cards by value in order to get a unique identifier for each distinct possible hand.

Since the original AHU sorts lists of integers, the main difficulty in making this substitution is to ensure that the complexity of multiplying lists of primes does not exceed that of sorting lists of integers.

In Section 3.1, we present our version of AHU algorithm which uses multiplication of primes; while Section 3.2 presents its complexity analysis and shows that, under the same Assumption 1 as original AHU, our algorithm is still linear.

### 3.1 AHU algorithm with primes

Suppose that each node  $u$  at level  $d$  has received a prime number  $c(u)$ , assuming that all nodes at that level and of the same class of equivalence have received the same number. Then, to a node  $u$  at level  $d$ , instead of associating the multiset  $\mathcal{C}_c(u) = \{\{c(v) : v \in \mathcal{C}(u)\}\}$ , we associate the number  $N(u) = \prod_{v \in \mathcal{C}(u)} c(v)$ . The nodes of level  $d$  are then renumbered with prime numbers – where each distinct number  $N(u)$  gets a distinct prime. The fundamental theorem of arithmetic ensures that two identical multisets  $\mathcal{C}_c(\cdot)$  receive the same number  $N(\cdot)$ . The pseudocode for this new version of AHU is presented in Algorithm 2.

This version of AHU differs from the ideal version presented in Algorithm 1 by only few lines: line 4, which defines additional variables required for this version; line 9 (respectively line 8 in Algorithm 1), which replaces the multiset  $\mathcal{C}_c(u)$  with the number  $N(u)$ ; and line 11 (respectively line 10), that replaces the increment of  $k$  by providing a new prime number  $p$ .

The subroutine `NEXTPRIME`, introduced in Algorithm 3, returns the next prime not already used at the current level; if there is no unassigned prime in the current list  $P$ , then new primes are generated using a segmented version of the sieve of Eratosthenes.

---

#### Algorithm 2: PRIMESAHU

---

```

Input:  $T_1, T_2$ 
Output:  $\top$  if and only if  $T_1 \simeq T_2$ 
1 if  $\mathcal{D}(T_1) \neq \mathcal{D}(T_2)$  then
2   | return  $\perp$ 
3 else
4   |  $P \leftarrow [2, 3, 5, 7, 11, 13]$  and  $N_{\text{sieve}} \leftarrow 16$ 
5   | for  $d$  from 0 to  $\mathcal{D}(T_1)$  do
6     | Let  $f : 1 \mapsto 2$ 
7     |  $p \leftarrow 2$ 
8     | for  $i \in \{1, 2\}$  and  $u \in T_i^d$  do
9       |  $N(u) \leftarrow \prod_{v \in \mathcal{C}(u)} c(v)$ 
10      | if  $f(N(u))$  is not defined then
11        |  $N_{\text{sieve}}, P, p \leftarrow \text{NEXTPRIME}(N_{\text{sieve}}, P, p)$ 
12        | Define  $f(N(u)) = p$ 
13      |  $c(u) \leftarrow f(N(u))$ 
14      | if  $\{\{c(u) : u \in T_1^d\}\} \neq \{\{c(u) : u \in T_2^d\}\}$  then
15        | return  $\perp$ 
16   | return  $\top$ 

```

---



---

#### Algorithm 3: NEXTPRIME

---

```

Input:  $N_{\text{sieve}}$ , the largest number for which the sieve of Eratosthenes has already been done, the list
          $P$  of primes  $\leq N_{\text{sieve}}$  in ascending order, with  $\text{length}(P) \geq 6$ , and a prime  $p \in P$ 
1 if  $p$  is the last of element of  $P$  then
2   |  $N_{\text{sieve}}, P \leftarrow \text{SIEVEOFERATOSTHENES}(N_{\text{sieve}}, P)$ 
   | /* At least one new prime has been added to  $P$  */
3 Let  $p'$  be the next prime after  $p$  in  $P$ 
4 return  $N_{\text{sieve}}, P, p'$ 

```

---

Let us denote  $p_n$  the  $n$ -th prime number. There are well known bounds on the value of  $p_n$  [13, 31] – with

In denoting the natural logarithm and  $n \geq 6$ :

$$n(\ln n + \ln \ln n - 1) < p_n < n(\ln n + \ln \ln n). \quad (1)$$

Suppose we have the list of all primes  $P \leq N_{\text{sieve}}$ , where  $N_{\text{sieve}}$  is the largest integer sieved so far. With  $\#P = n - 1$ , to generate  $p_n$ , we simply resume the sieve up to the integer  $\lceil n(\ln n + \ln \ln n) \rceil$ , starting from  $\lfloor n(\ln n + \ln \ln n - 1) \rfloor$  or  $N_{\text{sieve}} + 1$ , whichever is greater – to make sure there is no overlap between two consecutive segments of the sieve. With this precaution in mind, the total complexity of the segmented sieve is the same as if we had directly performed the sieve in one go [6]; i.e.,  $O(N \log \log N)$  for a sieve performed up to integer  $N$ . Therefore, to generate the first  $n$  prime numbers, according to (1), the complexity of the sieve can be evaluated as  $O(n \cdot (\log n + \log \log n) \cdot \log \log (n \cdot (\log n + \log \log n)))$ . We refer the reader to [29] for practical considerations on the implementation of the segmented sieve of Eratosthenes.

**Remark 2.** Note that other sieve algorithms exist, with better complexities – such as Atkin sieve [2] or the wheel sieve [30]; the sieve of Eratosthenes has the merit of being the simplest to implement and sufficient for our needs. Also, a better asymptotic complexity but with a worse constant can be counterproductive for producing small primes – which is rather our case since we generate the primes in order.

### 3.2 Complexity analysis

There are several points to address when analysing Algorithm 2: (i) the complexity for testing whether or not  $f(N)$  is defined in line 10; (ii) the number of primes required by the algorithm, and the complexity for generating them; and (iii) the complexity of multiplication in line 9. To simplify the notations, let us assume that we run the algorithm with  $T_1 = T_2 = T$  – this is the worst case, since, if  $T_1 \neq T_2$ , we do not visit all the levels.


(i) The question of determining in  $O(1)$  whether  $f(N)$  is defined or not is not trivial in Algorithm 1 where  $f(\cdot)$  hashes multisets; however, for integers it is possible by virtue of [15].

Note also that [15] also provides a way to create a table, which associates with integer  $i$  the  $i$ -th prime number, searchable in  $O(1)$ . An application of this table is found in line 14: one can use pigeonhole sort to compare the two lists, with complexity  $O(\#T^d + p_n)$ , where  $p_n$  is the biggest prime in the list; but many holes will be unnecessary (as  $c(u)$  is necessarily prime). Using the table, one can use only  $n$  holes, one for each prime number, which reduces the complexity to  $O(\#T^d + n)$ . Since the primes are reallocated at each level, at level  $d$  we need as many primes as there are different equivalence classes at that level – i.e.  $\#\{c(u) : u \in T^d\}$ . This number is  $\leq \#T^d$ , therefore the complexity of the sort collapses to  $O(\#T^d)$ .

(ii) As already discussed, to generate the first  $n$  primes, the sieve must be carried out up to the integer  $N = n \cdot (\ln n + \ln \ln n)$ , for total complexity  $O(N \log \log N)$ .

The number of primes required by Algorithm 2 at level  $d$  is  $\#\{c(u) : u \in T^d\}$ . Thus, in total, the number of primes needed is exactly  $\max_{d \in \llbracket 0, \mathcal{D}(T) \rrbracket} \#\{c(u) : u \in T^d\}$ . We call this number the *width* of  $T$  and denote it by  $\mathcal{W}(T)$ . We have the following key result.

**Proposition 2.** For any tree  $T$ ,  $\mathcal{W}(T) \cdot (\ln \mathcal{W}(T) + \ln \ln \mathcal{W}(T)) = O(\#T)$ .

**Proof.** The proof can be found in Appendix A. 

It follows that generating the primes required for Algorithm 2 is done in  $O(\#T \log \log \#T)$ .



(iii) Let us denote  $M(n)$  the complexity for multiplying two  $n$ -bits numbers, where  $M(n)$  depends on the algorithm used:  $O(n^2)$  for usual schoolbook algorithm,  $O(n^{1.585})$  with Karatsuba algorithm [20], and  $O(n \cdot \log n \cdot \log \log n)$  with Schönhage–Strassen algorithm [33]. The fastest known algorithm was recently introduced [17], with complexity  $O(n \cdot \log n)$  – even if this result is, by the authors’ own admission, primarily theoretical. For our complexity proof, we assume that the algorithm used is Schönhage–Strassen; note that, in practice, for small integers, algorithms with a worse complexity but a better constant are used. Nonetheless, we assume in the sequel that  $M(n) = n \cdot \log n \cdot \log \log n$ , for the purpose of the proof.

Multiplying two  $n$ -bits numbers together yields a  $2n$ -bits number. Therefore, if we sequentially multiply  $m$  numbers of  $n$  bits together, the total complexity can be evaluated as  $M(n)+M(2n)+\dots+M((m-1)n)$ , which is  $O(m \cdot M(mn))$ . A better way is to adopt a divide and conquer approach and multiply two numbers which themselves are the recursive product of  $m/2$  numbers. This strategy leads to a complexity of  $O(M(mn))$  by virtue of the Master Theorem [7].

**Conclusion** Combining all the above discussions, we get the following result.

**Proposition 3.** *Algorithm 2 runs in*

$$O(\#T \cdot \log p_{\mathcal{W}(T)} \cdot \log(\deg(T) \cdot \log p_{\mathcal{W}(T)}) \cdot \log \log(\deg(T) \cdot \log p_{\mathcal{W}(T)}) + \#T \log \log \#T)$$

where  $p_{\mathcal{W}(T)}$  is the largest prime needed by the algorithm.

**Proof.** By the previous discussion in (ii), we consider separately the generation of primes, whose total complexity is  $O(\#T \log \log \#T)$ .

Now, fix  $d \in \llbracket 0, \mathcal{D}(T) \rrbracket$  and  $u \in T^d$ . Computing  $N(u)$  implies multiplying  $\deg(u)$  primes with at most  $\log p_{\mathcal{W}(T)}$  bits, which is, following (iii),  $O(M(\deg(u) \cdot \log p_{\mathcal{W}(T)}))$  – with  $M(n) = n \cdot \log n \cdot \log \log n$  as stated earlier. Lines 10 to 13 are  $O(1)$  from the discussion in (i). Sorting the lists in line 14 is  $O(\#T^d)$  – also from (i).

Processing level  $d$  thus requires

$$O(\#T^{d-1} \cdot \log p_{\mathcal{W}(T)} \cdot \log(\deg(T) \cdot \log p_{\mathcal{W}(T)}) \cdot \log \log(\deg(T) \cdot \log p_{\mathcal{W}(T)}) + \#T^d),$$

noticing that  $\sum_{u \in T^d} \deg(u) = \#T^{d-1}$  and bounding other occurrences of  $\deg(u)$  by  $\deg(T)$ . Summing over  $d$  leads to the claim. 

As already stated, the original AHU algorithm is linear only under the assumption that trees are not too large – recall Assumption 1. The term  $\log \log \#T$  coming from the generation of primes immediately vanishes. We now analyse the term

$$\log p_{\mathcal{W}(T)} \cdot \log(\deg(T) \cdot \log p_{\mathcal{W}(T)}) \cdot \log \log(\deg(T) \cdot \log p_{\mathcal{W}(T)}).$$

First, since  $\deg(T) < \#T$ , we have  $\log \deg(T) = O(1)$ . Using (1) we have

$$p_{\mathcal{W}(T)} < \mathcal{W}(T) (\ln \mathcal{W}(T) + \ln \ln \mathcal{W}(T)),$$

where  $\ln$  is the natural logarithm. Using Proposition 2, we have  $p_{\mathcal{W}(T)} = O(\#T)$ . It follows immediately that  $\log p_{\mathcal{W}(T)} = O(1)$ . The nested logarithms follows without difficulty.

Finally, we have proven the following result.

**Theorem 2.** *Algorithm 2 runs in  $O(n)$  with  $n = \#T_1 = \#T_2$ .*

## 4 Numerical experiments

We established that PRIMESAHU is equivalent in complexity (under the same assumption) to the original version. However, this theoretical result would be of little interest, especially with regard to the “ease of implementation” argument, if the constant were much larger, resulting in disproportionately long calculation times compared with the original algorithm. We show here that this is not the case, by comparing on random trees two Python implementations of the algorithm, where (i) PRIMESAHU turns out to be faster in the case when the trees are isomorphic, and (ii) of comparable time in the case when they are not.

Section 4.1 provides some insights about our implementation of the two algorithms, while Section 4.2 presents the results obtained.

### 4.1 Comments on the implementation

We implemented in Python the two algorithms discussed in this paper, the original AHU (oAHU for short) and PRIMESAHU (pAHU for short), making extensive use of the `treex` library [3], designed to manipulate trees.

We tried, as far as possible, to re-implement all the auxiliary functions used by the algorithms, to avoid an unfair advantage, linked, for example, to the use of highly optimised functions in Python. Typically, to multiply lists of prime numbers, we have implemented and used the recursive procedure described in Section 3.2-(ii), even though the `numpy.prod` function is faster when lists become sufficiently large. Note, however, that for multiplication operations  $*$ , we let Python choose the appropriate algorithm (schoolbook for small numbers and Karatsuba for large numbers), whereas we chose Schönhage–Strassen for the purposes of our proof of complexity in Section 3.2 – despite being slower in practice for small numbers.

Concerning perfect hash function for integers, we used Python dictionaries. Note also our implementation of the segmented sieve of Eratosthenes ignores multiples of 2 and 3, thus making the sieve 6 times faster.

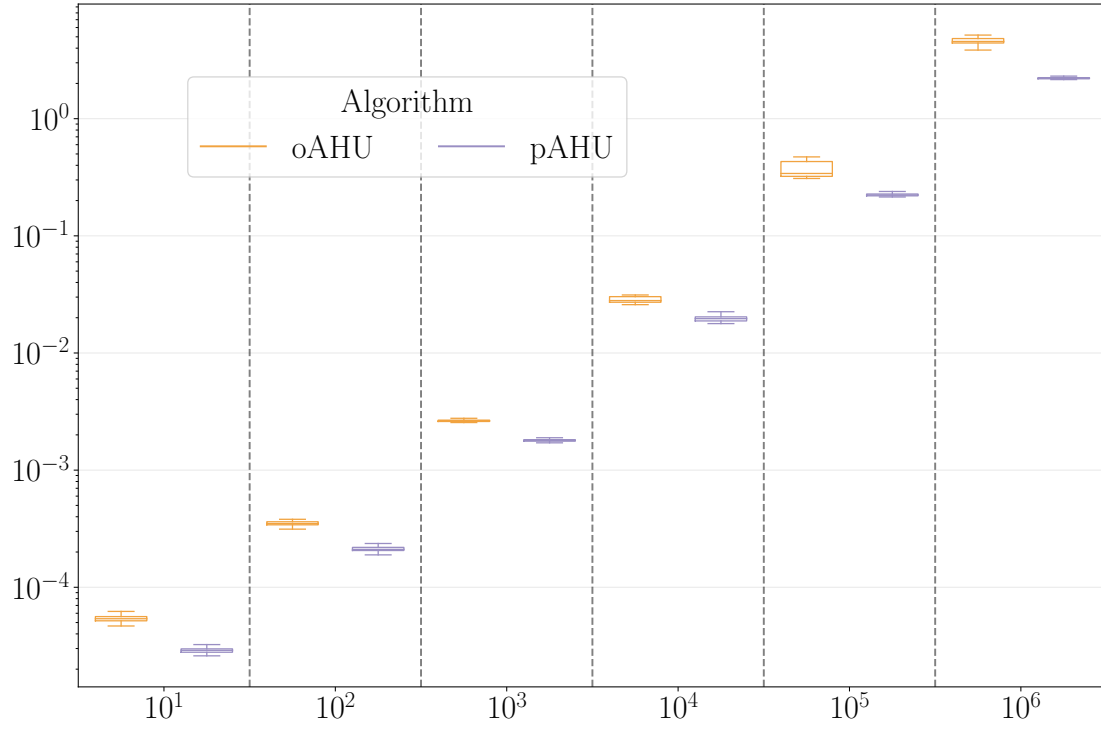
Experiments have been conducted on a HP Elite Notebook with 32 Go of RAM and Intel Core i7-1365U processor.

### 4.2 Results

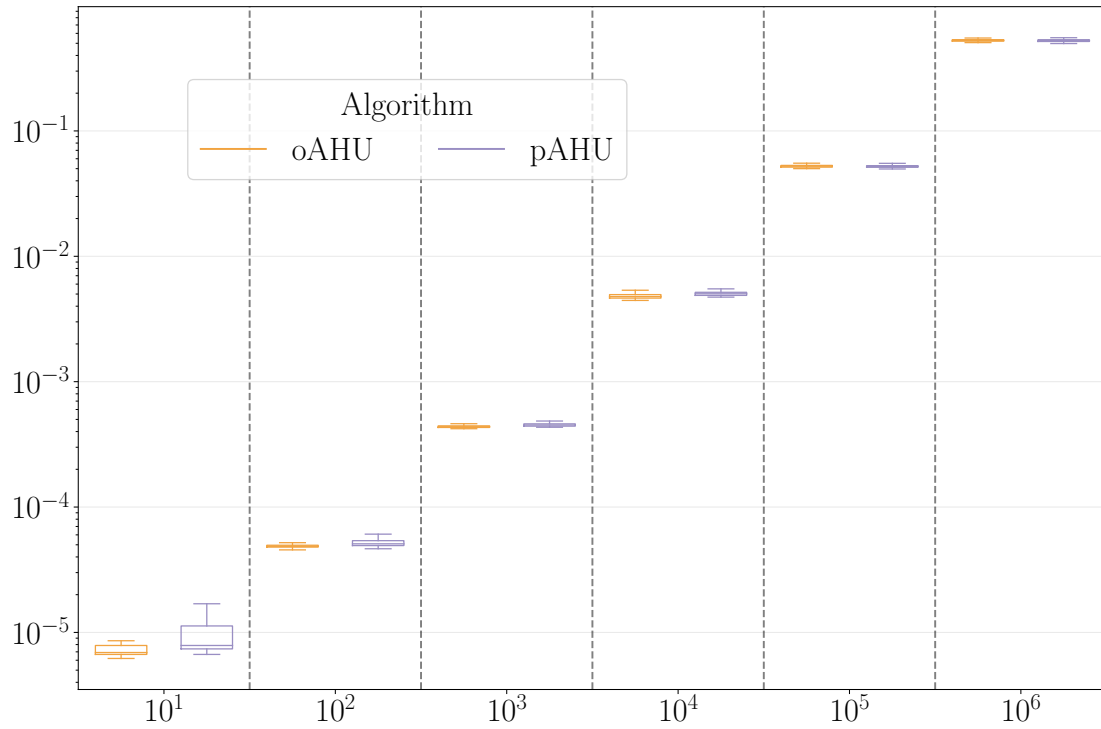
Provided an integer  $n$ , to treat the case  $T_1 \simeq T_2$ , we generate a random recursive tree  $T$  of size  $n$  [38], generate a copy  $T'$  of  $T$ , and measure the computation time taken for both algorithms – oAHU and pAHU – on the couple  $(T, T')$ . To treat the case  $T_1 \not\simeq T_2$ , we directly generate a pair  $(T_1, T_2)$  of random recursive trees of size  $n$  and measure the computation time for both algorithm.

We generated 100 couples for both cases and with  $n = 10^i$  for each  $i \in \llbracket 1, 6 \rrbracket$ . The results are depicted in Figure 3. As expected, both algorithms behave linearly; pAHU is faster in the case  $T_1 \simeq T_2$ , as can be seen in Figure 3a, whereas it achieves comparable time to oAHU in the case  $T_1 \not\simeq T_2$ , as soon as the trees are not too small – see Figure 3b. Note also that it is about 10 times faster to conclude  $\not\simeq$  than  $\simeq$  on the examples considered.

It is not our intention here to explain the differences in performance between the two algorithms – that is a topic requiring more detailed analysis – but suffice it to say that our algorithm, pAHU, is capable of performing just as well as oAHU in practice, which was our objective in this section.



(a)  $T_1 \approx T_2$  case



(b)  $T_1 \neq T_2$  case

Figure 3: Computation time (in seconds, log scale) for testing isomorphism of random recursive trees of sizes  $10^i$ ,  $i \in \{1, 2, 3, 4, 5, 6\}$ , using either oAHU or pAHU algorithm, with 100 couples of trees tested for each size and each case.

## Conclusion and perspectives

Following a remark by Campbell & Radford, who deplored the lack of clarity of the AHU algorithm (fundamental to understanding the tree isomorphism problem), we proposed a variant of this same algorithm: (i) with the same theoretical complexity; (ii) just as fast in practice, and (iii) intended to be simple to understand and implement. AHU works by sorting lists of integers, with the aim of computing a unique hash of multisets. We proposed instead to use an equally elementary concept, the multiplication of prime numbers, to also compute an invariant – thanks to the fundamental theorem of arithmetic.

We mentioned previously that the Weisfeiler-Lehman algorithms, used for graph isomorphism, use integer list sorts just like AHU to determine the next colour to assign to each node. This raises the question of whether these sorts can also be replaced by our idea of multiplying lists of prime numbers. While this issue is outside the scope of this paper, and remains to be investigated, let us nonetheless mention two points that may prove challenging.

First, the way Weisfeiler-Lehman operates can lead to processing as many colours as there are nodes in the graph, and therefore having to generate as many prime numbers – requiring the sieve of Eratosthenes to be run up to an integer supra-linear in the size of the graph.

Next, we would multiply lists whose size depends on the degree of the current node; in a dense or complete graph, this means lists whose size is comparable to the number of nodes in the graph. The complexity of performing these multiplications could prove far more expensive than for trees.

Since (1-dimensional) Weisfeiler-Lehman can be implemented in  $O((\#V + \#E) \log \#V)$  for a graph  $G = (V, E)$ , it remains to be investigated whether or not the additional complexities mentioned above exceeds that of the original algorithm. See [21, Section 3.1] and references therein for a more precise description of the Weisfeiler-Lehman algorithms.

## Acknowledgements

The author would like to thank Dr. Romain Azaïs and Dr. Jean Dupuy for their helpful suggestions on the first draft of the article.

## References

- [1] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *The Design and Analysis of Computer Algorithms*. Reading, Mass.: Addison-Wesley, 1974.
- [2] Arthur Atkin and Daniel Bernstein. “Prime sieves using binary quadratic forms”. In: *Mathematics of Computation* 73.246 (2004), pp. 1023–1030.
- [3] Romain Azaïs et al. “Treex: a Python package for manipulating rooted trees”. In: *Journal of Open Source Software* 4.38 (2019), p. 1351.
- [4] László Babai. “Graph isomorphism in quasipolynomial time”. In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 2016, pp. 684–697.
- [5] David A Barrington. “Bounded-width polynomial-size branching programs recognize exactly those languages in NC”. In: *Proceedings of the eighteenth annual ACM symposium on Theory of computing*. 1986, pp. 1–5.
- [6] Carter Bays and Richard H Hudson. “The segmented sieve of Eratosthenes and primes in arithmetic progressions to  $10^{12}$ ”. In: *BIT Numerical Mathematics* 17.2 (1977), pp. 121–127.
- [7] Jon Louis Bentley, Dorothea Haken, and James B Saxe. “A general method for solving divide-and-conquer recurrences”. In: *ACM SIGACT News* 12.3 (1980), pp. 36–44.

- [8] Paul E Black et al. “Dictionary of algorithms and data structures”. In: (1998).
- [9] Wayne D. Blizard et al. “Multiset theory”. In: *Notre Dame Journal of formal logic* 30.1 (1989), pp. 36–66.
- [10] Samuel R. Buss. “Alogtime algorithms for tree isomorphism, comparison, and canonization”. In: *Kurt Gödel Colloquium on Computational Logic and Proof Theory*. Springer. 1997, pp. 18–33.
- [11] Douglas M Campbell and David Radford. “Tree isomorphism algorithms: Speed vs. clarity”. In: *Mathematics Magazine* 64.4 (1991), pp. 252–261.
- [12] Dwaine Clarke et al. “Incremental multiset hash functions and their application to memory integrity checking”. In: *International conference on the theory and application of cryptology and information security*. Springer. 2003, pp. 188–207.
- [13] Pierre Dusart. “The  $k$ -th prime is greater than  $k(\ln k + \ln \ln k - 1)$  for  $k \geq 2$ ”. In: *Mathematics of computation* (1999), pp. 411–415.
- [14] Scott Fortin. “The graph isomorphism problem”. In: (1996).
- [15] Michael L Fredman, János Komlós, and Endre Szemerédi. “Storing a sparse table with 0 (1) worst case access time”. In: *Journal of the ACM (JACM)* 31.3 (1984), pp. 538–544.
- [16] Martin Grohe, Pascal Schweitzer, and Daniel Wiebking. “Deep weisfeiler leman”. In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 2600–2614.
- [17] David Harvey and Joris Van Der Hoeven. “Integer multiplication in time  $O(n \log n)$ ”. In: *Annals of Mathematics* 193.2 (2021), pp. 563–617.
- [18] Ningyuan Teresa Huang and Soledad Villar. “A Short Tutorial on The Weisfeiler-Lehman Test And Its Variants”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 8533–8537.
- [19] Florian Ingels and Romain Azaïs. “Isomorphic unordered labeled trees up to substitution ciphering”. In: *International Workshop on Combinatorial Algorithms*. Springer. 2021, pp. 385–399.
- [20] Anatolii Alekseevich Karatsuba and Yu P Ofman. “Multiplication of many-digital numbers by automatic computers”. In: *Doklady Akademii Nauk*. Vol. 145. 2. Russian Academy of Sciences. 1962, pp. 293–294.
- [21] Sandra Kiefer. “Power and limits of the Weisfeiler-Leman algorithm”. PhD thesis. Dissertation, RWTH Aachen University, 2020.
- [22] Donald E Knuth. *Art of Computer Programming, Volume 4, Fascicle 4, The: Generating All Trees–History of Combinatorial Generation*. Addison-Wesley Professional, 2013.
- [23] Donald E Knuth. “The Art of Computer Programming, Vol. 3: Sorting and Searching,” 1973”. In: *This book describes the binary search on (),* pp. 409–426.
- [24] Donald E Knuth. *The Art of Computer Programming: Fundamental Algorithms, volume 1*. Addison-Wesley Professional, 1997.
- [25] Pengyu Liu. “A tree distinguishing polynomial”. In: *Discrete Applied Mathematics* 288 (2021), pp. 1–8.
- [26] Jeremy Maitin-Shepard, Mehdi Tibouchi, and Diego F Aranha. “Elliptic curve multiset hash”. In: *The Computer Journal* 60.4 (2017), pp. 476–490.
- [27] Brendan D. McKay and Adolfo Piperno. “Practical graph isomorphism, II”. In: *Journal of symbolic computation* 60 (2014), pp. 94–112.
- [28] Shin-ichi Nakano and Takeaki Uno. “Efficient generation of rooted trees”. In: *National Institute for Informatics (Japan), Tech. Rep. NII-2003-005E* 8 (2003), pp. 4–63.
- [29] Melissa e O’neill. “The genuine sieve of Eratosthenes”. In: *Journal of Functional Programming* 19.1 (2009), pp. 95–106.
- [30] Paul Pritchard. “Linear prime-number sieves: A family tree”. In: *Science of computer programming* 9.1 (1987), pp. 17–35.

- [31] Barkley Rosser. “Explicit bounds for some functions of prime numbers”. In: *American Journal of Mathematics* 63.1 (1941), pp. 211–232.
- [32] Nitin Saxena. “Progress on Polynomial Identity Testing.” In: *Bull. EATCS* 99 (2009), pp. 49–79.
- [33] Arnold Schönhage and Volker Strassen. “Schnelle multiplikation grosser zahlen”. In: *Computing* 7.3 (1971), pp. 281–292.
- [34] Uwe Schöning. “Graph isomorphism is in the low hierarchy”. In: *Journal of Computer and System Sciences* 37.3 (1988), pp. 312–323.
- [35] Kevin Suffecool. *Cactus kev’s poker hand evaluator*. 2005.
- [36] Paul Tarau. “Towards a generic view of primality through multiset decompositions of natural numbers”. In: *Theoretical Computer Science* 537 (2014), pp. 105–124.
- [37] Gabriel Valiente. *Algorithms on trees and graphs*. Vol. 112. Springer, 2002.
- [38] Yazhe Zhang. “On the number of leaves in a random recursive tree”. In: (2015).

## A Proof of Proposition 2

Let  $k \geq 1$  be a fixed integer. A tree  $T$  such that  $\mathcal{W}(T) = k$  can be obtained by placing  $k$  trees  $T_i$ ,  $i \in \llbracket 1, k \rrbracket$ , under a common root, so that  $T_i \not\cong T_j$  for  $i \neq j$ . Note that this construction by no means encompasses all types of trees  $T$  with  $\mathcal{W}(T) = k$ . On the other hand, by cleverly choosing the  $T_i$ ’s, we can build a tree with the minimum number of nodes among all trees verifying  $\mathcal{W}(T) = k$ .

First,  $T_1$  would be the tree with a unique node. Then,  $T_2$  the only tree with two nodes. Then,  $T_3$  and  $T_4$  would be the two non-isomorphic trees with three nodes;  $T_5$  to  $T_8$  the four non-isomorphic trees with four nodes, and so on until we reach  $T_k$ . See Figure 4 for an example with  $k = 5$ . It should be clear that this construction ensures that  $\mathcal{W}(T) = k$  and  $\#T$  is minimal. The explicit generation of  $T_1, \dots, T_k$  can be carried following [28].

Following this construction, the total number of nodes in  $T$ , that we denote by  $t_k$ , is therefore closely related to the number of non-isomorphic trees and their cumulative sum. Let us denote  $a_n$  the number of non-isomorphic trees of size  $n$ , and  $b_n$  the number of non-isomorphic trees of size at most  $n$  – so that  $b_n = \sum_{i=1}^n a_i$ . Let  $n$  be the integer so that  $b_n < k \leq b_{n+1}$ . All trees with size up to  $n - 1$  are used in the construction, as well as  $k - b_n$  trees of size  $n + 1$  (no matter which ones). Therefore,

$$t_k = 1 + \sum_{i=1}^n i \cdot a_i + (n + 1)(k - b_n).$$

Table 1 provides the first values for  $a_n$ ,  $b_n$  and  $t_k$ .

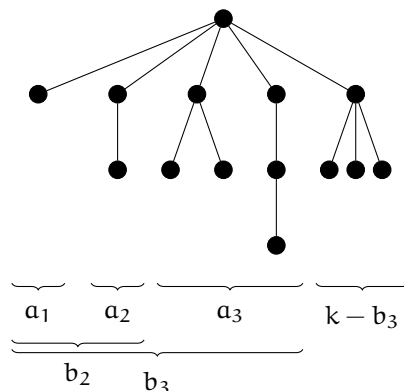


Figure 4: The smallest tree so that  $\mathcal{W}(T) = 5$ . We have  $b_3 < 5 \leq b_4$  and  $\#T = 1 + 1 \cdot a_1 + 2 \cdot a_2 + 3 \cdot a_3 + 4 \cdot (5 - b_3) = 14$ .

$n$	1	2	3	4	5	6	7	8
$a_n$	1	1	2	4	9	20	48	115
$b_n$	1	2	4	8	17	37	85	200
$t_n$	2	4	7	10	14	18	22	26

Table 1: First values of  $a_n$ ,  $b_n$  and  $t_n$ .  $a_n$  is sequence [A000081](#) in OEIS, and  $b_n$  sequence [A087803](#). See OEIS Foundation Inc. (2023), The On-Line Encyclopedia of Integer Sequences, Published electronically at <https://oeis.org>.

Following the previous discussion, we have the following result.

**Lemma 1.** For any tree  $T$ ,  $\#T \geq t_{\mathcal{W}(T)}$ .

Our goal here is to prove the following proposition. It should be clear that combining upcoming Proposition 4 and Lemma 1 yields Proposition 2.


**Proposition 4.** For any  $k \geq 1$ ,  $k \cdot (\ln k + \ln \ln k) = O(t_k)$ .

**A few asymptotic equivalents** Before starting the proof of Proposition 4, we require some preliminary results. We start by the following technical lemma.

**Lemma 2.** Let  $(u_n)_{n \in \mathbb{N}}$  be a sequence so that  $u_n \underset{+\infty}{\sim} c \cdot d^n \cdot n^{-\alpha}$ , with  $c, \alpha \geq 0$  and  $d > 1$ . Then  $\sum_{k=1}^n u_k \underset{+\infty}{\sim} c \cdot \frac{d}{d-1} \cdot d^n \cdot n^{-\alpha}$ .

**Proof.** Obviously the sequence  $\sum d^n \cdot n^{-\alpha}$  diverges, and therefore we have  $\sum_{k=1}^n u_k \underset{+\infty}{\sim} c \sum_{k=1}^n d^k \cdot k^{-\alpha}$ . Then,

$$\frac{c \sum_{k=1}^n d^k \cdot k^{-\alpha}}{c \cdot \frac{d}{d-1} \cdot d^n \cdot n^{-\alpha}} = \frac{d-1}{d} \sum_{k=1}^n d^{k-n} \left(\frac{k}{n}\right)^{-\alpha} \underset{j=n-k}{=} \frac{d-1}{d} \sum_{j=1}^n \left(1 - \frac{j}{n}\right)^{-\alpha} d^{-j}.$$

With bounds  $1 \leq \left(1 - \frac{j}{n}\right)^{-\alpha} \leq \left(1 - \frac{1}{n}\right)^{-\alpha}$ , it is easy to see that the right-hand term goes to 1 as  $n \rightarrow \infty$ . 

From [24, Section 2.3.4.4], we have  $a_n \underset{+\infty}{\sim} c \cdot d^n \cdot n^{-3/2}$  with  $c \approx 0.439924$  and  $d \approx 2.955765$ . From Lemma 2, we immediately derive  $b_n \underset{+\infty}{\sim} c \cdot \frac{d}{d-1} \cdot d^n \cdot n^{-3/2}$ . Finally, noticing that  $i \cdot a_i \underset{+\infty}{\sim} c \cdot d^i \cdot i^{-1/2}$ , we derive from Lemma 2 that  $\sum_{i=1}^n i \cdot a_i \underset{+\infty}{\sim} c \cdot \frac{d}{d-1} \cdot d^n \cdot n^{-1/2}$ .

**Proof of Proposition 4** Note that, by definition  $k \leq b_{n+1}$ , and  $t_k \geq \sum_{i=1}^n i \cdot a_i$ ; so it suffices to prove that  $b_{n+1} \cdot (\ln b_{n+1} + \ln \ln b_{n+1}) = O\left(\sum_{i=1}^n i \cdot a_i\right)$ . Having established this, some straightforward calculations allow to derive

$$b_{n+1} \cdot (\ln b_{n+1} + \ln \ln b_{n+1}) \underset{+\infty}{\sim} c \cdot \frac{d}{d-1} \cdot d^{n+1} \cdot (n+1)^{-1/2} \cdot \ln d.$$

Since  $\sum_{i=1}^n i \cdot a_i \underset{+\infty}{\sim} c \cdot \frac{d}{d-1} \cdot d^n \cdot n^{-1/2}$ , we can immediately conclude that

$$\frac{b_{n+1} \cdot (\ln b_{n+1} + \ln \ln b_{n+1})}{\sum_{i=1}^n i \cdot a_i} \xrightarrow{n \rightarrow \infty} d \ln d,$$

which completes the proof.