



**HAL**  
open science

## Enumeration of Unordered Forests

Florian Ingels, Romain Azaïs

► **To cite this version:**

| Florian Ingels, Romain Azaïs. Enumeration of Unordered Forests. 2021. hal-02511901v2

**HAL Id: hal-02511901**

**<https://hal.science/hal-02511901v2>**

Preprint submitted on 16 Mar 2021 (v2), last revised 13 Apr 2022 (v4)

**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.

# ENUMERATION OF UNORDERED FORESTS

Florian Ingels  
florian.ingels@inria.fr

Romain Azaïs  
romain.azais@inria.fr

Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS,  
INRAE, Inria, F-69342, Lyon, France

## Abstract

Reverse search is a convenient method for enumerating structured objects, that can be used both to address theoretical issues and to solve data mining problems. This method has already been successfully developed to handle unordered trees. If the literature proposes solutions to enumerate singletons of trees, we study in this article a more general, higher combinatorial problem, the enumeration of sets of trees – forests. By compressing each forest into a Directed Acyclic Graph (DAG), we develop a reverse search like method to enumerate DAG compressing forests. Remarkably, we prove that these DAG are in bijection with the row-Fishburn matrices, a well-studied class of combinatorial objects. In a second step, we derive our forest enumeration to provide algorithms for tackling two related problems : (i) the enumeration of “subforests” of a forest, and (ii) the frequent “subforest” mining problem. All the methods presented in this article enumerate each item uniquely, up to isomorphism.

**keywords:** Directed Acyclic Graph, Reverse Search, Unordered Trees, Enumeration, Forest

## 1 Introduction

### 1.1 Context of the work

Enumeration of trees is a long-term problem, where Cayley was the first to propose a formula for counting unordered trees in the mid-19th century [9, I.5.2]. The exhaustive enumeration of ordered and unordered trees was successfully tackled in the early 00’s by Nakano and Uno in [17, 18]. In the unordered case, an extension of the algorithm has been proposed to solve the problem of frequent substructure mining [1]. Moreover, in the field of machine learning, we have recently demonstrated that exhaustive enumeration of the subtrees of a tree makes it possible to design classification algorithms significantly more efficient than their counterpart without such enumeration [5].

Our ambition in this article is to take these two problems of enumeration – trees and subtrees – to a higher order, i.e. to enumerate sets of trees instead of singletons. Specifically, we call a forest a set of trees that contains no repetition – in the sense that no tree is a subtree of another. This condition of non-repetition is in line with a parsimonious enumeration approach, where the objects considered are all different and enumerated up to isomorphism. Besides, this condition is not restrictive since one can always introduce repetition afterwards. We are therefore interested in the problem of enumerating forests of unordered trees, and then, given a tree or forest, to enumerate all its “subforests” – as forests of subtrees. The latter has already been discussed in the literature,

but without consideration on isomorphism [20]. We re-emphasize that we aim to enumerate these various items – forests and subforests – up to isomorphism.

Such an ambition immediately raises a number of obstacles. First of all, the trees are indeed unordered, but so are the sets of trees. For the former the literature has introduced the notion of the canonical form of a tree [18, 1], which is a unique ordered representation of an unordered tree. The enumeration therefore focuses only on these canonical trees. Unfortunately, if it is possible to order a set of vertices, there is no total order on the set of trees, to the best of our knowledge. In addition, forests belong to the powerset of trees, that have an exponentially stronger combinatorics; the condition of non-repetition filters the powerset of trees in a non-trivial way, making, a priori, the enumeration problem trickier.

Enumeration problems are recurrent in many fields, notably combinatorial optimization and data mining. They involve the exhaustive listing of a subset of the elements of a search set (possibly all of them), e.g. graphs, trees or vertices of a simplex. Given the possibly high combinatorial nature of these elements, it is essential to adopt clever exploration strategies as opposed to brute-force enumeration, typically to avoid areas of the search set not belonging to the objective subset.

One proven way of proceeding is to provide the search set with an enumeration tree structure; starting from the root, the branches of the tree are explored recursively, eliminating those that do not address the problem. Based on this principle, we can notably mention the well-known “branch and bound” method in combinatorial optimization [16] and the gSpan algorithm for frequent subgraph mining in data mining [24]. Another of these methods is the so-called reverse search technique, which requires that the search set has a partial order structure, and which has solved a large number of enumeration problems since its introduction [2] until very recently [23]. Actually, the algorithms previously introduced in the literature to enumerate trees are based on this technique [17, 18, 1].

In the present paper, we restrict ourselves to reverse search methods, for which the following formalism is adapted from the one that can be found in [19, p. 45-51], and slightly differ from the original definition by Avis and Fukuda [2]. We refer the reader to these two references for further details.

Let  $(\mathcal{S}, \subseteq)$  be a partially ordered set, and  $g : \mathcal{S} \rightarrow \{\top, \perp\}$  be a *property*, satisfying anti-monotonicity

$$\forall s, t \in \mathcal{S} : (s \subseteq t) \wedge g(t) = \top \implies g(s) = \top.$$

The *enumeration problem* for the property  $g$  is the problem of listing all elements of  $E_{\mathcal{S}}(g) = \{s \in \mathcal{S} : g(s) = \top\}$ . An enumeration algorithm is an algorithm that returns  $E_{\mathcal{S}}(g)$ .

The reverse search technique relies on inverting a *reduction rule*  $f : \mathcal{S} \setminus \emptyset \rightarrow \mathcal{S}$ , where  $f$  satisfies the two properties of (i) *covering*:  $\forall s \in \mathcal{S} \setminus \emptyset, f(s) \subset s$  and (ii) *finiteness*:  $\forall s \in \mathcal{S} \setminus \emptyset, \exists k \in \mathbb{N}^*, f^k(s) = \emptyset$ . Then, the *expansion rule* is defined as  $f^{-1}(t) = \{s \in \mathcal{S} : f(s) = t\}$ . This defines an enumeration tree rooted in  $\emptyset$ , and repeated call to  $f^{-1}$  can therefore enumerates all the elements of  $\mathcal{S}$ .

The reverse search algorithm is shown in Algorithm 1.  $E_{\mathcal{S}}(g)$  can be obtained from the call of REVERSESEARCH( $(\mathcal{S}, \subseteq), f^{-1}, g, \emptyset$ ). As  $g$  is anti-monotone, if  $g(s) = \perp$ , then all elements  $s \subseteq t$  also have  $g(t) = \perp$ , and therefore pruning the enumeration tree in  $s$  does not miss any element of  $E_{\mathcal{S}}(g)$ .

---

**Algorithm 1:** REVERSESEARCH

---

**Input:**  $(\mathcal{S}, \subseteq), f^{-1}, g, s_0 \in \mathcal{S} - \text{s.t. } g(s_0) = \top$   
**1 output**  $s_0$   
**2 for**  $t \in \{s \in f^{-1}(s_0) | g(s) = \top\}$  **do**  
**3**      $\lfloor$  REVERSESEARCH( $(\mathcal{S}, \subseteq), f^{-1}, g, t$ )

---

## 1.2 Precise formulation of the problem

A rooted tree  $T$  is a connected graph with no cycle such that there exists a unique vertex called the root, which has no parent, and any vertex different from the root has exactly one parent. Rooted trees are said unordered if the order between the sibling vertices of any vertex is not significant. As such, the set of children of a vertex  $v$  is considered as a multiset and denoted by  $\mathcal{C}(v)$ . The leaves  $\mathcal{L}(T)$  are all the vertices without children. The height of a vertex  $v$  of a tree  $T$  can be recursively defined as

$$\mathcal{H}(v) = \begin{cases} 0 & \text{if } v \in \mathcal{L}(T), \\ 1 + \max_{u \in \mathcal{C}(v)} \mathcal{H}(u) & \text{otherwise.} \end{cases} \quad (1)$$

The height  $\mathcal{H}(T)$  of the tree  $T$  is defined as the height of its root. The outdegree of a vertex  $v \in T$  is defined as  $\deg(v) = \#\mathcal{C}(v)$ ; the outdegree of  $T$  is then defined as  $\deg(T) = \max_{v \in T} \deg(v)$ . The depth of a vertex  $v$  is the number of edges on the path from  $v$  to the root of the tree.

Two trees  $T_1$  and  $T_2$  are isomorphic if there exists a one-to-one correspondance  $\phi$  between the vertices of the trees such that (i)  $u \in \mathcal{C}(v)$  in  $T_1 \iff \phi(u) \in \mathcal{C}(\phi(v))$  in  $T_2$  and (ii) the roots are mapped together. For any vertex  $v$  of  $T$ , the subtree  $T[v]$  rooted in  $v$  is the tree composed of  $v$  and all its descendants – denoted by  $\mathcal{D}(v)$ .  $\mathcal{S}(T)$  denotes the set of all distinct subtrees of  $T$ , which is the quotient set of  $\{T[v] : v \in T\}$  by the tree isomorphism relation. In this article, we consider only unordered rooted trees that will simply be called trees in the sequel. We denote by  $\mathcal{T}$  the set of all trees.

As mentioned before, we are interested in this paper in the enumeration of *forests*. The literature acknowledges two definitions for a forest [6, p. 172]: (i) an undirected graph in which any two vertices are connected by at most one path or (ii) a disjoint union of trees. We adopt a variation of the latter one, that forbids repetitions inside the forest.

**Definition 1.1.** *A set  $\{T_1, \dots, T_n\}$  of trees is a forest if and only if*

$$\forall i \neq j, T_i \notin \mathcal{S}(T_j). \quad (2)$$

We denote by  $\mathcal{F}$  the set of all forests. Our goal is therefore to provide a reverse search like method that outputs  $\mathcal{F}$ . As already stated, this goal raises two major difficulties: firstly, the twofold unordered nature of forests (the set of trees and the trees themselves), and secondly, the higher combinatorics of forests compared to trees combined with the non-trivial condition of non-repetition. While the latter problem is intrinsic, the main idea of this paper to address the former is to resort to the reduction of a forest into a Directed Acyclic Graph (DAG).

DAG reduction is a method meant to eliminate internal repetitions in the structure of trees and forests of trees. Beginning with [21], DAG representations of trees are also much used in computer graphics where the process of condensing a tree into a graph is called object instancing [12]. A precise definition of DAG reduction of trees, together with algorithms to compute it, are provided in [10], whereas one technique to extend those algorithms to forests is presented in [5, Section 3.2]. DAG reduction can be interpreted as the construction of the quotient graph of a forest by the tree isomorphism relation. However, in this paper, we provide the general idea of DAG reduction as a vertex coloring procedure.

Consider a forest  $F = \{T_1, \dots, T_n\}$  to reduce. Each vertex of each tree is given a color such that if two distinct vertices  $u, v$  belonging respectively to  $T_i, T_j$  (not necessarily distinct) have the same color, then  $T_i[u]$  and  $T_j[v]$  are isomorphic. Reciprocally, if two subtrees are isomorphic, their roots have to be identically colored. Let us denote  $c(\cdot)$  the function that associates a color to any vertex. Then, we build a directed graph  $D = (V, A)$  with as many vertices as colors used, i.e.  $\#V = \#\text{Im}(c)$ . For any two vertices  $u, v$  in the forest, if  $u \in \mathcal{C}(v)$ , then we create an arc  $c(v) \rightarrow c(u)$  in  $D$ . Note that this definition implies that multiples arcs are possible in  $D$ , as if there exist  $u, u' \in \mathcal{C}(v)$ , for  $v \in T$ , such that  $T[u]$  and  $T[u']$  are isomorphic, then the arcs  $c(v) \rightarrow c(u)$  and  $c(v) \rightarrow c(u')$  are identical. The graph  $D$  is a DAG [10, Proposition 1], i.e. a connected directed graph without cycles. We refer to Figure 1 for an example of DAG reduction.

In this paper,  $\mathfrak{R}(F)$  denotes the DAG reduction of  $F$ . It is crucial to notice that the function  $\mathfrak{R}$  is a one-to-one correspondence [10, Proposition 4], which means that DAG reduction is a lossless compression algorithm. Since  $F$  fulfills condition (2), no tree of  $F$  is a subtree of another. If this were the case, say  $T_i \in \mathcal{S}(T_j)$ , then  $\mathfrak{R}(T_i)$  would be a subDAG of  $\mathfrak{R}(T_j)$ , and therefore the numbers of roots in  $\mathfrak{R}(F)$  would be strictly less than  $\#F$ . Since such a situation can not occur, there are exactly as many roots in  $\mathfrak{R}(F)$  as there are elements in  $F$ : no information is lost. In other words,  $F$  can be reconstructed from  $\mathfrak{R}(F)$  and  $\mathfrak{R}^{-1}$  stands for the inverse function.

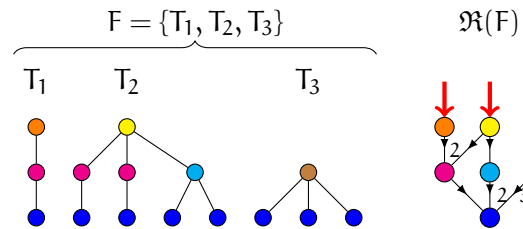


Figure 1: A forest  $F$  (left) and its DAG reduction (right). Roots of isomorphic subtrees are identically colored, as well as the corresponding vertex of the DAG. Red arrows indicate on the DAG the roots of the trees of  $F$ . Arcs with multiplicity greater than one are displayed with their multiplicity number.

The DAG structure inherits of some properties of trees. For a vertex  $v$  in a DAG  $D$ , we will denote by  $\mathcal{C}(v)$  the set of children of  $v$ .  $\mathcal{H}(v)$  and  $\text{deg}(v)$  are inherited as well. Similarly to trees, we denote by  $D[v]$  the subDAG rooted in  $v$  composed of  $v$  and all its descendants  $\mathcal{D}(v)$ . Note that since  $D[v]$  has a unique root  $v$ , it compresses a forest made of a single tree. For the simplicity of notation, we use  $\mathfrak{R}^{-1}(D[v])$  to designate the tree compressed by  $D[v]$  – instead of the singleton.

In the sequel, DAG compressing forests are called FDAG, to distinguish them from general directed acyclic graphs.

Since DAG compression is lossless, and since a forest can be reconstructed from its DAG reduction, it should be clear that enumerating all forests is equivalent to enumerating all FDAG. Yet, the latter approach has the merit of transforming set of trees into unique objects, which makes it possible, if able to design a canonical representation – like the trees in [18, 1], to get rid of the twofold unordered nature of forests, as claimed earlier. Indeed, any ordering of the vertices of the DAG induces an order on the roots of the DAG, and therefore on the elements of the forest, as well on the vertices of the trees themselves.

### 1.3 Aim of the paper

To the best of our knowledge, the enumeration of DAG has never been considered in the literature. The aim of this article is twofold, i.e (i) to open the way by presenting a reverse search algorithm

enumerating FDAG, in Section 2, and (ii) to derive from it an algorithm for enumerating substructures in Section 3. The frequent pattern mining problem is a classical data mining problem – see [11] for a survey on that question – and we provide in Section 4 a slight variation of the algorithm of Section 3 to tackle this issue. In more details, our outline is as follows:

- The first step is to introduce a canonical form for FDAG. For trees [18, Section 3], this consisted in associating an integer (its depth) to each vertex, and maximizing the sequence by choosing an appropriate ordering over the vertices. The notion of depth does not apply to FDAG, which forces us to find another strategy. DAG are characterized by the existence of a topological ordering [15], and we introduce in Subsection 2.2 a topological ordering that is unique if and only if a DAG compresses a forest. This canonical ordering is defined so that the sequence of children of the vertices is strictly increasing, where the multisets of children are ordered by the lexicographical order. In fact, these ordered multisets of children are considered as formal words, which brings us to a detour through the theory of formal languages in Subsection 2.1 to introduce useful results for the rest of the article. Compared to trees, we have here a first gain in complexity insofar as we maximize a sequence of words instead of a sequence of integers.
- The expansion rule used for trees [18, Section 4] is to add a new vertex in the tree as a child of some other vertex, so that the depth-sequence remains maximal. Consequently, a single arc is also added. On the other hand, for a FDAG, we want to be able to add either vertices or arcs independently. In Subsection 2.3, we define three expansion rules, reflecting the full spectrum of possible operations, so that the DAG obtained afterward is still a FDAG. Specifically, the branching rule allows to add an arc, where the elongation and widening rules add vertices at different height. We show in Proposition 2.10 that the rules preserve the canonicalness and in Proposition 2.11 that they are “bijective”: any FDAG can be reached by applying the expansion rules to a unique FDAG. In Subsection 2.4 we derive from them an enumeration tree covering the set of FDAG.
- Notably, a bijection between FDAG and row-Fishburn matrices, a class of combinatorial objects much exploited in the literature [13, Section 2], is shown in Theorem 2.13 – which proof lies in Appendix A. The asymptotic behavior of these matrices being well known [14, 7], this allows us to derive from it the behavior of the enumeration tree. In return, since our bijection is constructive, the enumeration tree can be used to enumerate row-Fishburn matrices – and all the objects they are in bijection with – via the reverse search method. Remarkably, this bijection operates between two objects that, at first sight, have little in common.
- It has been mentioned above that the set of FDAG belongs to the powerset of  $\mathcal{T}$ . For an enumeration algorithm to have any practical interest, it is necessary that the associated enumeration tree has a “reasonable” growth – with regard to the combinatorics of the explored space. This is the case for our algorithm since we prove, in Subsection 2.5, that a FDAG with  $n$  vertices has a number of successors in the enumeration tree in the order of  $\Theta(n)$  – and that those successors can be computed in quasi-quadratic time in the worst case.
- Finally, Subsection 2.6 concludes on enumeration by proposing sets of constraints that make the enumeration tree finite. Indeed, since the rules only allow to increase the height, degree

or number of vertices, it is sufficient to set maximum values for some of these parameters to achieve this goal; however the combination of parameters has to be wisely chosen, as we show it.

- Since the structures we enumerate are forests, it is natural that the substructures we are interested in are “subforests”. A precise definition of the latter is given in Section 3, i.e. forests of subtrees, and are referred to as subFDAG. An algorithm to enumerate all subFDAG appearing in a FDAG is also provided. The frequent subFDAG mining problem is finally addressed in Section 4.

Concluding remarks concerning the implementation of our results in the Python library `treex` [3] are briefly mentioned at the end of the article. In Appendix B, the interested reader will find an index of frequent notations used throughout the paper.

## 2 Exhaustive enumeration of FDAG

In this section, we introduce our main result, that is, a reverse search algorithm for the enumeration of FDAG. As we will consider the multisets of children of vertices as formal words built on the alphabet formed by the set of vertices, we introduce in Subsection 2.1 some definitions and results on formal languages that will be useful for the sequel. We characterize unambiguously in Subsection 2.2 our objects of study, through the lens of topological orderings, defining a canonical topological ordering for DAG, that is unique if and only if a DAG compresses a forest of unordered trees, i.e. it is a FDAG – see Theorem 2.4. We then define three expansion rules that are meant to extend the structures of FDAG in Subsection 2.3. In Subsection 2.4, we show with Theorem 2.12 that these expansion rules define an enumeration tree on the set of FDAG; the growth of this tree is then analysed in Subsection 2.5. Notably, we exhibit a bijection – Theorem 2.13 – with a class of combinatorial objects from the literature, allowing us to obtain an asymptotic expansion of the growth of the tree. In addition, we show that any FDAG has a linear number of children in that tree in Theorem 2.15, and that the time complexity to construct those children has quadratic magnitude in the worst case – see Proposition 2.16. Finally, options to constrain the enumeration tree – on maximum number of vertices, height or outdegree – and making it finite are proposed in Subsection 2.6.

### 2.1 Preliminary: a detour through formal languages

We present in this subsection some definitions and results on formal languages that will be useful for the sequel of Section 2.

Let  $\mathcal{A}$  be a totally ordered finite set, called alphabet, whose elements are called letters. A word is a finite sequence of letters of  $\mathcal{A}$ . The length of a word  $w$  is equal to its number of letters and is denoted by  $\#w$ . There is a unique word with no letter called the empty word and denoted by  $\epsilon$ . The set of all words is denoted by  $\mathcal{A}^*$ . Words can be concatenated to create a new word whose length is the sum of the lengths of the original words;  $\epsilon$  is the neutral element of this concatenation operation.

The lexicographical order over  $\mathcal{A}^*$ , denoted by  $<_{\text{lex}}$ , is defined as follows. Let  $w_1 = a_0 \cdots a_p$  and  $w_2 = b_0 \cdots b_q$  be two words, with  $a_i, b_j \in \mathcal{A}$ . If  $\#w_1 = \#w_2$ , then  $w_1 <_{\text{lex}} w_2$  if and only if



$\exists k \in \llbracket 0, p \rrbracket, a_i = b_i \forall i < k$  and  $a_k < b_k$ . Otherwise, let  $m = \min(p, q)$ ;  $w_1 <_{\text{lex.}} w_2$  if and only if either (i)  $a_0 \cdots a_m <_{\text{lex.}} b_0 \cdots b_m$  or (ii)  $a_0 \cdots a_m =_{\text{lex.}} b_0 \cdots b_m$  and  $m < q$  – that is,  $p < q$ . Note that, by convention,  $\epsilon <_{\text{lex.}} w$  for any word  $w$ .

Let  $w \in \mathcal{A}^*$ . We define the suffix-cut operator  $SC(w)$ , which removes the last letter of  $w$ :

$$SC(w) = \begin{cases} w' & \text{if } w = w'a \text{ with } a \in \mathcal{A} \text{ and } w' \in \mathcal{A}^*, \\ \epsilon & \text{otherwise.} \end{cases} \quad (3)$$

A *language* is a set of words satisfying some construction rules. We introduce hereafter two languages that will be useful in the sequel of the paper.

**Definition 2.1.** *The language of decreasing words is defined as*

$$\Lambda = \{w = a_0 \cdots a_m \in \mathcal{A}^* : a_i \geq_{\text{lex.}} a_{i+1} \forall i \in \llbracket 0, m-1 \rrbracket\}.$$

**Definition 2.2.** *Let  $\bar{w} \in \Lambda$ . The language of decreasing words bounded by  $\bar{w}$  is defined as*

$$\Lambda^{\bar{w}} = \{w \in \Lambda : w >_{\text{lex.}} \bar{w}\}.$$

*Any word  $w \in \Lambda^{\bar{w}}$  is said to be minimal if and only if  $w \in \Lambda^{\bar{w}}$  but  $SC(w) \notin \Lambda^{\bar{w}}$ .*

As an example, if  $\mathcal{A} = \{0, 1, 2, 3\}$ , then  $\bar{w} = 211 \in \Lambda$ , whereas  $121 \notin \Lambda$ . In addition,  $\Lambda^{\bar{w}}$  contains words such as 31, 22, 21110, etc. 22 is a minimal word of  $\Lambda^{\bar{w}}$  as  $22 >_{\text{lex.}} 211$  but  $SC(22) = 2 <_{\text{lex.}} 211$ .

Our focus is now on the construction of the minimal words of  $\Lambda^{\bar{w}}$ . Let  $\bar{w} = a_0 \cdots a_p$  and  $w = b_0 \cdots b_q \in \Lambda^{\bar{w}}$ . Taking into account that  $w >_{\text{lex.}} \bar{w}$  and that they both are decreasing words, there are only two possibles cases:

- (i)  $w$  and  $\bar{w}$  share a common prefix  $a_0 \cdots a_m$ . Then  $w = a_0 \cdots a_m b_{m+1} \cdots b_q$ , and the word  $a_0 \cdots a_m b_{m+1}$  is minimal by applying successive suffix-cut operations.
- (ii)  $w$  and  $\bar{w}$  do not share a common prefix. Necessarily  $b_0 >_{\text{lex.}} a_0$ , and then the word  $b_0$  is minimal by applying several suffix-cut operations.

From the above, we deduce a method for constructing all minimal word of  $\Lambda^{\bar{w}}$ . First, we partition  $\mathcal{A}$  into disjoint – potentially empty – subsets:

$$\begin{aligned} \mathcal{A}^0 &= \{a \in \mathcal{A} : a >_{\text{lex.}} a_0\}, \\ \mathcal{A}^i &= \{a \in \mathcal{A} : a_{i-1} \geq_{\text{lex.}} a >_{\text{lex.}} a_i\} \quad 1 \leq i \leq p, \\ \mathcal{A}^{p+1} &= \{a \in \mathcal{A} : a_p \geq_{\text{lex.}} a\}. \end{aligned}$$

It then follows that – empty  $\mathcal{A}^i$ 's not being considered,

- $\forall b \in \mathcal{A}^0$ , the word  $b$  is minimal,
- $\forall b \in \mathcal{A}^i$  with  $i \in \{1, \dots, p\}$ , the word  $a_0 \cdots a_{i-1} b$  is minimal,
- $\forall b \in \mathcal{A}^{p+1}$ , the word  $\bar{w}b$  is minimal.



As we partitioned  $\mathcal{A}$ , we have proved the following proposition.

**Proposition 2.3.** *The number of minimal words of  $\Lambda^{\bar{w}}$  is exactly  $\#\mathcal{A}$ .*

As a follow-up of the example some lines ago, with  $\mathcal{A} = \{0, 1, 2, 3\}$  and  $\bar{w} = 211$ , we apply the proposed method to find the minimal elements of  $\Lambda^{\bar{w}}$ . We partition  $\mathcal{A}$  into:  $\mathcal{A}^0 = \{3\}$ ,  $\mathcal{A}^1 = \{2\}$ ,  $\mathcal{A}^2 = \emptyset$ ,  $\mathcal{A}^3 = \{0, 1\}$ . The four minimal words are therefore 3, 22, 2111 and 2110.

Although the previous result is completely general, if we require that  $\mathcal{A} = \{0, \dots, n\}$ , then the partition method described above can be rewritten into Algorithm 2. While this is not included in the pseudocode provided, note that the algorithm should return an empty list if  $a_0 > n$ , as in this case there would be no minimal word to look for.

## 2.2 Canonical FDAG

FDAG are unordered objects, like the trees they compress, and therefore their enumeration requires to reflect this nature. In practice, finding a systematic way to order them makes it possible to design a simpler reduction rule, as done for trees [18], ignoring the combinatorics of permutations. The purpose of this subsection is to provide a unique way to order FDAG. We show that such an order exists in Theorem 2.4, unambiguously characterizing FDAG. The approach chosen is based on the notion of topological order.

**Topological ordering** Let  $D$  be a directed graph, where multiple arcs are allowed. A topological ordering on  $D$  is an ordering of the vertices of  $D$  such that for every arc  $uv$  from vertex  $u$  to vertex  $v$ ,  $u$  comes after  $v$  in the ordering. Formally,  $\psi : D \rightarrow \llbracket 0, \#D - 1 \rrbracket$  is a topological ordering if and only if  $\psi$  is bijective and  $\psi(u) > \psi(v)$  for all  $u, v \in D$  such that there exists at least one arc  $uv$  in  $D$ . A well known result establishes that  $D$  is a DAG if and only if it admits a topological ordering [15]. Nonetheless, when a topological ordering exists, it is in general not unique – see Figure 2. A reverse search enumeration of topological orderings of a given DAG can actually be found in [2, Section 3.5].

**Constrained topological ordering** We aim to reduce the number of possible topological orderings of a DAG by constraining them. Let  $D$  be a DAG and  $\psi$  a topological ordering. Taking advantage

---

### Algorithm 2: MINIMALWORDS

---

**Input:**  $\bar{w} = a_0 \dots a_p$ ,  $\mathcal{A} = \{0, \dots, n\}$   
**Output:** All minimal words of  $\Lambda^{\bar{w}}$

- 1 Set  $L$  to the empty list
- 2 **if**  $a_0 < n$  **then**
- 3     **for**  $i \in \{a_0 + 1, \dots, n\}$  **do**
- 4         Add the word  $i$  to  $L$
- 5 **for**  $k \in \{1, \dots, p\}$  **do**
- 6     **if**  $a_k < a_{k-1}$  **then**
- 7         **for**  $i \in \{a_k + 1, \dots, a_{k-1}\}$  **do**
- 8             Add the word  $a_0 \dots a_{k-1}i$  to  $L$
- 9 **for**  $i \in \{0, \dots, a_p\}$  **do**
- 10     Add the word  $a_0 \dots a_pi$  to  $L$
- 11 **return**  $L$

---

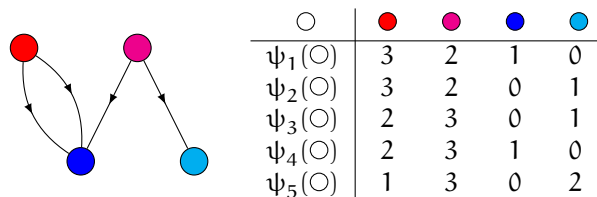


Figure 2: The DAG on the left admits five topological orderings, which are shown in the table.

of the vertical hierarchy of DAG, our first constraint is

$$\forall (u, v) \in D^2, \quad \mathcal{H}(u) > \mathcal{H}(v) \implies \psi(u) > \psi(v). \quad (4)$$

Applying (4) to the topological orderings presented in Figure 2,  $\psi_5$  must be removed, as  $\psi_5(\bullet) > \psi_5(\circ)$  and  $\mathcal{H}(\bullet) > \mathcal{H}(\circ)$ .

For any vertex  $v$ , and any  $u \in \mathcal{C}(v)$ , by definition,  $\mathcal{H}(v) > \mathcal{H}(u)$ . Therefore, there can be no arcs between vertices at same height. Any arbitrary order on them leads to a different topological ordering. The next constraint we propose relies on the lexicographical order:

$$\forall (u, v) \in D^2, \quad \mathcal{H}(u) = \mathcal{H}(v) \text{ and } \mathcal{C}_\psi(u) >_{\text{lex}} \mathcal{C}_\psi(v) \implies \psi(u) > \psi(v), \quad (5)$$

where  $\mathcal{C}_\psi(v)$  is the list  $[\psi(v_i) : v_i \in \mathcal{C}(v)]$  sorted by decreasing order w.r.t. the lexicographical order. In other words,  $\mathcal{C}_\psi(v)$  is a decreasing word – see Definition 2.1 – on the alphabet  $\mathcal{A} = \llbracket 0, \#D - 1 \rrbracket$ . Table 1 illustrates the behavior of (5) on the followed example of Figure 2.

$\circ$	$\bullet$	$\color{magenta}\bullet$	(5)
$\mathcal{C}_{\psi_1}(\circ)$	11	10	✓
$\mathcal{C}_{\psi_2}(\circ)$	00	10	✗
$\mathcal{C}_{\psi_3}(\circ)$	00	10	✓
$\mathcal{C}_{\psi_4}(\circ)$	11	10	✗

Table 1: Application of (5) to the remaining topological orderings of Figure 2 that satisfy (4). As  $\mathcal{C}_\psi(\bullet) = \mathcal{C}_\psi(\circ)$ , we only need to consider vertices  $\bullet$  and  $\color{magenta}\bullet$ . As  $\psi_i(\bullet) > \psi_i(\color{magenta}\bullet) \iff i \in \{1, 2\}$ , the only orderings that are kept are  $\psi_1$  and  $\psi_3$ .

The combination of those two constraints imposes uniqueness in all cases except when there exists  $(u, v) \in D^2$  such that  $\mathcal{C}_\psi(u) = \mathcal{C}_\psi(v)$  and  $u \neq v$ . It should be clear that if we impose the upcoming condition (6), such a pathological case can not occur.

$$\forall (u, v) \in D^2, \quad u \neq v \implies \mathcal{C}(u) \neq \mathcal{C}(v) \quad (6)$$

Upcoming Theorem 2.4 establishes that a DAG compresses a forest if and only if the topological order constrained by (4) and (5) is unique. In other words, an unambiguous characterization of FDAG is exhibited.

**Theorem 2.4.** *The following statements are equivalent:*

- (i)  $D$  fulfills (6),
- (ii) there exists a unique topological ordering  $\psi$  of  $D$  that satisfies both constraints (4) and (5),
- (iii) there exists a unique forest  $F \in \mathcal{F}$  – cf. (2) – such that  $D = \mathfrak{R}(F)$ ,

where  $\mathfrak{R}$  is the DAG reduction operation defined in Subsection 1.2.

**Proof.** (i)  $\iff$  (ii) follows from the above discussion. (iii)  $\implies$  (i) follows from the definition of  $\mathfrak{R}$ . Indeed, if there was two distinct vertices  $(u, v) \in D^2$  with the same multiset of children, they would have been compressed as a unique vertex in the reduction. We now prove that (i)  $\implies$  (iii).

In the first place, if  $D$  fulfills (6), then  $D$  must admit a unique leaf, denoted by  $\mathcal{L}(D)$ . Indeed, if there were two leaves  $l_1$  and  $l_2$ , we would have  $\mathcal{H}(l_1) = \mathcal{H}(l_2) = 0$  but also  $\mathcal{C}(l_1) = \mathcal{C}(l_2) = \emptyset$ , which

would violate (6). Let  $r_1, \dots, r_k$  be the vertices in  $D$  that have no parent. We define  $D_1, \dots, D_k$  as the subDAG rooted respectively in  $r_1, \dots, r_k$ . Then, we define  $T_i = \mathfrak{R}^{-1}(D_i)$  and  $F = \{T_1, \dots, T_k\}$ . The  $T_i$ 's are well defined as all vertices in  $D$  (consequently in  $D_i$ ) have a different multiset of children, and therefore compress distinct subtrees – i.e.  $F$  fulfills (2), therefore  $F \in \mathcal{F}$ . Moreover,  $D = \mathfrak{R}(F)$ .  $\color{red}{\spadesuit}$

In the sequel of the article, we shall only consider FDAG. Consequently, from Proposition 2.4, they admit a unique topological ordering  $\psi$  satisfying both constraints (4) and (5), called *canonical ordering*. Thus, for any FDAG  $D$ , the associated canonical ordering  $\psi$  will be implicitly defined. The vertices will be numbered accordingly to their ordering, i.e.  $D = (v_0, \dots, v_n)$  with  $\psi(v_i) = i$ . Finally, as a consequence of constraints (4) and (5), note that  $D$  can be partitioned in subsets of vertices with same height, each of them containing only consecutive numbered vertices. Figure 3 provides an example of a FDAG and its canonical ordering.

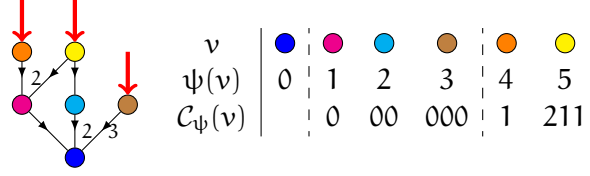


Figure 3: A FDAG  $D$  (left) and its canonical ordering  $\psi$  (right). Vertices that are at the same height are enclosed in the table between the dashed lines. Red arrows indicates the roots of the trees of the forest that is compressed by  $D$ .

### 2.3 Expansion rules

Reverse search techniques implies finding reduction rules, and then inverse them. Equally, we will define instead three expansion rules, of which inverse will be reduction rules. An expansion rule takes a FDAG and create a new DAG, that is “expanded” in the sense of having either more vertices or more arcs. Our rules are analysed at the end of the subsection, where notably we prove in Proposition 2.10 that expansion rules preserve the canonicalness. Moreover, we show in Proposition 2.11 that they are “bijective”: any FDAG is in the image of a unique FDAG through the expansion rules. We begin with a preliminary definition.

**Definition 2.5.** Let  $D$  be a FDAG, with  $D = (v_0, \dots, v_n)$ . We define the two following alphabets

$$\begin{aligned} \mathcal{A}_= &= \{\psi(v) : v \in D, \mathcal{H}(v) = \mathcal{H}(v_n)\} = \{p + 1, \dots, n\}, \\ \mathcal{A}_< &= \{\psi(v) : v \in D, \mathcal{H}(v) < \mathcal{H}(v_n)\} = \{0, \dots, p\}, \end{aligned}$$

where  $p \in \llbracket 0, n - 1 \rrbracket$  and  $\psi(\cdot)$  is the canonical ordering of  $D$ .

In other words,  $\mathcal{A}_=$  contains the indices of all vertices that have the same height as the vertex with the highest index according to  $\psi$ , and  $\mathcal{A}_<$  the indices of all vertices that have an inferior height. The FDAG presented in Figure 3 will serve as a guideline example all along this subsection. Here, we have  $\mathcal{A}_= = \{4, 5\}$  and  $\mathcal{A}_< = \{0, 1, 2, 3\}$ .

The three expansion rules are now introduced. Let  $D = (v_0, \dots, v_n)$ . Each of these rules is associated with an explicit symbol, which may be used, when necessary, to designate the rule afterward. It is worth noting that all of these rules will operate according to the vertex of highest index,  $v_n$ .

**Branching rule** ( $\color{red}{\spadesuit}$ ) This rule adds an arc between  $v_n$  and a vertex below. The end vertex of the new arc is chosen such that  $\mathcal{C}_\psi(v_n)$  remains a decreasing word. In Figure 4, ( $\color{red}{\spadesuit}$ ) is applied on our guideline example.

**Definition 2.6.** ( $\curvearrowright$ ) Let  $\mathcal{C}_\psi(v_n) = a_0 \cdots a_m$ . Choose  $a_{m+1} \in \mathcal{A}_<$  such that  $a_m \geq_{\text{lex}} a_{m+1}$  and add an arc between  $\psi^{-1}(a_{m+1})$  and  $v_n$ .



Figure 4: Branching rule applied to the FDAG of Figure 3. As  $\mathcal{C}_\psi(v_5) = 211$ , the only letters  $a$  we can pick from  $\mathcal{A}_< = \{0, 1, 2, 3\}$ , satisfying  $a \leq_{\text{lex}} 1$ , are 0 and 1. The only two possible outcomes of ( $\curvearrowright$ ) are the words (a) 2111 and (b) 2110.

**Elongation rule** ( $\uparrow$ ) This rule adds a new vertex  $v_{n+1}$  such that  $\mathcal{H}(v_{n+1}) = \mathcal{H}(v_n) + 1$ . Consequently, the alphabets change and become  $\mathcal{A}_= = \{n + 1\}$  and  $\mathcal{A}_< = \{0, \dots, n\}$ . Note that after using this rule, it is not possible to ever add a new vertex at height  $\mathcal{H}(v_n)$ . See Figure 5 for an illustration of this rule on the guideline example.

**Definition 2.7.** ( $\uparrow$ ) Add new vertex  $v_{n+1}$  such that  $\mathcal{C}_\psi(v_{n+1}) = a_0 \in \mathcal{A}_=$ .

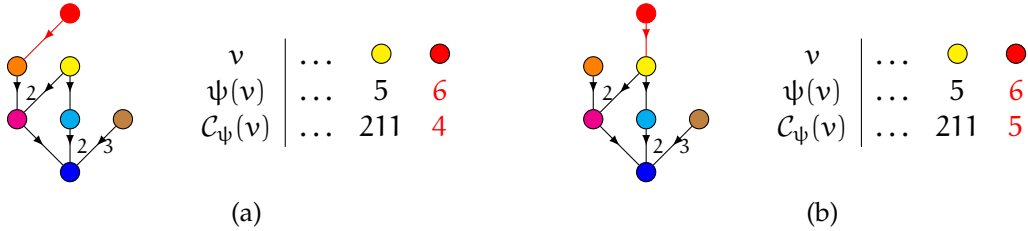


Figure 5: Elongation rule applied to the FDAG of Figure 3. As  $\mathcal{A}_= = \{4, 5\}$ , there are only two choices leading to (a)  $\mathcal{C}_\psi(v_6) = 4$  and (b)  $\mathcal{C}_\psi(v_6) = 5$ . The alphabets become  $\mathcal{A}_< = \{0, \dots, 5\}$  and  $\mathcal{A}_= = \{6\}$ .

**Widening rule** ( $\curvearrowleft$ ) This rule adds a new vertex  $v_{n+1}$  at height  $\mathcal{H}(v_n)$ . The vertex is added with children that respects the canonicalness of the DAG, that is, such that  $\mathcal{C}_\psi(v_{n+1}) >_{\text{lex}} \mathcal{C}_\psi(v_n)$  – as in condition (5). In other terms, denoting  $\Lambda_<$  the language of decreasing words on alphabet  $\mathcal{A}_<$ , and with  $\bar{w} = \mathcal{C}_\psi(v_n)$ ,  $\mathcal{C}_\psi(v_{n+1})$  must be chosen in  $\Lambda_<^{\bar{w}}$  – see Definition 2.2. However, this set is infinite, so we restrict  $\mathcal{C}_\psi(v_{n+1})$  to be chosen among the minimal words of  $\Lambda_<^{\bar{w}}$ . It follows from the definition of suffix-cut operator  $\text{SC}(\cdot)$  that, by inverting the said operator, the other words in  $\Lambda_<^{\bar{w}}$  can be obtained by performing repeated ( $\curvearrowright$ ) operations. Finally, this new vertex is added to  $\mathcal{A}_=$ .

**Definition 2.8.** ( $\curvearrowleft$ ) Add new vertex  $v_{n+1}$  such that

$$\mathcal{C}_\psi(v_{n+1}) \in \{w \in \Lambda_<^{\bar{w}} : w \text{ is a minimal word of } \Lambda_<^{\bar{w}}\}$$

with  $\bar{w} = \mathcal{C}_\psi(v_n)$ .

From Proposition 2.3 we now that such minimal words exist. We prove in the upcoming lemma that, as claimed,  $\mathcal{H}(v_{n+1}) = \mathcal{H}(v_n)$ .

**Lemma 2.9.** *Any element of  $\Lambda_{\leq}^{\bar{w}}$  defines a new vertex  $v_{n+1}$  such that  $\mathcal{H}(v_{n+1}) = \mathcal{H}(v_n)$ .*

**Proof.** From the definition of  $\mathcal{H}(\cdot)$  – (1), it suffices to prove that  $v_{n+1}$  admits at least one child at height  $h = \mathcal{H}(v_n) - 1$ . Let us denote  $b_0$  and  $a_0$  the first letter of, respectively,  $\mathcal{C}_\psi(v_{n+1})$  and  $\mathcal{C}_\psi(v_n)$ . Denoting  $v = \psi^{-1}(b_0)$  and  $u = \psi^{-1}(a_0)$ , we already know that  $\mathcal{H}(u) = h - 1$  as  $\psi$  respects (5) and  $\mathcal{C}_\psi(v_n)$  is a decreasing word. Therefore, as by construction  $\mathcal{C}_\psi(v_{n+1}) >_{\text{lex}} \mathcal{C}_\psi(v_n)$ , either (i)  $b_0 = a_0$  and therefore  $v = u$ , either (ii)  $b_0 >_{\text{lex}} a_0$ . In the latter, as  $\psi$  respects (4) and (5),  $\mathcal{H}(v) \geq \mathcal{H}(u) = h$ . But, as  $b_0 \in \mathcal{A}_{<}$ ,  $\mathcal{H}(v) < \mathcal{H}(v_n) = h + 1$ . In both cases,  $\mathcal{H}(v) = h$ .  $\color{red}{\spadesuit}$

Figure 6 illustrates the use of the widening rule on the followed example. It should be noted that the possible outcomes of  $(\spadesuit)$  are obtained by using Algorithm 2, applied to  $w = \mathcal{C}_\psi(v_n)$  and  $p$  – with  $\mathcal{A}_{<} = \{0, \dots, p\}$ .

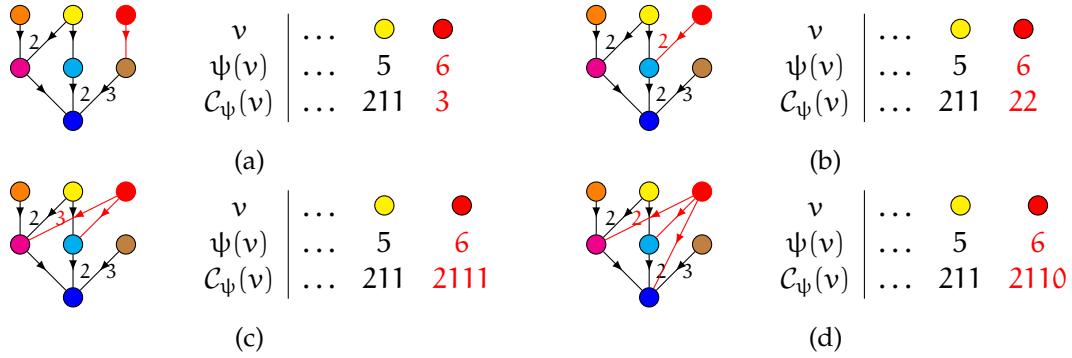



Figure 6: We apply  $(\spadesuit)$  to the FDAG of Figure 3. Here,  $\mathcal{A}_{<} = \{0, 1, 2, 3\}$  and  $\bar{w} = 211$ . As seen in Subsection 2.1, the minimal words of  $\Lambda_{\leq}^{\bar{w}}$  are 3, 22, 2111 and 2110. Therefore, there are 4 ways to add a new vertex  $v_6$  via the widening rule, that are such that (a)  $\mathcal{C}_\psi(v_6) = 3$ , (b)  $\mathcal{C}_\psi(v_6) = 22$ , (c)  $\mathcal{C}_\psi(v_6) = 2111$  or (d)  $\mathcal{C}_\psi(v_6) = 2110$ . Finally, we update  $\mathcal{A}_{\leq}$  to be equal to  $\{4, 5, 6\}$ .

**Analysis of the rules** Since our goal is to enumerate FDAG, it is required that the expansion rules indeed construct FDAG. This is achieved by virtue of the following proposition.

**Proposition 2.10.** *The expansion rules preserve the canonicalness property.*

**Proof.** Let  $D = (v_0, \dots, v_n)$  be a FDAG. The proposition follows naturally from the definitions:

- ( $\heartsuit$ ) Let  $a$  be the letter added to  $w = \mathcal{C}_\psi(v_n)$ . As  $wa >_{\text{lex}} w >_{\text{lex}} \mathcal{C}_\psi(v_{n-1})$ , the ordering is unchanged.
- ( $\dagger$ ) The new vertex  $v_{n+1}$  is such that  $\mathcal{H}(v_{n+1}) > \mathcal{H}(v_n)$ , so condition (4) is still met.
- ( $\spadesuit$ ) The new vertex  $v_{n+1}$  is chosen so that  $\mathcal{H}(v_{n+1}) = \mathcal{H}(v_n)$  and  $\mathcal{C}_\psi(v_{n+1}) >_{\text{lex}} \mathcal{C}_\psi(v_n)$ , so condition (5) is also still met.

Therefore, any DAG obtained from  $D$  is still a FDAG. 

Secondly, since our goal is to provide the FDAG space with an enumeration tree, which will be explored via the expansion rules, it is important that these expansion rules are “bijective” in the following sense: for any FDAG  $D$ , there exists a unique FDAG  $D'$  such that  $D$  is obtained from  $D'$  via one of the three rules  $(\Leftarrow)$ ,  $(\Uparrow)$  or  $(\Downarrow)$ .

Such  $D'$  can be constructed via Algorithm 3 as shown in upcoming Proposition 2.11. Conditional expressions applied to  $D$  are used to determine which modification should be applied to construct  $D'$ . The gray symbol (in the algorithm) next to these modifications indicates which expansion rule allows to retrieve  $D$  from  $D'$ .

**Proposition 2.11.** *Algorithm 3 applied to any FDAG constructs the unique antecedent of this FDAG.*

---

**Algorithm 3: ANTECEDENT**

---

**Input:**  $D = (v_0, \dots, v_n); w = \mathcal{C}_\psi(v_n); w' = \mathcal{C}_\psi(v_{n-1})$

```

1 if  $v_n$  is the only vertex of height  $\mathcal{H}(v_n)$  then
2   if  $\#w = 1$  then
3      $(\Uparrow)$  Delete vertex  $v_n$ 
4   else
5      $(\Leftarrow)$   $w \leftarrow SC(w)$ 
6 else
7   if  $w$  is a minimal word of  $\Lambda_{<}^{w'}$  then
8      $(\Downarrow)$  Delete vertex  $v_n$ 
9   else
10     $(\Leftarrow)$   $w \leftarrow SC(w)$ 


```

---

$SC(\cdot)$  is the suffix-cut operator defined in (3).

**Proof.** Let  $D = (v_0, \dots, v_n)$  be a FDAG. Let  $w = \mathcal{C}_\psi(v_n)$  and  $w' = \mathcal{C}_\psi(v_{n-1})$ . Two cases can occur: (i) either  $v_n$  is the only vertex at height  $\mathcal{H}(v_n)$ , (ii) or it is not.

- (i) It is clear in this case that  $D$  can not be obtained from any FDAG via the rule  $(\Downarrow)$  – otherwise  $v_n$  would not be alone at its height. Concerning  $(\Leftarrow)$  and  $(\Uparrow)$ , let us look at the number of children of  $v_n$ .
  - (a) If  $v_n$  admits only one child, it must come from an  $(\Uparrow)$  step, since  $(\Leftarrow)$  would imply that  $\#w \geq 2$ . Therefore, in this case,  $D$  can be retrieved among the outcomes of rule  $(\Uparrow)$  applied to  $D' = (v_0, \dots, v_{n-1})$ .
  - (b) Otherwise, when  $\#w > 1$ ,  $D$  can not come from an  $(\Uparrow)$  step, and must therefore come from  $(\Leftarrow)$ . Denoting  $v'_n$  the vertex with list of children  $SC(w)$  – see (3),  $D$  is one of the outcomes of  $D' = (v_0, \dots, v_{n-1}, v'_n)$  via  $(\Leftarrow)$ .
- (ii) following the same logic as (i),  $D$  can not be obtained via  $(\Uparrow)$ . We discriminate between rules  $(\Downarrow)$  and  $(\Leftarrow)$  by comparing  $w$  and  $w'$ . If  $w$  is a minimal word of  $\Lambda_{<}^{w'}$ , then  $D$  can not be obtained from  $(\Leftarrow)$  – this would break the canonical order. Therefore, in this case,  $D$  is an outcome of rule  $(\Downarrow)$  applied to  $D' = (v_0, \dots, v_{n-1})$ . Otherwise, if  $w$  is not a minimal word, then it can not be obtained from  $(\Downarrow)$ , and must come from a  $(\Leftarrow)$  step, applied to  $D' = (v_0, \dots, v_{n-1}, v'_n)$  where  $\mathcal{C}_\psi(v'_n) = SC(w)$ .


Whatever the case among those evoked, they correspond exactly to the conditional expressions of the Algorithm 3, which therefore constructs the correct antecedent of  $D$ , which is unique by virtue of the previous discussion. 

## 2.4 Enumeration tree

In this subsection, we construct the enumeration tree of FDAG derived from the expansion rules of Subsection 2.3. As aimed, their inverse is indeed a reduction rule.

**Theorem 2.12.** *Algorithm 3 is a reduction rule, as defined in Subsection 1.1.*

**Proof.** Let us denote  $f(D)$  the output of Algorithm 3 applied to a FDAG  $D$ . We need to prove that: (i)  $f(D) \subset D$  and (ii) for any  $D \neq D_0$ , there exists an integer  $k$  such that  $f^k(D) = D_0$ , where  $D_0$  is the FDAG with one vertex and no arcs.

- (i) Since Algorithm 3 deletes either one vertex and its leaving arcs, or just one arc,  $f(D)$  is indeed a subgraph of  $D$ .
- (ii) The sequence of general term  $f^k(D)$  is made of discrete objects whose size is strictly decreasing, therefore the sequence is finite and reaches  $D_0$ . 

The associated expansion rule is exactly, in light of Proposition 2.11, the union of the three expansion rules  $(\uparrow)$ ,  $(\swarrow)$  and  $(\mathcal{V})$ . Since  $D_0$ , the DAG with one vertex and no arcs, is a FDAG, by virtue of what precedes and with Algorithm 1 – here with  $g(\cdot) = \top$ , we just defined an enumeration tree covering the whole set of FDAG, whose root is  $D_0$ . A fraction of this enumeration tree is shown in Figure 7, illustrating the path from the root  $D_0$  to the FDAG of Figure 3. Unexplored branches are ignored, but are still indicated by their respective root.

## 2.5 Analysis of the growth of the tree

In this subsection, we investigate the growth of the enumeration tree defined in Subsection 2.4. First, we show that FDAG are in bijection with a set of particular matrices, whose combinatorial properties are known and give us access to an asymptotic expansion of the enumeration tree growth. Then, we prove that any FDAG has a linear number of successors<sup>1</sup> in this tree, which shows that its growth, although exponential, is reasonably controlled. Finally, we discuss the time complexity for actually constructing the tree.

Let us denote  $E_k$  the set of all FDAG that are accessible from  $D_0$  in exactly  $k$  steps in the enumeration tree – with  $E_0 = \{D_0\}$ ; then Table 2 depicts the values of  $\#E_k$  for the first nine values of  $k$ <sup>2</sup>.

$k$	0	1	2	3	4	5	6	7	8
$\#E_k$	1	1	3	12	61	380	2,815	24,213	237,348

Table 2: Number of FDAG accessible from  $D_0$  in  $k$  steps in the enumeration tree.

Actually, the terms of Table 2 coincide with the first terms of OEIS sequence A158691<sup>3</sup>, which counts the number of *row-Fishburn matrices*, that are upper-triangular matrices with at least one nonzero entry in each row. The *size* of such a matrix is equal to the sum of its entries.

**Theorem 2.13.** *There exists a bijection  $\Phi$  between the set of FDAG and the set of row-Fishburn matrices, such that if  $D$  is a FDAG and  $M = \Phi(D)$ , then*

$$D \in E_k \iff \text{size}(M) = k.$$

<sup>1</sup>“successor” in the sense of “children in the enumeration tree”. We make the distinction to avoid confusion with the children denoted by  $\mathcal{C}(\cdot)$ .

<sup>2</sup>These numbers were obtained numerically (cf. “Implementation” at the end of the article).

<sup>3</sup>OEIS Foundation Inc. (2021), The On-Line Encyclopedia of Integer Sequences, <http://oeis.org/A158691>.



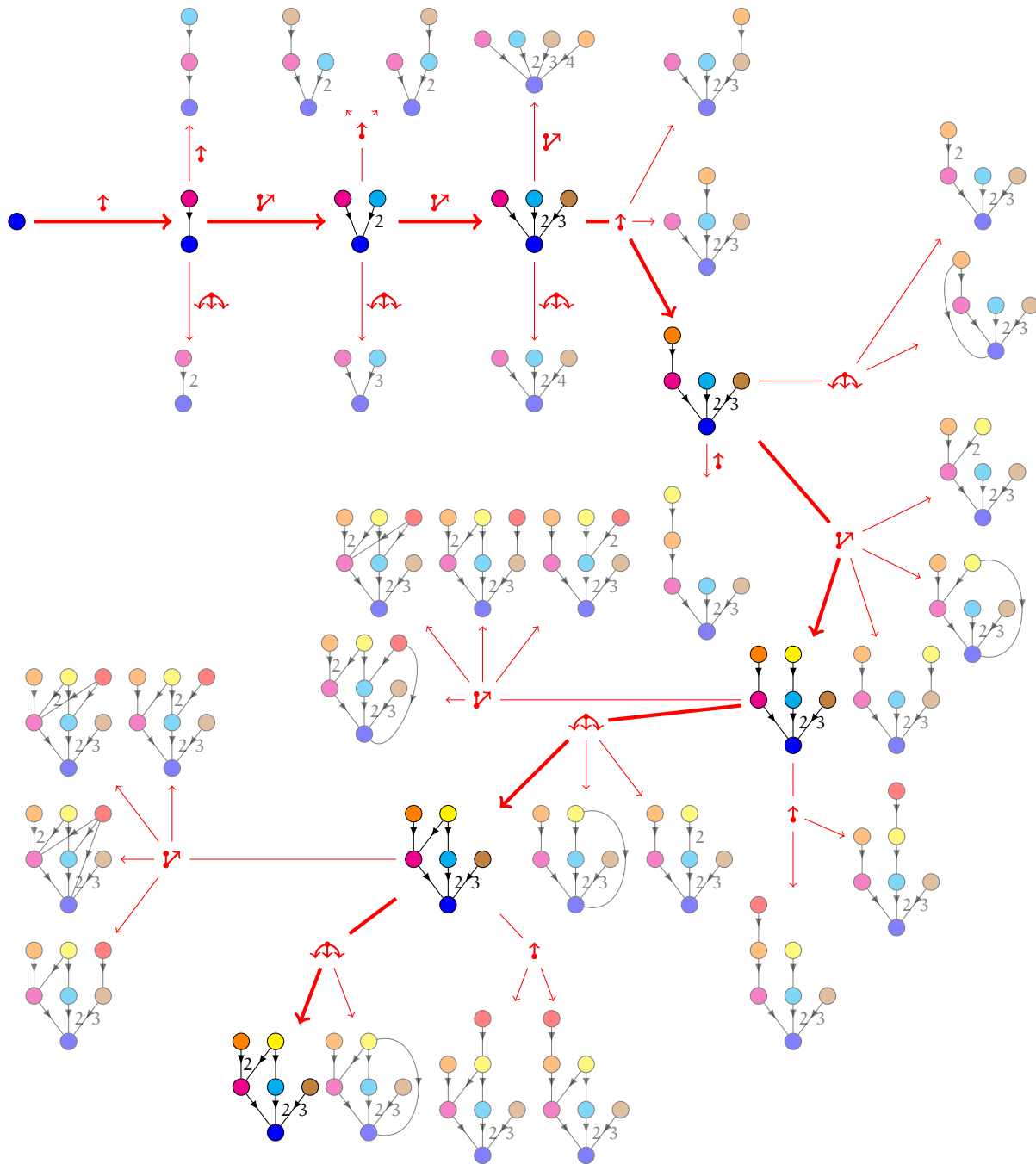



Figure 7: The path (in bold) in the FDAG enumeration tree leading to the FDAG of Figure 3. The unexplored branches are only displayed by their root, which are shown partially transparent. The order of insertion of the vertices of each FDAG is always the same, and follows the color code (in the order of insertion):  $\bullet$ ,  $\bullet$ ,  $\bullet$ ,  $\bullet$ ,  $\bullet$ ,  $\bullet$  and  $\bullet$ . With respect to the canonical ordering, they are numbered 0 to 6 in the same order.

**Proof.** The proof lies in Appendix A. 

This connection is to our advantage since Fishburn matrices (in general) are combinatorial objects widely explored in the literature as they are in bijection with many others – see [13, Section 2] for a general overview. Notably, the asymptotic expansion of the number of row-Fishburn matrices has been conjectured first by Jelínek [14] and then proved by Bringmann et al. [7].

**Proposition 2.14** (Jelínek, Bringmann et al.). *As  $k \rightarrow \infty$ ,*

$$\#E_k = k! \left( \frac{12}{\pi^2} \right)^k \left( \beta + O\left( \frac{1}{k} \right) \right)$$


*with  $\beta = \frac{6\sqrt{2}}{\pi^2} e^{\pi^2/24} = 1.29706861206\dots$*

Given the overall combinatorics of FDAG, it is no surprise that the enumeration tree grows extremely fast. However, despite this combinatorial explosion, the growth of the tree is controlled, as proven by the upcoming Theorem 2.15.

**Theorem 2.15.** *Any FDAG  $D$  has  $\Theta(\#D)$  successors in the FDAG enumeration tree.*

**Proof.** Let  $D = (v_0, \dots, v_n)$  be a FDAG. We denote  $\mathcal{C}_\psi(v_n) = a_0 \cdots a_m$ . Depending on the rule chosen:

- ( $\curvearrowright$ )  $a_{m+1}$  belongs to  $\mathcal{A}_< = \{0, \dots, p\}$ , so the maximum number of successors is at most  $p + 1$ , and at least 1, depending on the condition  $a_m \geq_{\text{lex}} a_{m+1}$ .
- ( $\uparrow$ ) The child of the new vertex is taken from  $\mathcal{A}_= = \{p + 1, \dots, n\}$  so the number of successors is exactly  $n - p$ .
- ( $\curvearrowleft$ ) Following Proposition 2.3, the number of successors is exactly  $\#\mathcal{A}_< = p + 1$ .

Combining everything, the number of successors is at least  $n + 2$  and at most  $n + p + 2 \leq 2n + 1$  (as  $p \leq n - 1$ , with equality for FDAG obtained just after using ( $\uparrow$ ) rule). 

In the previous proof, we have shown that the number of successors of a FDAG with  $n$  vertices is between  $n + 1$  and  $2n - 1$ . Figure 8 illustrates that these boundaries are tight, on 1 000 randomly generated FDAG. The random generation of a FDAG is as follows. Starting from  $D_0$  – the root, construct iteratively  $D_i$  as a successor of  $D_{i-1}$  in the enumeration tree, picked at random. We stop this random exploration at step  $k$ , and keep  $D_k$ . In Figure 8, we have generated 10 trajectories for each  $k \in \{1, \dots, 100\}$ .

It is indeed a suitable property that any FDAG admits a linear number of successors; but it would be of little use if the time required to compute those successors is too important. We demonstrate in the following proposition that temporal complexity is manageable.

**Proposition 2.16.** *Computing the successors of any FDAG  $D$  has worst case complexity  $O((\#D \deg(D))^2)$ .*

**Proof.** Let  $D = (v_0, \dots, v_n)$  be a FDAG with  $n + 1$  vertices. The explicit construction of the successors involves copying the structure of  $D$ . Therefore, the time complexity is divided into two parts: for each successor, (i) copy the structure of  $D$  and (ii) compute the new vertex or new arc.

Copying the structure of  $D$  requires to copy the vertices and their children, therefore the complexity is of the order of  $\sum_{i=0}^n (1 + \deg(v_i))$ , which can be roughly bounded by  $(n + 1)(\deg(D) + 1)$ .

The alphabets  $\mathcal{A}_<$  and  $\mathcal{A}_=$  can be computed – once and for all successors – within a linear traversal of  $D$ . This linear cost is neglected in the sequel of the proof since we aim for a quadratic complexity. Depending on the expansion rule, the complexity for computing the new vertex or new arc varies:

- (↔) The last letter of  $\mathcal{C}_\psi(v_n)$  determines the number of successors – but it is no more than  $p + 1$ . Since the new letter is added at the end of  $\mathcal{C}_\psi(v_n)$ , which was already copied, the complexity is bounded by  $p + 1$ .
- (†) Each successor is obtained by picking one element of  $\mathcal{A}_= = \{p + 1, \dots, n\}$ . The complexity is exactly (up to a constant)  $n - p$ .
- (↗) The successors are obtained by Algorithm 2, involving copying subwords of  $\mathcal{C}_\psi(v_n)$  – the overall complexity is bounded by  $(p + 1) \deg(v_n)$ .

The overall complexity is therefore of the order of  $(n+1)(\deg(D)+1) [(p + 1)(\deg(v_n) + 1) + n - p]$ . Using rough bounds, with  $\deg(v_n) \leq \deg(D)$  and  $p \leq n$ , we end up with the stated complexity. ✂

**Remark 2.17.** *In the proof of Proposition 2.16, we stated that the alphabets  $\mathcal{A}_<$  and  $\mathcal{A}_=$  can be computed in linear time. Nevertheless, it is more economical to keep track of them throughout the enumeration. Actually, it suffices to know the couple  $(n, p)$ ; how to update these indices has already been presented in Subsection 2.3, when introducing each expansion rule.*

We measured the time required to compute the successors of the randomly generated FDAG of Figure 8. The results are depicted in Figure 9, where we plotted (in blue) the total time  $t_D$  for computing all successors of a given FDAG  $D$ , and (in red) what we call *amortized time*, i.e.  $t_D / (\#D \deg(D))^2$ . As expected from Proposition 2.16, one can observe an asymptotic quadratic behaviour for the total time (in blue); concerning amortized time (in red), despite some variability, the upper bound seems to be constant.

## 2.6 Constraining the enumeration

In [18], the authors propose an algorithm to enumerate all trees with at most  $n$  vertices. They simply check whether the current tree has  $n$  vertices or not, and as their expansion rule adds one vertex at a time, they decide to cut a branch in the enumeration tree once they have reached  $n$  vertices. Similarly, adding a vertex to a tree can only increase its height or outdegree, so we can proceed in the same way to enumerate all trees with maximal height  $H$  and maximal outdegree  $d$ . Indeed, the number of trees satisfying those constraints is finite [4, Appendix D.2].

This property also holds in our approach : following one of the three expansion rules, we can only increase the height, outdegree or number of vertices of the FDAG. So, it makes sense to define similar constraints on the enumeration. However, for this constrained enumeration to generate a finite number of FDAG, constraints must be chosen wisely, as shown in the following proposition.

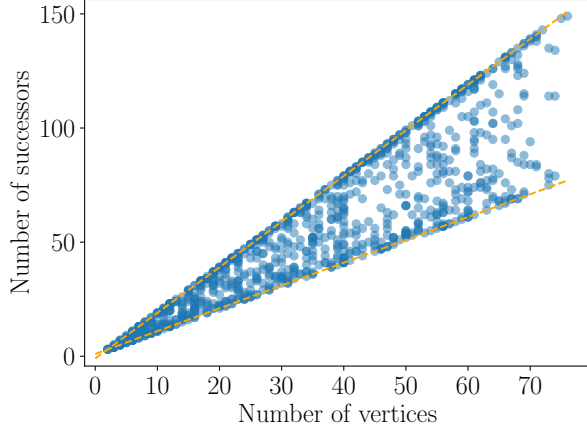


Figure 8: Numbers of successors of 1,000 random FDAG in the enumeration tree, according to their number of vertices. Orange lines have equations  $y = n + 1$  and  $y = 2n - 1$ .

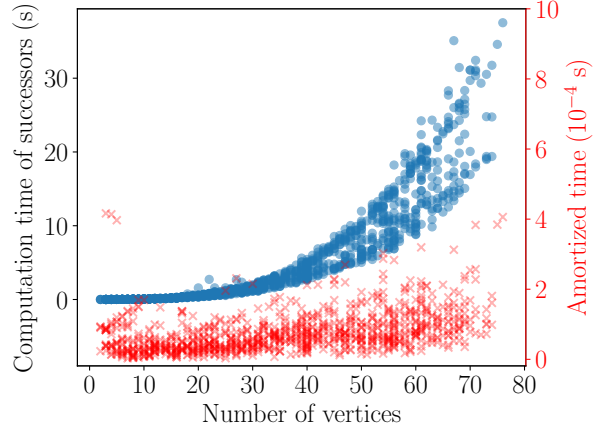


Figure 9: Computation times for the successors of the 1,000 random FDAG of Figure 8, according to their number of vertices: total computation time (in blue) and amortized time (in red).

**Proposition 2.18.** *The FDAG enumeration tree is finite if at least one of those set of constraints is chosen:*

- (i) *maximum number of vertices  $n$  and maximum outdegree  $d$ ,*
- (ii) *maximum height  $H$  and maximum outdegree  $d$ .*

**Proof.** As  $(\heartsuit)$  allows to add arcs indefinitely without changing the numbers of vertices, constraining on the maximum outdegree is mandatory in both cases. As the two others rules add vertices, constraining by the number of vertices leads to a finite enumeration tree – (i) is proved. To conclude, we only need to prove that  $(\heartsuit)$  can not be repeated an infinite number of times, i.e. there is only a finite number of new vertices that can be added at a given height, up to the maximum outdegree. This is achieved by virtue of the upcoming lemma.

Let  $H > 2$  and  $d \geq 1$ . Let  $D$  be the FDAG constructed so that for each  $0 \leq h \leq H$ ,  $D$  has the maximum possible number  $n_h$  of vertices of height  $h$  and with maximum outdegree  $d$ . Initial values are  $n_0 = 1$  and  $n_1 = d$ .

**Lemma 2.19.**  $\forall 2 \leq h \leq H, \quad n_h = \sum_{k=1}^d \binom{k + n_{h-1} - 1}{k} \binom{d - k + n_0 + \dots + n_{h-2}}{d - k}.$

Let  $h \geq 2$  be fixed. To lighten the notation, let  $n = n_{h-1}$  and  $m = n_0 + \dots + n_{h-2}$ . Let  $v$  be a vertex to be added at height  $h$ . For any vertex  $v_i$  at height  $h - 1$ , let  $x_i$  be the multiplicity of  $v_i$  in  $\mathcal{C}(v) - 0$  if  $v_i \notin \mathcal{C}(v)$ . Similarly, for any vertex  $v_j$  with  $\mathcal{H}(v_j) \leq h - 2$ ,  $y_j$  is the multiplicity of  $v_j$  in  $\mathcal{C}(v) -$  possibly 0. By definition of  $\mathcal{H}(\cdot)$  – see (1), at least one  $x_i$  is non-zero. Therefore, there exist  $k \in \llbracket 1, d \rrbracket$  such that:

$$\begin{aligned} x_1 + \dots + x_n &= k \\ y_1 + \dots + y_m &\leq d - k \end{aligned}$$

By virtue of the stars and bars theorem, for a fixed  $k$ , there are  $\binom{k+n-1}{k}$  choices for variables  $x_i$ , and  $\binom{d-k+m}{d-k}$  for variables  $y_j$ . Summing upon all values for  $k$  proves the claim. ✍

**Remark 2.20.** In the constrained enumeration proposed in [18], all the trees with  $n$  vertices are the leaves of the enumeration tree. To get all trees with  $n + 1$  vertices, it suffices to add to the enumeration all children of these leaves, i.e. trees obtained by adding a single vertex to them. This property – moving from one parameter value to the next by enumerating just one step further – does not hold anymore as soon as our set of constraints involve the maximum outdegree  $d$ , both for trees and FDAG. For instance, from a FDAG of height  $H$ , one can obtain FDAG of height  $H + 1$  by using  $(\uparrow)$  once and repeating  $(\curvearrowright)$  up to  $d - 1$  times.

### 3 Enumeration of forests of subtrees

Once the reverse search scheme has been set up to enumerate a certain type of structure, it is natural to move to a finer scale by using the same scheme to enumerate substructures. However, the notion of “substructure” is not obvious to derive from the main structure, as several choices are possible – e.g. for trees one can think of subtrees [22, 5], subset trees [8], etc. From a practical point of view, the enumeration of substructures permits to solve the frequent pattern mining problem – which will be tackled in Section 4.

In this section we define forests of subtrees, which will be our substructures. Compressed as FDAG, these objects will be called subFDAG. We then address the problem of enumerating all subFDAG appearing in an FDAG  $D$  – similar as the one of enumerating all subtrees of a tree.

**Forests of subtrees** Similarly to forest being tuple of trees, *forests of subtrees* are tuple of subtrees, satisfying (2). Formally:

**Definition 3.1.** Let  $F$  and  $f$  be two forests.  $f$  is a forest of subtrees of  $F$  if and only if

$$\forall t \in f, \exists T \in F, t \in \mathcal{S}(T).$$

Forest of subtrees can be directly constructed from FDAG, as shown by the upcoming proposition. Let  $D$  be a FDAG, and  $V$  be a subset of vertices of  $D$ .

**Proposition 3.2.** If  $\forall v \in V, \mathcal{C}(v) \subseteq V$ , then  $V$  defines a FDAG  $\Delta$ , such that  $\mathfrak{R}^{-1}(\Delta)$  is a forest of subtrees of  $\mathfrak{R}^{-1}(D)$ .

**Proof.** We recall from Subsection 1.2 that the notation  $D[v]$  stands for the subDAG of  $D$  rooted in  $v$  composed of  $v$  and all its descendants  $\mathcal{D}(v)$ . The notation  $\mathfrak{R}^{-1}(D[v])$  stands for the tree compressed by  $D[v]$ . The demonstration is in two steps. (i) Remove from  $D$  the vertices that does not belong to  $V$ ; as there are no arcs that leave  $V$  by hypothesis, end up with a FDAG. Let us call  $\Delta$  this FDAG. (ii) Let  $\rho$  be a root of  $\Delta$ . By construction,  $\rho$  is also a vertex in  $D$ . Among all roots of  $D$ , there exists a root  $r$  such that  $\rho \in \mathcal{D}(r)$ . Therefore,  $D[\rho]$  is a subDAG of  $D[r]$ , and then  $t = \mathfrak{R}^{-1}(D[\rho])$  is a subtree of  $T = \mathfrak{R}^{-1}(D[r])$  – with  $T \in F = \mathfrak{R}^{-1}(D)$ . As  $\Delta[\rho]$  and  $D[\rho]$  are isomorphic,  $t \in f = \mathfrak{R}^{-1}(\Delta)$ . Therefore we have proved that  $\forall t \in f, \exists T \in F, t \in \mathcal{S}(T)$ . ✍

We say that the FDAG  $\Delta$  is a *subFDAG*<sup>4</sup> of  $D$ . Figure 10 provides an example of such a construction.

<sup>4</sup>Not to be confused with *subDAG*, introduced in Subsection 1.2. A subDAG admits a single root and therefore compresses a single tree, whereas a subFDAG admits several roots and compresses a forest.

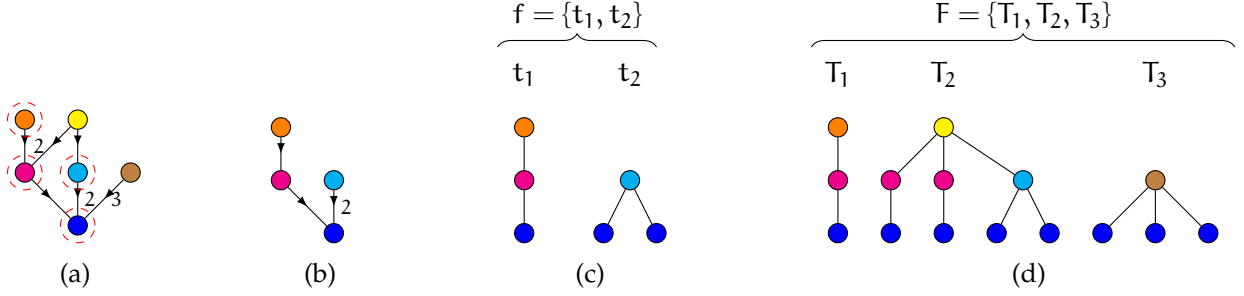


Figure 10: Construction of a forest of subtrees from FDAG. (a) A FDAG  $D$ . The set  $V$  is circled in red. (b) The FDAG  $\Delta$  (c) The forest  $f$  compressed by  $\Delta$ . (d) The forest  $F$  compressed by  $D$ . One can spot that  $t_1 \in \mathcal{S}(T_1)$ ,  $t_2 \in \mathcal{S}(T_2)$  so  $f$  is a forest of subtrees of  $F$ , and  $\Delta$  a subFDAG of  $D$ .

**Enumeration of subFDAG** We now solve the following enumeration problem: given a forest  $F$ , find all forests of subtrees of  $F$ . Equally, given a FDAG  $D$ , find all subFDAG of  $D$ . To address this, we make extensive use of the reverse search technique, adapting the one presented in Section 2.

Since a subFDAG is also a FDAG, it admits successors in the enumeration tree defined in Section 2. We are interested in those of these successors who are also subFDAG (if any). In fact, since a subFDAG can be defined from a set of vertices, all one has to do is determine which new vertex can be chosen to expand an existing subFDAG – corresponding to a  $(\uparrow)$  or  $(\uparrow')$  step. The covering of all added new arcs is implicit in this construction and corresponds to some steps of  $(\downarrow)$ .

Let  $\Delta$  be a subFDAG of  $D$  and  $v$  its last inserted vertex – it is also the vertex with the largest ordering number in  $D$ . We denote by  $S(\Delta)$  the set of all vertices  $v' \in D$  that can be added to  $\Delta$  to expand it to a new subFDAG. Let us call  $S(\Delta)$  the *set of candidate vertices* of  $\Delta$ . More precisely:

**Lemma 3.3.**  $S(\Delta)$  is the set of vertices  $v' \in D$  that satisfies both:

- (i)  $\mathcal{C}(v') \subseteq \Delta$
- (ii)  $\psi(v') > \psi(v)$

where  $\psi(\cdot)$  is the canonical ordering of  $D$ .

**Proof.** (i) This condition is necessary so that  $\Delta' = \Delta \cup \{v'\}$  fulfill the requirements for Proposition 3.2. (ii) This condition is necessary so that  $\Delta'$  remains a FDAG. As  $\psi(v') > \psi(v)$ , either  $\mathcal{H}(v') = \mathcal{H}(v) + 1$  – then it is a  $(\uparrow)$  step – or  $\mathcal{H}(v') = \mathcal{H}(v)$  and  $\mathcal{C}_\psi(v') >_{\text{lex}} \mathcal{C}_\psi(v)$  – for a  $(\uparrow')$  step. ✍

When  $S(\Delta)$  is not empty, picking  $s \in S(\Delta)$  ensure that  $\Delta' = \Delta \cup \{s\}$  is a subFDAG of  $D$ . With respect to the enumeration tree of Section 2,  $\Delta$  is an ancestor of  $\Delta'$  – but not necessarily its parent, since the steps of  $(\downarrow)$  are implicit.  $\Delta'$  is called an *heir* of  $\Delta$ . We can in turn calculate  $S(\Delta')$ , by updating  $S(\Delta)$ : (i) remove from  $S(\Delta)$  all vertices  $v'$  such that  $\psi(s) > \psi(v')$ ; (ii) in  $D$ , look only after the vertices  $v'$  such that  $s \in \mathcal{C}(v') \subseteq \Delta \cup \{s\}$  and add them to  $S(\Delta')$ .

If  $\Delta'$  is an heir of  $\Delta$ , then by removing the last inserted vertex of  $\Delta'$ , one can retrieve  $\Delta$ . This define a reduction rule  $f$ , and therefore an enumeration tree. Algorithm 4 is meant to construct the set  $f^{-1}(\Delta)$ . Applying Algorithm 1 together with it, and starting from  $\Delta = \mathcal{L}(D)$ <sup>5</sup> – in this case,  $S(\Delta)$  is

<sup>5</sup>where  $\mathcal{L}(D)$  designates the leaf of  $D$ , i.e. the only vertex without children.

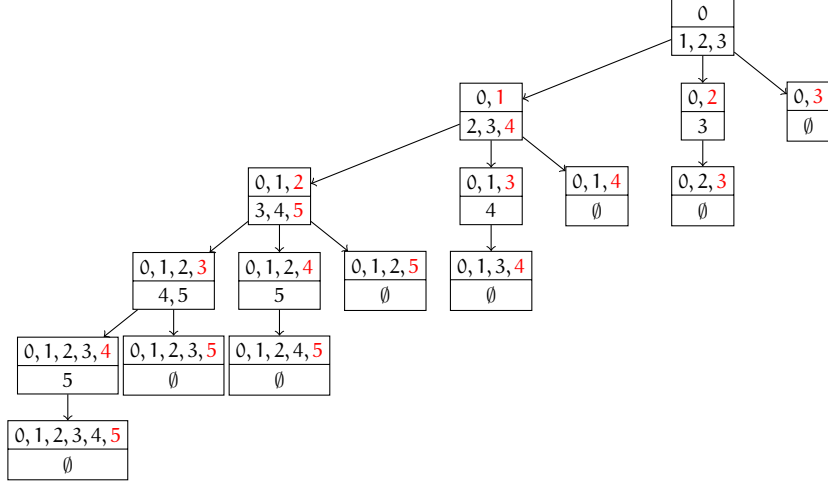


Figure 11: Enumeration tree of the subFDAG of the FDAG of Figure 3, using both Algorithm 1 and Algorithm 4. The indices of the vertices correspond to the canonical ordering defined in Figure 3. In each vertex, the upper part corresponds to the current subFDAG  $\Delta$  whereas the lower part stands for the set  $S(\Delta)$ . Numbers in red indicate what changes for an heir compared to its parent.

the set of parents of  $\mathcal{L}(D)$  of height 1 – permits to enumerate all subFDAG of  $D$ . Figure 11 provide an example by enumerating all subFDAG of the FDAG of Figure 3.

## 4 Frequent subFDAG mining problem

Using the reverse search formalism introduced in Subsection 1.1, the *frequent pattern mining problem* can be formulated as follow: from a dataset  $X = \{s_1, \dots, s_n\}$  with  $s_i \in \mathcal{S}$ , and a fixed threshold  $\sigma$ , find all elements  $s \in \mathcal{S}$  that satisfies  $\text{freq}(s, X) \geq \sigma$ , where  $\text{freq}(\cdot)$  is a function, to be defined, that counts the frequency of appearance of  $s$  in the dataset  $X$ . This problem can be solved using Algorithm 1 with  $g(s, X, \sigma) = (\text{freq}(s, X) \geq \sigma)$ . The function  $g$  is trivially anti-monotone.

---

### Algorithm 4: HEIRS

---

**Input:**  $D, [\Delta, S(\Delta)]$

- 1 Set  $L$  to the empty list
- 2 **for**  $s \in S(\Delta)$  **do**
- 3     Let  $S'$  be a copy of  $S(\Delta)$
- 4      $S' \leftarrow S' \setminus \{v' \in S' : \mathcal{C}_\psi(v') \leq_{\text{lex}} \mathcal{C}_\psi(s)\}$
- 5      $S' \leftarrow S' \cup \{v' \in D : s \in \mathcal{C}(v') \subseteq \Delta \cup \{s\}\}$
- 6     Add  $[\Delta \cup \{s\}, S']$  to  $L$
- 7 **return**  $L$

---

We emphasize here that each possible definition of “ $s$  appears in  $X$ ” leads to a different data mining problem. The choice of this definition is therefore of prime importance. In particular, this choice should induce a way of calculating  $\text{freq}(s, X)$  that reflects the specificity of the chosen reduction rule  $f$ , so that  $\{s \in f^{-1}(s_0) | g(s) = \top\}$  can be constructed directly, instead of first generating  $f^{-1}(s_0)$  and then filtering according to the value of  $g$ . Indeed, if  $g$  is too restrictive, and  $f^{-1}(s_0)$  too large, one would have to enumerate objects that are not relevant to the problem, which is not desirable.

In this article, the problem we consider is the following: given a set of trees  $X = \{T_1, \dots, T_n\}$ , account for forests of subtrees that appear simultaneously in different  $T_i$ 's. In other words, if we denote



$\mathcal{F}_i$  the set of all forests of subtrees appearing in the forest formed by  $\{T_i\}$ , we are interested in the study of  $\bigcap_{i \in I_\sigma} \mathcal{F}_i$  where  $I_\sigma \subseteq \llbracket 1, n \rrbracket$ , such that  $\#I_\sigma \geq \sigma \cdot n$ .

A first, naive strategy would be to first build the  $\mathcal{F}_i$ 's, e.g. by using Algorithm 4 on  $\mathfrak{R}(T_i)$ , and then construct  $\bigcap_{i \in I_\sigma} \mathcal{F}_i$  for all possible choices of  $I_\sigma$ . Obviously, this approach has its weaknesses: (i) many subFDAG will be enumerated for nothing or in several copies, and (ii) it does not take into account that  $X$  is itself a forest. Our aim is to propose a variant of Algorithm 4 that, applied to  $\mathfrak{R}(X)$ , would enumerate only subFDAG appearing in the  $\mathfrak{R}(T_i)$ 's with a large enough frequency.

Given a forest  $F = \{T_1, \dots, T_n\}$  and its DAG compression  $D = \mathfrak{R}(F)$ , we have to retrieve, for each vertex in  $D$ , their origin in the dataset, that is, from which tree they come from. This issue has already been addressed in a previous article [5, Section 3.3], and has led to the concept of *origin*. For any vertex  $v \in D$ , the origin of  $v$  is defined as

$$o(v) = \{i \in \llbracket 1, n \rrbracket : \mathfrak{R}^{-1}(D[v]) \in \mathcal{S}(T_i)\},$$

where  $\mathfrak{R}^{-1}(D[v])$  designates the tree compressed by the subDAG  $D[v]$  rooted in  $v$ . In others words,  $o(v)$  represents the set of trees for which  $\mathfrak{R}^{-1}(D[v])$  is a subtree. We state in [5, Proposition 3.4] that origins can be iteratively computed in one exploration of  $D$ . The proof lies in the property that if  $i \in o(v)$ , then for all  $v' \in \mathcal{D}(v)$ ,  $i \in o(v')$  – as  $\mathfrak{R}^{-1}(D[v']) \in \mathcal{S}(\mathfrak{R}^{-1}(D[v]))$ .

Let  $\Delta$  be a subFDAG of  $D$ . For  $\Delta$  to compress a forest of subtrees of a tree  $T_i$ , it is necessary that  $i \in o(v)$  for all  $v \in \Delta$ . Therefore, the set of trees for which  $\Delta$  compress a forest of subtrees – the *origin* of  $\Delta$ , denoted by  $o(\Delta)$  – is equal to

$$o(\Delta) = \bigcap_{v \in \Delta} o(v).$$

If  $\Delta' = \Delta \cup \{s\}$  is an heir of  $\Delta$  – as defined earlier, then  $o(\Delta') = o(\Delta) \cap o(s)$ . Algorithm 4 can therefore be refined so that  $\Delta'$  should be ignored if  $o(\Delta') = \emptyset$  – as  $\Delta'$  does not anymore compresses any forest of subtrees actually present in the trees of  $F$ .

So far we neglected the threshold  $\sigma$ . We only want to keep subFDAG that appear in at least  $\sigma\%$  of the data. If  $\#o(\Delta)/\#F < \sigma$ , then the successors of  $\Delta$  are not investigated. Indeed, as  $o(\cdot)$  is a decreasing function, successors of  $\Delta$  can not exceed the threshold again.

We can finally introduce Algorithm 5 that solves the frequent subFDAG mining problem for trees. With the notations of Subsection 1.1, this algorithm builds the set  $\{\Delta' \in f^{-1}(\Delta) \mid \text{freq}(\Delta', F) \geq \sigma\}$ , with  $\text{freq}(\Delta', F) = \#o(\Delta')/\#F$ . The set is also built directly, without any posterior filtering, which is suitable as discussed at the beginning of the present subsection.

---

**Algorithm 5:** FREQUENTHEIRS

---

**Input:**  $D = \mathfrak{R}(F)$ ,  $[\Delta, S(\Delta), o(\Delta)]$ ,  $\sigma$

1 Set  $L$  to the empty list

2 **for**  $s \in S(\Delta)$  **do**

3     **if**  $o(\Delta) \cap o(s) \neq \emptyset$  **and**  $\#(o(\Delta) \cap o(s)) \geq \sigma \cdot \#F$  **then**

4         Let  $S'$  be a copy of  $S(\Delta)$

5          $S' \leftarrow S' \setminus \{v' \in S' : \mathcal{C}_\psi(v') \leq_{\text{lex}} \mathcal{C}_\psi(s)\}$

6          $S' \leftarrow S' \cup \{v' \in D : s \in \mathcal{C}(v') \subseteq D_0 \cup \{s\}\}$

7         Add  $[\Delta \cup \{s\}, S', o(\Delta) \cap o(s)]$  to  $L$

8 **return**  $L$

---

We stated earlier that we wanted to avoid generating unnecessary or multiple copies of subFDAG, which is achieved with Algorithm 5. We now empirically study what we have gained from this, by

comparing the use of Algorithm 5 on  $D = \mathfrak{R}(F)$ , with the use of Algorithm 4 on each  $\mathfrak{R}(T_i)$ . As in Subsection 2.5, we generated 1 000 random DAG, 10 repetitions for each  $k \in \{1, \dots, 100\}$ , creating  $D_k$  as a random exploration in  $k$  steps of the enumeration tree defined in Section 2. We assume  $D_k = \mathfrak{R}(f)$  where  $f = \{\mathfrak{R}^{-1}(D_k[r]) : r \text{ root of } D_k\}$ . For each  $D_k$ , we have computed the quotient

$$Q(D_k) = \frac{\text{number of subFDAG of } D_k \text{ enumerated via Algorithm 5}}{\sum_{r \text{ root of } D_k} \text{number of subFDAG of } D_k[r] \text{ enumerated via Algorithm 4}}$$

with parameter  $\sigma = 0$  when using Algorithm 5. The results are provided in Figure 12. Despite a rather marked variability, there is a general trend of decreasing as the number of vertices increases. We obtain fairly low quotients, around 20%, quite quickly. Given the combinatorial explosion of the objects to be enumerated, such an advantage is of the greatest interest.

## Implementation

The `treex` library for Python [3] is designed to manipulate rooted trees, with a lot of diversity (ordered or not, labeled or not). It offers options for random generation, visualization, edit operations, conversions to other formats, and various algorithms. The enumeration of Section 2 and the algorithms of Section 3 and 4 have been implemented as a module of `treex` so that the interested reader can manipulate the concepts discussed in this paper in a ready-to-use manner. Installing instructions and the documentation of `treex` can be found from [3].

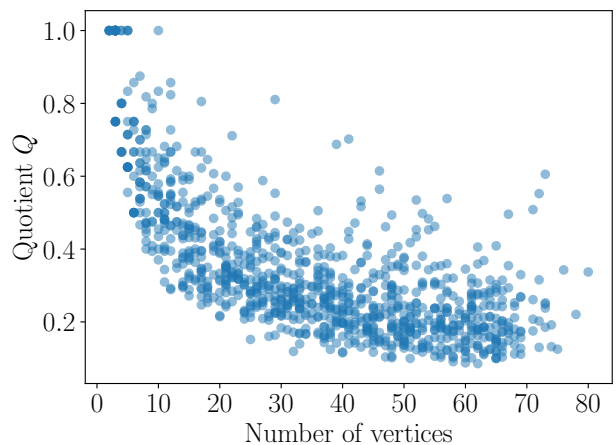


Figure 12: Quotient  $Q(D)$  according to the number of vertices of  $D$ . Here 1 000 random FDAG are displayed.

## Acknowledgments

This work has been supported by the European Union’s H2020 project ROMI. The authors would like to thank Dr. Arnaud Mary for his helpful suggestions on the draft of the article. Finally, the authors would like to thank an anonymous reviewer for his/her valuable comments, which help them to significantly improve the quality of their manuscript.

## References

- [1] Tatsuya Asai, Hiroki Arimura, Takeaki Uno, and Shin-Ichi Nakano. Discovering frequent substructures in large unordered trees. In *International Conference on Discovery Science*, pages 47–61. Springer, 2003.
- [2] David Avis and Komei Fukuda. Reverse search for enumeration. *Discrete Applied Mathematics*, 65(1-3):21–46, 1996.
- [3] Romain Azaïs, Guillaume Cerutti, Didier Gemmerlé, and Florian Ingels. `treex`: a python package for manipulating rooted trees. *The Journal of Open Source Software*, 4, 2019.

- [4] Romain Azaïs, Jean-Baptiste Durand, and Christophe Godin. Approximation of trees by self-nested trees. In *ALLENEX 2019 - Algorithm Engineering and Experiments*, pages 1–24, San Diego, United States, January 2019. URL: <https://hal.archives-ouvertes.fr/hal-01294013>, doi:10.10860.
- [5] Romain Azaïs and Florian Ingels. The weight function in the subtree kernel is decisive. *Journal of Machine Learning Research*, 21:1–36, 2020.
- [6] Edward A Bender and S Gill Williamson. *Lists, Decisions and Graphs*. S. Gill Williamson, 2010.
- [7] Kathrin Bringmann, Yingkun Li, and Robert C Rhoades. Asymptotics for the number of row-fishburn matrices. *European Journal of Combinatorics*, 41:183–196, 2014.
- [8] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Advances in neural information processing systems*, pages 625–632, 2002.
- [9] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. cambridge University press, 2009.
- [10] Christophe Godin and Pascal Ferraro. Quantifying the degree of self-nestedness of trees: application to the structural analysis of plants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(4):688–703, 2010.
- [11] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15(1):55–86, 2007.
- [12] John C. Hart and Thomas A. DeFanti. Efficient antialiased rendering of 3-d linear fractals. *SIGGRAPH Comput. Graph.*, 25(4):91–100, July 1991. URL: <http://doi.acm.org/10.1145/127719.122728>, doi:10.1145/127719.122728.
- [13] Hsien-Kuei Hwang and Emma Yu Jin. Asymptotics and statistics on fishburn matrices and their generalizations. *arXiv preprint arXiv:1911.06690*, 2019.
- [14] Vít Jelínek. Counting general and self-dual interval orders. *Journal of Combinatorial Theory, Series A*, 119(3):599–614, 2012.
- [15] Arthur B Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.
- [16] Ailsa H Land and Alison G Doig. An automatic method for solving discrete programming problems. In *50 Years of Integer Programming 1958-2008*, pages 105–132. Springer, 2010.
- [17] Shin-Ichi Nakano. Efficient generation of plane trees. *Information Processing Letters*, 84(3):167–172, 2002.
- [18] Shin-ichi Nakano and Takeaki Uno. Efficient generation of rooted trees. *National Institute for Informatics (Japan), Tech. Rep. NII-2003-005E*, 8, 2003.
- [19] Sebastian Nowozin. *Learning with structured data: applications to computer vision*. PhD thesis, Berlin Institute of Technology, 2009.
- [20] Benno Schwikowski and Ewald Speckenmeyer. On enumerating all minimal solutions of feedback problems. *Discrete Applied Mathematics*, 117(1-3):253–265, 2002.
- [21] Ivan E. Sutherland. Sketchpad: A man-machine graphical communication system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, AFIPS '63 (Spring)*, pages 329–346, New York, NY, USA, 1963. ACM. URL: <http://doi.acm.org/10.1145/1461551.1461591>, doi:10.1145/1461551.1461591.
- [22] S.V.N. Vishwanathan and Alexander J Smola. Fast kernels on strings and trees. *Advances on Neural Information Processing Systems*, 14, 2002.
- [23] Kazuaki Yamazaki, Toshiki Saitoh, Masashi Kiyomi, and Ryuhei Uehara. Enumeration of nonisomorphic interval graphs and nonisomorphic permutation graphs. *Theoretical Computer Science*, 806:310–322, 2020.
- [24] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724. IEEE, 2002.

## A A bijection between FDAG and row-Fishburn matrices

This section is dedicated to the proof of Theorem 2.13, which is in two steps. First, we recall the natural bijection between FDAG and their adjacency matrices; the latter are then put into bijection with the row-Fishburn matrices.

**FDAG  $\leftrightarrow$  Reduced adjacency matrix** Let  $D = (v_0, \dots, v_n)$  be a FDAG constructed in  $k$  steps from  $D_0$  in the enumeration tree defined in Subsection 2.4. The adjacency matrix of  $D$  is defined as  $A = (A_{i,j})_{i,j \in [n,0]^2}$  where, if  $m$  is the multiplicity of  $v_j$  in  $\mathcal{C}(v_i)$ , then  $A_{i,j} = m$  – possibly 0 if  $v_j \notin \mathcal{C}(v_i)$ . By construction of  $D$ , as  $v_n$  is the last inserted vertex, it has no parents, so  $A_{n,\cdot}$  is a column of zeros; and as  $v_0$  is a leaf, it has no children, so  $A_{\cdot,0}$  is a row of zeros. We define the *reduced* adjacency matrix  $M$  as the matrix  $A$  deprived of this column and this row. Therefore,  $M = (A_{i,j})_{i \in [n,1], j \in [n-1,0]}$ . As a vertex can not be a parent to any vertex introduced after it, we have  $A_{i,j} = 0$  for all  $i \leq j$  – so that  $M$  is an upper-triangular matrix. In addition, as all vertices except  $v_0$  have at least one child, there is at least one non-zero entry in each row of  $M$ . Therefore,  $M$  is a row-Fishburn matrix. However, we have no guarantee that this matrix verifies  $\text{size}(M) = k$ .

**Reduced adjacency matrix  $\rightarrow$  Incremental adjacency matrix** Let  $D = (v_0, \dots, v_n)$  be a FDAG, and  $M$  its reduced adjacency matrix. Let  $M_i$  be the row of  $M$  corresponding to  $\mathcal{C}_\psi(v_i)$ . The incremental adjacency matrix  $\hat{M}$  is defined as:

$$\begin{cases} \hat{M}_1 = M_1 \\ \hat{M}_{i+1} = M_{i+1} \ominus M_i \end{cases}$$

where the  $\ominus$  operation is defined as follow: given two rows  $a_0 \cdots a_n$  and  $b_0 \cdots b_n$ , then denoting  $j = \min\{i : a_i \neq b_i\}$ , and  $c = a_j - b_j$ ,

$$\begin{array}{r} a_0 \cdots a_{j-1} \quad a_j \quad a_{j+1} \cdots a_n \\ \ominus b_0 \cdots b_{j-1} \quad b_j \quad b_{j+1} \cdots b_n \\ \hline = 0 \cdots 0 \quad c \quad a_{j+1} \cdots a_n \end{array} .$$

We claim that this new matrix  $\hat{M}$  is a row-Fishburn matrix of size  $k$ , if  $D \in E_k$ . Actually, since  $M$  was already a row-Fishburn matrix, we just have to check that the size is correct. Let us consider  $v_i$  and  $v_{i+1}$ . The vertex  $v_{i+1}$  has been constructed from  $v_i$  by using either  $(\uparrow)$  or  $(\uparrow')$ , and potentially several  $(\uparrow\uparrow)$  after that – let us say  $p \geq 0$  times. Therefore, if the claim is correct, the sum over  $\hat{M}_{i+1}$  should be exactly  $p + 1$ . Consider the operation by which  $v_{i+1}$  was added in the first place:

- ( $\uparrow$ )  $\mathcal{C}_\psi(v_{i+1})$  is reduced to a single element  $a$ , such that  $a >_{\text{lex}} \mathcal{C}_\psi(v_i)$ . Therefore, the index  $j$  of the first non-zero coefficient of  $M_{i+1}$  is ahead of the one of  $M_i$  so that the coefficient  $c$  of  $\ominus$  is equal to the  $j$ -th coefficient of  $M_{i+1}$  minus zero. Since the  $(\uparrow\uparrow)$  rule adds children to respect decreasing words, the  $p$  extra coefficients are added to the right of the  $j$ -th coefficient (including it) and therefore they are kept unchanged in the  $\ominus$  operation. Eventually, the sum over  $M_{i+1}$  is  $p + 1$  and so is the sum over  $\hat{M}_{i+1}$ .

( $\heartsuit$ )  $\mathcal{C}_\psi(v_{i+1})$  is built from  $\mathcal{C}_\psi(v_i)$  with Algorithm 2, and therefore they (i) share a common prefix, possibly empty and (ii) then differ by a single letter. The index of that letter in  $M_{i+1}$  corresponds to the index  $j$  defined in  $\ominus$ . Therefore, the coefficient  $\mathbf{c}$  is – before any ( $\heartsuit$ ) – equal to one. The argument of ( $\heartsuit$ ) letters being added to the right of  $j$  still hold and therefore the sum over  $\widehat{M}_{i+1}$  is also  $p + 1$ .

To conclude the proof, we have to exhibit the inverse function of the mapping we just defined. This will prove that this mapping is indeed a bijection, and then the theorem holds.

**Incremental adjacency matrix  $\rightarrow$  Reduced adjacency matrix** Let  $M$  and  $\widehat{M}$  be constructed as before. From  $\widehat{M}$ , we can define a matrix  $M'$  as:

$$\begin{cases} M'_1 = \widehat{M}_1 \\ M'_{i+1} = M'_i \oplus \widehat{M}_{i+1} \end{cases}$$

where the  $\oplus$  operation is defined as follow: given two words  $a_0 \cdots a_n$  and  $b_0 \cdots b_n$ , then denoting  $j = \min\{i : b_i \neq 0\}$ , and  $c = a_j + b_j$ ,

$$\begin{array}{r} \begin{array}{|c|c|c|} \hline a_0 \cdots a_{j-1} & a_j & a_{j+1} \cdots a_n \\ \hline \end{array} \\ \oplus \begin{array}{|c|c|c|} \hline b_0 \cdots b_{j-1} & b_j & b_{j+1} \cdots b_n \\ \hline \end{array} \\ \hline = \begin{array}{|c|c|c|} \hline a_0 \cdots a_{j-1} & \mathbf{c} & b_{j+1} \cdots b_n \\ \hline \end{array} \end{array} .$$

By construction,  $\oplus$  is the inverse operation of  $\ominus$ , so that we have the following lemma:

**Lemma A.1.** *The following properties hold:*

- $M_i \oplus (M_{i+1} \ominus M_i) = M_{i+1}$
- $(M_i \oplus \widehat{M}_{i+1}) \ominus M_i = \widehat{M}_{i+1}$

Therefore,  $M = M'$ .

The FDAG of Figure 3 is reproduced below to illustrates the stages of the proof. This FDAG is constructed in 7 steps, that are (in this order): ( $\uparrow$ ), ( $\heartsuit$ ), ( $\heartsuit$ ), ( $\uparrow$ ), ( $\heartsuit$ ), ( $\heartsuit$ ) and ( $\heartsuit$ ). The matrices  $A$ ,  $M$  and  $\widehat{M}$  are given in Figure 13. One can see that  $\widehat{M}$  is of size 7, as expected.

**Remark A.2.** *It should be noted that (general) Fishburn matrices, with at least one non-zero entry on each row and column, are in bijection with FDAG compressing forests made of a unique tree. Indeed, via the bijection above, as such FDAG have a unique root, it must be the last inserted vertex, and therefore, each column admits at least one non-zero entry (otherwise it would be another root).*

**Remark A.3.** *It is possible to enumerate row-Fishburn matrices by using the previous bijection and the FDAG enumeration tree together. Nevertheless, things are a little simpler in this case and the equivalent of the operations ( $\uparrow$ ) and ( $\heartsuit$ ) can be merged, giving two rules for matrix expansion:*

- (R1) Increase one coefficient to the (inclusive) left of the rightmost nonzero coefficient of the top row by 1.
- (R2) Increase the dimension of the matrix by 1 (to the left and top), all new coefficients set to zero. Set one coefficient of the top row to 1.

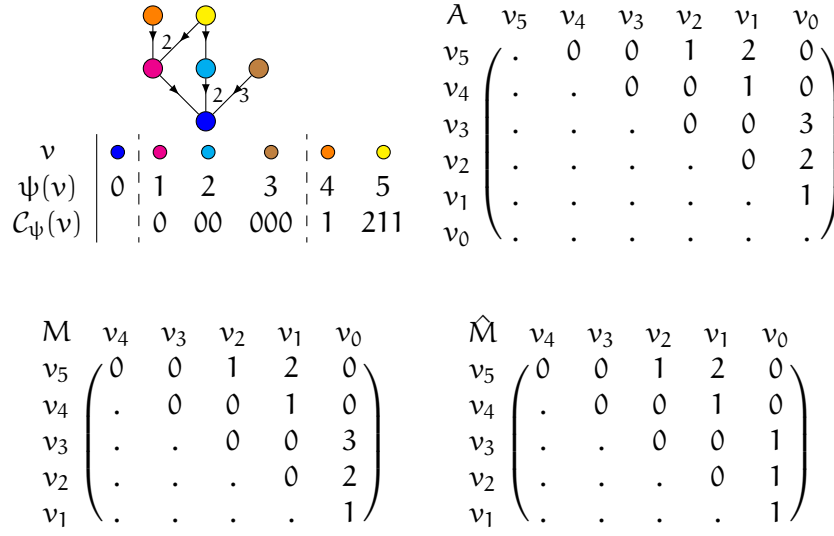


Figure 13: The FDAG of Figure 3 reproduced (top left), its adjacency matrix  $A$  (top right), its reduced adjacency matrix  $M$  (bottom left) and its incremental adjacency matrix  $\hat{M}$  (bottom right). Dots represent zeros corresponding to  $A_{i,j}$  elements with  $i \leq j$ .

## B Index of frequent notations

**Trees & DAG**  $v$  designates indifferently a vertex of a tree  $T$  or a DAG  $D$ .

$\mathcal{C}(v)$  *children* of  $v$ : all vertices connected to an arc leaving  $v$

$\text{deg}(v)$  *outdegree* of  $v$ : number of children of  $v$

$\mathcal{D}(v)$  *descendants* of  $v$ : children of  $v$ , their children, and so on

$\mathcal{H}(v)$  *height* of  $v$ : length of the longest path from  $v$  to a leaf

$\#T, \#D$  number of vertices

$T[v], D[v]$  *subtree/subDAG* rooted in  $v$  and composed of  $v$  and  $\mathcal{D}(v)$

$\mathcal{L}(T), \mathcal{L}(D)$  *leaves*: vertices without any children

$\text{deg}(T), \text{deg}(D)$  *outdegree*: maximum outdegree among all vertices

$\mathcal{S}(T)$  the set of all distinct subtrees of  $T$

$\mathcal{T}$  the set of all trees

$\mathcal{F}$  the set of all *forests*, i.e. sets of trees such that no tree is a subtree of another

**DAG reduction**  $D$  designates a FDAG,  $F$  a forest.

$\mathfrak{R}(F)$  DAG reduction of the forest  $F$

$\mathfrak{R}^{-1}(D)$  the forest  $F$  compressed by  $D$ , so that  $\mathfrak{R}(F) = D$

$\mathfrak{R}^{-1}(D[v])$  the tree  $T$  compressed by  $D[v]$ , so that  $\mathfrak{R}(\{T\}) = D[v]$

**Canonical FDAG** Let  $D$  be a fixed FDAG,  $v$  any vertex of  $D$  and  $v_n$  the vertex with highest index in the canonical ordering.

$\psi(\cdot)$  canonical topological ordering of  $D$

$\mathcal{C}_\psi(v)$   $\mathcal{C}(v)$  sorted by decreasing order on the indices defined by  $\psi(\cdot)$

$\mathcal{A}_=$  the indices of all vertices of  $D$  with same height as  $v_n$

$\mathcal{A}_<$  the indices of all vertices of  $D$  with strictly inferior height as  $v_n$

$\Lambda_<$  the set of all *decreasing words* on  $\mathcal{A}_<$ , i.e. where each letter is greater than or equal to those who follow, w.r.t. the lexicographical order

$\Lambda_<^{\bar{w}}$  the set of all decreasing words *bounded by*  $\bar{w}$ , i.e. all words in  $\Lambda_<$  that are greater than or equal to  $\bar{w}$ , w.r.t. the lexicographical order

$SC(w)$  *suffix-cut operator*: the word  $w$  deprived of its last letter

### Enumeration tree

$D_0$  the FDAG with one vertex and no arcs

$E_k$  the set of FDAG that are accessible in exactly  $k$  steps from  $D_0$  in the FDAG enumeration tree

**Forests of subtrees** Let  $\Delta$  be a subFDAG of  $D$ , and  $v$  any vertex. Let  $F = \mathfrak{R}^{-1}(D)$ .

$S(\Delta)$  *candidate vertices of*  $\Delta$ : the set of all vertices  $v'$  of  $D$  so that  $\Delta \cup \{v'\}$  is still a subFDAG of  $D$

$o(v)$  *origin of*  $v$ : the set of indices  $i$  so that  $\mathfrak{R}^{-1}(D[v])$  is a subtree of  $T_i \in F$ .

$o(\Delta)$  *origin of*  $\Delta$ : the set of indices  $i$  so that  $\mathfrak{R}^{-1}(\Delta)$  is a forest of subtrees of  $T_i \in F$ .