On the Count of Trees

Everardo Barcenas, Pierre Genevès, Nabil Layaïda, Alan Schmitt

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

On the Count of Trees

Everardo Bárcenas — Pierre Genevès — Nabil Layaïda — Alan Schmitt

N° 7251 — version 2

initial version April 2010 — revised version August 2010

Knowledge and Data Representation and Management

ISSN 0249-6399 ISRN INRIA/RR--7251--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

On the Count of Trees

Everardo Bárcenas , Pierre Genevès , Nabil Layaïda , Alan Schmitt

Theme : Knowledge and Data Representation and Management
Équipes-Projets WAM et SARDES

Rapport de recherche n° 7251 — version 2 — initial version April 2010 —
revised version August 2010 — 34 pages

Abstract: Regular tree grammars and regular path expressions constitute core constructs widely used in programming languages and type systems. Nevertheless, there has been little research so far on frameworks for reasoning about path expressions where node cardinality constraints occur along a path in a tree. We present a logic capable of expressing deep counting along paths which may include arbitrary recursive forward and backward navigation. The counting extensions can be seen as a generalization of graded modalities that count immediate successor nodes. While the combination of graded modalities, nominals, and inverse modalities yields undecidable logics over graphs, we show that these features can be combined in a decidable tree logic whose main features can be decided in exponential time. Our logic being closed under negation, it may be used to decide typical problems on XPath queries such as satisfiability, type checking with relation to regular types, containment, or equivalence.

Key-words: Modal Logic, XML, XPath, Schema

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

On the Count of Trees

Résumé : Ce document introduit une logique d'arbre décidable en temps exponentielle et qui est capable d'exprimer des contraintes de cardinalité sur chemins multidirectionnelle.

Mots-clés : Logique Modal, XML, XPath, Schema

1 Introduction

A fundamental peculiarity of XML is the description of regular properties. For example, in XML schema languages the content types of element definitions rely on regular expressions. In addition, selecting nodes in such constrained trees is also done by means of regular path expressions (à la XPath). In both cases, it is often interesting to be able to express conditions on the frequency of occurrences of nodes.

Even if we consider simple strings, it is well known that some formal languages easily described in English may require voluminous regular expressions. For instance, as pointed out in [13], the language L_{2a2b} of all strings over $\Sigma = \{a, b, c\}$ containing at least two occurrences of a *and* at least two occurrences of b requires a large expression, such as:

$$\begin{array}{ccccc} & \Sigma^*a\Sigma^*a\Sigma^*b\Sigma^*b\Sigma^* & \cup & \Sigma^*a\Sigma^*b\Sigma^*a\Sigma^*b\Sigma^* \\ \cup & \Sigma^*a\Sigma^*b\Sigma^*b\Sigma^*a\Sigma^* & \cup & \Sigma^*b\Sigma^*b\Sigma^*a\Sigma^*a\Sigma^* \\ \cup & \Sigma^*b\Sigma^*a\Sigma^*b\Sigma^*a\Sigma^* & \cup & \Sigma^*b\Sigma^*a\Sigma^*a\Sigma^*b\Sigma^*. \end{array}$$

If we add $\cap$ to the operators for forming regular expressions, then the language L_{2a2b} can be expressed more concisely as $(\Sigma^*a\Sigma^*a\Sigma^*) \cap (\Sigma^*b\Sigma^*b\Sigma^*)$. In logical terms, conjunction offers a dramatic reduction in expression size, which is crucial when the complexity of the decision procedure depends on formula size.

If we now consider a formalism equipped with the ability to describe numerical constraints on the frequency of occurrences, we get a second (exponential) reduction in size. For instance, the above expression can be formulated as $(\Sigma^*a\Sigma^*)^2 \cap (\Sigma^*b\Sigma^*)^2$. We can even write $(\Sigma^*a\Sigma^*)^{2^n} \cap (\Sigma^*b\Sigma^*)^{2^n}$ (for any natural n) instead of a (much) larger expression.

Different extensions of regular expressions with intersection, counting constraints, and interleaving have been considered over strings, and for describing content models of sibling nodes in XML type languages [4, 9, 15]. The complexity of the inclusion problem over these different language extensions and their combinations typically ranges from polynomial time to exponential space (see [9] for a survey). The main distinction between these works and the work presented here is that we focus on counting nodes located along deep and recursive paths in trees.

When considering regular *tree* languages instead of regular *string* languages, succinct syntax such as the one presented above is even more useful, as branching results in a higher combinatorial complexity. In the case of trees, it is often useful to express cardinality constraints not only on the sequence of children nodes, but also in a particular region of a tree, such as a subtree. Suppose, for instance, that we want to define a tree language over Σ where there is no more than 2 "b" nodes. This requires a quite large regular tree type expression such

as:

$$\begin{array}{lcl} x_{\text{root}} & \rightarrow & b[x_{b\leq 1}] \mid c[x_{b\leq 2}] \mid a[x_{b\leq 2}] \\ x_{b\leq 2} & \rightarrow & x_{\neg b}, b[x_{\neg b}], x_{\neg b}, b[x_{\neg b}], x_{\neg b} \mid x_{\neg b}, b[x_{b\leq 1}], x_{\neg b} \\ & & \mid x_{\neg b}, a[x_{b\leq 2}], x_{\neg b} \mid x_{\neg b}, c[x_{b\leq 2}], x_{\neg b} \mid x_{b\leq 1} \\ x_{b\leq 1} & \rightarrow & x_{\neg b} \mid x_{\neg b}, b[x_{\neg b}], x_{\neg b} \mid a[x_{b\leq 1}] \mid c[x_{b\leq 1}] \\ x_{\neg b} & \rightarrow & (a[x_{\neg b}] \mid c[x_{\neg b}])^* \end{array}$$

where x_{root} is the starting non-terminal; $x_{\neg b}, x_{b\leq 1}, x_{b\leq 2}$ are non-terminals; the notation $a[x_{\neg b}]$ describes a subtree whose root is labeled a and in which there is no b node; and "," is concatenation.

More generally, the widely adopted notations for regular tree grammars produce very verbose definitions for properties involving cardinality constraints on the nesting of elements[1].

The problem with regular tree (and even string) grammars is that one is forced to fully expand all the patterns of interest using concatenation, union, and Kleene star. Instead, it is often tempting to rely on another kind of (formal) notation that just describes a simple pattern and additional constraints on it, which are intuitive and compact with respect to size. For instance, one could imagine denoting the previous example as follows, where the additional constraint is described using XPath notation:

$$(x \rightarrow (a[x] \mid b[x] \mid c[x])^*) \ \wedge \ \text{count}(/\text{descendant-or-self::}b) \leq 2$$

Although this kind of counting operators does not increase the expressive power of regular tree grammars, it can have a drastic impact on succinctness, thus making reasoning over these languages harder (as noticed in [7] in the case of strings). Indeed, reasoning on this kind of extensions without relying on their expansion (in order to avoid syntactic blow-ups) is often tricky [8]. Determining satisfiability, containment, and equivalence over these classes of extended regular expressions typically requires involved algorithms with higher complexity [22] compared to ordinary regular expressions.

In the present paper, we propose a succinct logical notation, equipped with a satisfiability checking algorithm, for describing many sorts of cardinality constraints on the frequency of occurrence of nodes in regular tree types. Regular tree types encompass most of XML types (DTDs, XML Schemas, RelaxNGs) used in practice today.

XPath is the standard query language for XML documents, and it is an important part of other XML technologies such as XSLT and XQuery. XPath expressions are regular path expressions interpreted as sets of nodes selected from a given context node. One of the reasons why XPath is popular for web programming resides in its ability to express multidirectional navigation. Indeed, XPath expressions may use recursive navigation, to access descendant nodes, and also backward navigation, to reach previous siblings or ancestor

[1]This is typically the reason why the standard DTD for XHTML does not syntactically prevent the nesting of anchors, whereas this nesting is actually prohibited in the XHTML standard.

nodes. Expressing cardinality restrictions on nodes accessible by recursive multidirectional paths may introduce an extra-exponential cost [11, 27], or may even lead to undecidable formalisms [27, 6]. We present in this paper a decidable framework capable of succinctly expressing cardinality constraints along deep multidirectional paths.

A major application of this logical framework is the decision of problems found in the static analysis of programming languages manipulating XML data. For instance, since the logic is closed under negation, it can be used to solve subtyping problems such as XPath containment in the presence of tree constraints. Checking that a query q is contained in a query p with this logical approach amounts to verifying the validity of $q \Rightarrow p$, or equivalently, the unsatisfiability of $q \wedge \neg p$.

Contributions We extend a tree logic with a succinct notation for counting operators. These operators allow arbitrarily deep and recursive counting constraints. We present a sound and complete algorithm for checking satisfiability of logical formulas. We show that its complexity is exponential in the size of the succinct form.

Outline We introduce the logic in Section 2. Section 3 shows how the logic can be applied in the XML setting, in particular for the static analysis of XPath expressions and of common schemas containing constraints on the frequency of occurrence of nodes. The decision procedure and the proofs of soundness, completeness, and complexity are presented in Section 4. Finally, we review related work in Section 5 before concluding in Section 6.

2 Counting Tree Logic

We introduce our syntax for trees, define a notion of trails in trees, then present the syntax and semantics of logical formulas.

2.1 Trees

We consider finite trees which are node-labeled and sibling-ordered. Since there is a well-known bijective encoding between n–ary and binary trees, we focus on binary trees without loss of generality. Specifically, we use the encoding represented in Figure 1, where the binary representation preserves the first child of a node and append sibling nodes as second successors.

The structure of a tree is built upon modalities "$\triangledown$" and "$\triangleright$". Modality "$\triangledown$" labels the edge between a node and its first child. Modality "$\triangleright$" labels the edge between a node and its next sibling. Converse modalities "$\triangle$" and "$\triangleleft$" respectively label the same edges in the reverse direction.

We define a Kripke semantics for our tree logic, similar to the one of modal logics [29]. We write $M = \{\triangledown, \triangleright, \triangle, \triangleleft\}$ for the set of modalities. For $m \in M$ we denote by $\overline{m}$ the corresponding inverse modality ($\overline{\triangledown} = \triangle, \overline{\triangleright} = \triangleleft, \overline{\triangle} = \triangledown, \overline{\triangleleft} = \triangleright$).

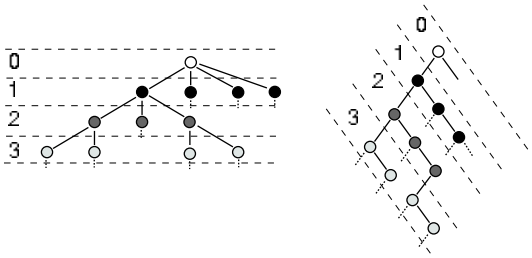

Figure 1: n–ary to binary trees

We also consider a countable alphabet P of *propositions* representing names of nodes. A node is always labeled with exactly one proposition.

A tree is defined as a tuple (N, R, L), where N is a finite set of nodes; R is a partial mapping from $N \times M$ to N that defines a tree structure;[2] and L is a labeling function from N to P.

2.2 Trails

Trails are defined as regular expressions formed by modalities, as follows:

$$\alpha ::= \alpha_0 \mid \alpha_0^\star \mid \alpha_0^\star, \alpha$$
$$\alpha_0 ::= m \mid \alpha_0, \alpha_0 \mid \alpha_0 \dashv \alpha_0$$

We restrict trails to sequences of repeated subtrails (which themselves contain no repetition) followed by a subtrail (with no repetition). Since we do not consider infinite paths, we also disallow trails where both a subtrail and its converse occurs under the scope of the recursion operator, thus ensuring cycle-freeness (see Section 2.5). These restrictions on trails allow us to prove the completeness of our approach while retaining the ability to express many counting formulas, such as the ones of XPath.

Trails are interpreted as sets of *paths*. A path, written ρ, is a sequence of modalities that belongs to the regular language denoted by the trail, written $\rho \in \alpha$.

In a given tree, we say that there is a *trail α from the node n_0 to the node n_k*, written $n_0 \xrightarrow{\alpha} n_k$, if and only if there is a sequence of nodes $n_0, \ldots, n_k$ and a path $\rho = m_1, \ldots, m_k$ such that $\rho \in \alpha$, and $R(n_j, m_{j+1}) = n_{j+1}$ for every $j = 0, \ldots, k-1$.

[2] For all $n, n' \in N, m \in M$, $R(n,m) = n' \iff R(n', \overline{m}) = n$; for all $n \in N$ except one (the *root*), exactly one of $R(n, \triangle)$ or $R(n, \triangleleft)$ is defined; for the root, neither $R(n, \triangle)$ nor $R(n, \triangleleft)$ is defined.

$\Phi \ni \phi$	::=			formula
		$\top \mid \neg\top$		true, false
	$\mid$	$p \mid \neg p$		atomic prop (negated)
	$\mid$	x		recursion variable
	$\mid$	$\phi \vee \phi$		disjunction
	$\mid$	$\phi \wedge \phi$		conjunction
	$\mid$	$\langle m\rangle\phi$	$\mid \neg\langle m\rangle\top$	modality (negated)
	$\mid$	$\langle\alpha\rangle_{\leq k}\psi$	$\mid \langle\alpha\rangle_{>k}\psi$	counting
	$\mid$	$\mu x.\psi$		fixpoint operator
ψ	::=			
		$\top \mid \neg\top$		
	$\mid$	$p \mid \neg p$		
	$\mid$	x		
	$\mid$	$\psi \vee \psi$		
	$\mid$	$\psi \wedge \psi$		
	$\mid$	$\langle m\rangle\psi$	$\mid \neg\langle m\rangle\top$	
	$\mid$	$\mu x.\psi$		

Figure 2: Syntax of Formulas (in Normal Form).

$$\neg\langle m\rangle\phi \equiv \neg\langle m\rangle\top \vee \langle m\rangle\neg\phi \qquad \neg\mu x.\psi \equiv \mu x.\neg\psi\{^{x}/_{\neg x}\}$$
$$\neg\langle\alpha\rangle_{\leq k}\psi \equiv \langle\alpha\rangle_{>k}\psi \qquad \neg\langle\alpha\rangle_{>k}\psi \equiv \langle\alpha\rangle_{\leq k}\psi$$

Figure 3: Reduction to Negation Normal Form.

2.3 Syntax of Logical Formulas

The syntax of logical formulas is given in Figure 2, where $m \in M$ and $k \in \mathbb{N}$. Formulas written ϕ may contain counting subformulas, whereas formulas written ψ cannot. We thus disallow counting under counting or under fixpoints. We also restrict formulas to *cycle-free* formulas, as detailed in Section 2.5. The syntax is shown in negation normal form. The negation of any closed formula (i.e., with no free variable) built using the syntax of Figure 2 may be transformed into negation normal form using the usual De Morgan rules together with rules given in Figure 3. When we write $\neg\phi$, we mean its negated normal form.

Defining an *equality* operator for counting formulas is straightforward using the other counting operators.

$$\langle\alpha\rangle_{=k}\psi \equiv \langle\alpha\rangle_{>(k-1)}\psi \wedge \langle\alpha\rangle_{\leq k}\psi \qquad \text{if } k > 0$$
$$\langle\alpha\rangle_{=0}\psi \equiv \langle\alpha\rangle_{\leq 0}\psi$$

$$
\begin{array}{lcl}
[\![\top]\!]^T_V & = & N \\
[\![\neg\top]\!]^T_V & = & \varnothing \\
[\![p]\!]^T_V & = & \{n \mid L(n) = p\} \\
[\![\neg p]\!]^T_V & = & \{n \mid L(n) \neq p\} \\
[\![x]\!]^T_V & = & V(x) \\
[\![\phi_1 \vee \phi_2]\!]^T_V & = & [\![\phi_1]\!]^T_V \cup [\![\phi_2]\!]^T_V \\
[\![\phi_1 \wedge \phi_2]\!]^T_V & = & [\![\phi_1]\!]^T_V \cap [\![\phi_2]\!]^T_V \\
[\![\langle m\rangle\phi]\!]^T_V & = & \{n \mid R(n,m) \in [\![\phi]\!]^T_V\} \\
[\![\neg\langle m\rangle\top]\!]^T_V & = & \{n \mid R(n,m) \text{ undefined}\} \\
[\![\langle\alpha\rangle_{\leq k}\psi]\!]^T_V & = & \{n \mid \ |\{n' \in [\![\psi]\!]^T_V \mid n \xrightarrow{\alpha} n'\}| \leq k\} \\
[\![\langle\alpha\rangle_{>k}\psi]\!]^T_V & = & \{n \mid \ |\{n' \in [\![\psi]\!]^T_V \mid n \xrightarrow{\alpha} n'\}| > k\} \\
[\![\mu x.\psi]\!]^T_V & = & \bigcap\{N' \mid [\![\psi]\!]^T_{V[N'/x]} \subseteq N'\}
\end{array}
$$

Figure 4: Semantics of Formulas.

2.4 Semantics of Logical Formulas

A formula is interpreted as a set of nodes in a tree. A model of a formula is a tree such that the formula denotes a non-empty set of nodes in this tree. A counting formula $\langle\alpha\rangle_{>k}\psi$ satisfied at a given node n means that there are at least $k+1$ nodes satisfying ψ that can be reached from n through the trail α. A counting formula $\langle\alpha\rangle_{>k}\psi$ is thus interpreted as the set of nodes such that, for each of them, the previously described condition holds. For example, the formula $p_1 \wedge \langle\triangledown\rangle\langle\triangleright^*\rangle_{>5}p_2$, denotes p_1 nodes with strictly more than 5 children nodes named p_2.

In order to present the formal semantics of formulas, we introduce valuations, written V, which relate variables to sets of nodes. We write $V[^{N'}/_x]$, where N' is a subset of the nodes, for the valuation defined as $V[^{N'}/_x](y) = V(y)$ if $x \neq y$, and $V[^{N'}/_x](x) = N'$. Given a tree $T = (N, R, L)$ and a valuation V, the formal semantics of formulas is given in Figure 4.

Note that the function $f : Y \to [\![\psi]\!]^T_{V[Y/x]}$ is monotone, and the denotation of $\mu x.\psi$ is a fixed point [26].

Intuitively, propositions denote the nodes where they occur; negation is interpreted as set complement; disjunction and conjunction are respectively set union and intersection; the least fixpoint operator performs finite recursive navigation; and the counting operator denotes nodes such that the ones accessible from this node through a trail fulfill a cardinality restriction. A logical formula is said to be *satisfiable* iff it has a model, i.e., there exists a tree for which the semantics of the formula is not empty.

2.5 Cycle-Freeness

Formal definition of cycle-freeness can be found in [10]. Intuitively, in a cycle-free formula, fixpoint variables must occur under a modality but cannot occur in the scope of both a modality and its converse. For instance, the formula $\mu x.\langle\triangledown\rangle x \vee \langle\triangle\rangle x$ is not cycle-free. In a cycle-free formula, the number of modality cycles (of the form $m\overline{m}$) is bound independently of the number of times fixpoints are unfolded (i.e., by replacing a fixpoint variable with the fixpoint itself). A fundamental consequence of the restriction to cycle-free formulas is that, when considering only finite trees, the interpretations of the greatest and smallest fixpoints coincide. This greatly simplifies the logic.

Here, we also restrict our approach to cycle-free formulas. We thus need to extend this notion to the counting operators, and more precisely to the trails that occur in them. Cycle-free trails are trails where both a subtrail and its converse do not occur under the scope of the recursion operator. We thus restrict the formulas under consideration to cycle-free formulas whose counting operators contain cycle-free trails.

Lemma 2.1. *Let ϕ be a cycle-free formula, and T be a tree for which $[\![\phi]\!]^T_\varnothing \neq \varnothing$. Then there is a finite unfolding ϕ' of the fixpoints of ϕ such that $[\![\phi'\{^{\neg\top}/_{\mu x.\psi}\}]\!]^T_\varnothing = [\![\phi]\!]^T_\varnothing$.*

Proof. As cycle-free counting formulas may be translated into (exponentially larger) cycle-free non-counting formulas, the proof is identical to the one in [10]. □

As a consequence, our logic is closed under negation even without greatest fixpoints.

2.6 Global Counting Formulas and Nominals

To conclude this section, we turn to an illustration of the expressive power of our logic. An interesting consequence of the inclusion of backward axes in trails is the ability to reach every node in the tree from any node of the tree, using the trail $(\triangle|\triangleleft)^\star, (\triangledown|\triangleright)^\star$.[3] We can thus select some nodes depending on some global counting property. Consider the following formula, where # stands for one of the comparison operators $\leq$, $>$, or $=$.

$$\langle(\triangle|\triangleleft)^\star, (\triangledown|\triangleright)^\star\rangle_{\#k}\phi_1$$

Intuitively, this formula counts how many nodes in the whole tree satisfy ϕ_1. For each node of the tree, it selects it if and only if the count is compatible with the comparison considered. The interpretation of this formula is thus either every node of the tree, or none. It is then easy to restrict the selected nodes to some that satisfy another formula ϕ_2, using intersection.

$$(\langle(\triangle|\triangleleft)^\star, (\triangledown|\triangleright)^\star\rangle_{\#k}\phi_1) \wedge \phi_2$$

[3] Note that this trail is cycle-free.

This formula select every node satisfying ϕ_2 if and only if there are $\#k$ nodes satisfying ϕ_1, which we write as follows.

$$\phi_1 \# k \implies \phi_2$$

We can now express existential properties, such as "select every node satisfying ϕ_2 if there exists a node satisfying ϕ_1".

$$\phi_1 > 0 \implies \phi_2$$

We can also express universal properties, such as "select every node satisfying ϕ_2 if every node satisfies ϕ_1".

$$(\neg\phi_1) \leq 0 \implies \phi_2$$

Another way to interpret global counting formulas is as a generalization of the so-called nominals in the modal logics community [24]. Nominals are special propositions whose interpretation is a singleton (they occur exactly once in the model). They come for free with the logic. A nominal, denoted "@n", corresponds to the following global counting formula:

$$[\langle (\triangle|\triangleleft)^\star, (\triangledown|\triangleright)^\star \rangle_{=1} n] \wedge n$$

where n is a new fresh atomic proposition.

One may need for nominals to occur in the scope of counting formulas. As we disallow counting under counting, we propose the following alternative encoding of nominals in these cases:

$$\begin{aligned} @n \equiv n \wedge \neg[&\text{descendant}(n) \vee \text{ancestor}(n) \vee \\ &\text{anc–or–self}(\text{siblings}(\text{desc–or–self}(n)))], \end{aligned}$$

where:

$$\begin{aligned} \text{descendant}(\psi) &= \langle\triangledown\rangle \mu x.\psi \vee \langle\triangledown\rangle x \vee \langle\triangleright\rangle x; \\ \text{foll–sibling}(\psi) &= \mu x.\langle\triangleright\rangle \psi \vee \langle\triangleright\rangle x; \\ \text{prec–sibling}(\psi) &= \mu x.\langle\triangleleft\rangle \psi \vee \langle\triangleleft\rangle x; \\ \text{desc–or–self}(\psi) &= \mu x.\psi \vee \langle\triangledown\rangle \mu y.x \vee \langle\triangleright\rangle y; \\ \text{ancestor}(\psi) &= \mu x.\langle\triangle\rangle(\psi \vee x) \vee \langle\triangleleft\rangle x; \\ \text{anc–or–self}(\psi) &= \mu x.\psi \vee \mu y.\langle\triangle\rangle(y \vee x) \vee \langle\triangleleft\rangle y; \\ \text{siblings}(\psi) &= \text{foll–sibling}(\psi) \vee \text{prec–sibling}(\psi). \end{aligned}$$

3 Application to XML Trees

3.1 XPath Expressions

XPath [3] was introduced as part of the W3C XSLT transformation language to have a non-XML format for selecting nodes and computing values from an XML

document (see [10] for a formal presentation of XPath). Since then, XPath has become part of several other standards, in particular it forms the "navigation subset" of the XQuery language.

In their simplest form XPath expressions look like "directory navigation paths". For example, the XPath

```
/company/personnel/employee
```

navigates from the root of a document through the top-level "company" node to its "personnel" child nodes and on to its "employee" child nodes. The result of the evaluation of the entire expression is the set of all the "employee" nodes that can be reached in this manner. At each step in the navigation, the selected nodes for that step can be filtered with a predicate test. Of special interest to us are the predicates that count nodes or that test the position of the selected node in the previous step's selection. For example, if we ask for

```
/company/personnel/employee[position()=2]
```

then the result is *all* employee nodes that are the *second* employee node (in document order) among the employee child nodes of each personnel node selected by the previous step.

XPath also makes it possible to combine the capability of searching along "axes" other than the shown "children of" with counting constraints. For example, if we ask for

```
/company[count(descendant::employee)<=300]/name
```

then the result consists of the company names with less than 300 employees in total (the axis "descendant" is the transitive closure of the default – and often omitted – axis "child").

The syntax and semantics of Core XPath expressions are respectively given on Figure 5 and Figure 6. An XPath expression is interpreted as a relation between nodes. The considered XPath fragment allows absolute and relative paths, path union, intersection, composition, as well as node tests and qualifiers with counting operators, conjunction, disjunction, negation, and path navigation. Furthermore, it supports all XPath axes allowing multidirectional navigation.

It was already observed in [11, 27] that using positional information in paths reduces to counting (at the cost of an exponential blow-up). For example, the expression

```
child::a[position()=5]
```

first selects the "`a`" nodes occurring as children of the current context node, and then keeps those occurring at the 5th position. This expression can be rewritten into the semantically equivalent expression:

```
child::a[count(preceding-sibling::a)=4]
```

$$
\begin{aligned}
\text{Axis} &::= \text{self} \mid \text{child} \mid \text{parent} \mid \text{descendant} \mid \text{ancestor} \mid \\
&\quad \text{following-sibling} \mid \text{preceding-sibling} \mid \\
&\quad \text{following} \mid \text{preceding} \\
\text{NameTest} &::= \text{QName} \mid * \\
\text{Step} &::= \text{Axis::NameTest} \\
\text{PathExpr} &::= \text{PathExpr/PathExpr} \mid \text{PathExpr[Qualifier]} \mid \text{Step} \\
\text{Qualifier} &::= \text{PathExpr} \mid \text{CountExpr} \mid \text{not Qualifier} \mid \\
&\quad \text{Qualifier and Qualifier} \mid \text{Qualifier or Qualifier} \mid @n \\
\text{CountExpr} &::= \text{count}(\text{PathExpr}') \text{ Comp } k \\
\text{PathExpr}' &::= \text{PathExpr}'/\text{PathExpr}' \mid \text{PathExpr}'[\text{Qualifier}'] \mid \text{Step} \\
\text{Qualifier}' &::= \text{PathExpr}' \mid \text{not Qualifier}' \mid \text{Qualifier}' \text{ and Qualifier}' \\
&\quad \mid \text{Qualifier}' \text{ or Qualifier}' \mid @n \\
\text{Comp} &::= \leq \mid > \mid \geq \mid < \mid = \\
\text{XPath} &::= \text{PathExpr} \mid /\text{PathExpr} \mid \text{XPath union PathExpr} \mid \\
&\quad \text{XPath intersect PathExpr} \mid \text{XPath except PathExpr}
\end{aligned}
$$

Figure 5: Syntax of Core XPath Expressions.

which constraints the number of preceding siblings named "a" to 4, so that the qualifier becomes true only for the 5th child "a". A general translation of positional information in terms of counting operators [11, 27] is summarized on Figure 7, where $\ll$ denotes the document order (depth-first left-to-right) relation in a tree. Note that translated path expressions can in turn be expressed into the core XPath fragment of Figure 5 (at the cost of another exponential blow-up). Indeed, expressions like $\text{PathExpr}/(\text{PathExpr}_2 \text{ except } \text{PathExpr}_3)/\text{PathExpr}_4$ must be rewritten into expressions where binary connectives for paths occur only at top level, as in:

$$
\begin{aligned}
&\text{PathExpr}/\text{PathExpr}_2/\text{PathExpr}_4 \text{ except} \\
&\text{PathExpr}/\text{PathExpr}_3/\text{PathExpr}_4
\end{aligned}
$$

We focus on Core XPath expressions involving the counting operator (see Figure 5). The XPath fragment without the counting operator (the navigational fragment) was already linearly translated into μ-calculus in [10]. The contributions presented in this paper allow to equip this navigational fragment with counting features such as the ones formulated above. Logical formulas capture the aforementioned XPath counting constraints. For example, consider the following XPath expression:

```
child::a[count(descendant::b[parent::c])>5]
```

$$
\begin{aligned}
[\![\text{Axis::NameTest}]\!] &= \{(x,y) \in N^2 \mid x(\text{Axis})y \text{ and } \\
&\quad\ y \text{ satisfies NameTest}\} \\
[\![\text{/PathExpr}]\!] &= \{(r,y) \in [\![\text{PathExpr}]\!] \mid \\
&\quad\ r \text{ is the root}\} \\
[\![P_1/P_2]\!] &= [\![P_1]\!] \circ [\![P_2]\!] \\
[\![P_1 \text{ union } P_2]\!] &= [\![P_1]\!] \cup [\![P_2]\!] \\
[\![P_1 \text{ intersect } P_2]\!] &= [\![P_1]\!] \cap [\![P_2]\!] \\
[\![P_1 \text{ except } P_2]\!] &= [\![P_1]\!] \setminus [\![P_2]\!] \\
[\![\text{PathExpr[Qualifier]}]\!] &= \{(x,y) \in [\![\text{PathExpr}]\!] \mid \\
&\quad\ y \in [\![\text{Qualifier}]\!]_{\text{Qualif}}\}
\end{aligned}
$$

$$
\begin{aligned}
[\![\text{PathExpr}]\!]_{\text{Qualif}} &= \{x \mid \exists y.(x,y) \in [\![\text{PathExpr}]\!]\} \\
[\![\text{count(PathExpr) Comp } k]\!]_{\text{Qualif}} &= \{x \in N \mid \\
&\quad\ |\{y \mid (x,y) \in [\![\text{PathExpr}]\!]\}| \\
&\quad\ \text{satisfies Comp } k\} \\
[\![\text{not } Q]\!]_{\text{Qualif}} &= N \setminus [\![Q]\!]_{\text{Qualif}} \\
[\![Q_1 \text{ and } Q_2]\!]_{\text{Qualif}} &= [\![Q_1]\!]_{\text{Qualif}} \cap [\![Q_1]\!]_{\text{Qualif}} \\
[\![Q_1 \text{ or } Q_2]\!]_{\text{Qualif}} &= [\![Q_2]\!]_{\text{Qualif}} \cup [\![Q_2]\!]_{\text{Qualif}}
\end{aligned}
$$

Figure 6: Semantics of Core XPath Expressions

$$
\begin{aligned}
\text{PathExpr[position() = 1]} &\equiv \text{PathExpr except (PathExpr/} \ll) \\
\text{PathExpr[position() = } k+1\text{]} &\equiv (\text{PathExpr intersect} \\
&\quad\ (\text{PathExpr}[k]/\ll))[\text{position()} = 1] \\
\ll &\equiv (\text{descendant::*}) \text{ union (a-o-s::*/} \\
&\quad\ \text{following-sibling::*/d-or-s::*)} \\
\text{a-or-s::*} &\equiv \text{ancestor::* union self::*} \\
\text{d-or-s::*} &\equiv \text{descendant::* union self::*}
\end{aligned}
$$

Figure 7: Positional Information as Syntactic Sugars [11, 27]

Path	Logical formula
γ/self::*	γ
γ/child::*	$\langle \triangleleft^*, \triangle \rangle \gamma$
γ/parent::*	$\langle \triangledown \rangle \langle \triangleright^* \rangle \gamma$
γ/descendant::*	$\langle (\triangleleft \mid \triangle)^*, \triangle \rangle \gamma$
γ/ancestor::*	$\langle \triangledown \rangle \langle (\triangledown \mid \triangleright)^* \rangle \gamma$
γ/following-sibling::*	$\langle \triangleleft \rangle \langle \triangleleft^* \rangle \gamma$
γ/preceding-sibling::*	$\langle \triangleright \rangle \langle \triangleright^* \rangle \gamma$

Figure 8: XPath axes as modalities over binary trees.

This expression selects the children nodes named "`a`" provided they have more than 5 descendants which (1) are named "`b`" and (2) whose parent is named "`c`". The logical formula denoting the set of children nodes named "`a`" is:

$$\psi = a \wedge \langle \triangleleft^*, \triangle \rangle \top$$

The logical translation of the above XPath expression is:

$$\psi \wedge \langle \triangledown \rangle \langle (\triangledown | \triangleright)^* \rangle_{>5} (b \wedge \mu x. \langle \triangle \rangle c \vee \langle \triangleleft \rangle x)$$

This formula holds for nodes selected by the XPath expression. A correspondence between the main XPath axes over unranked trees and modal formulas over binary trees is given in Figure 8. In this figure, each logical formula holds for nodes selected by the corresponding XPath axis from a context γ.

Let consider another example (XPath expression e_1):

```
child::a/child::b[count(child::e/descendant::h)>3]
```

Starting from a given context in a tree, this XPath expression navigates to children nodes named "a" and selects their children named "b". Finally, it retains only those "b" nodes for which the qualifier between brackets holds. The first path can be translated in the logic as follows:

$$\vartheta = b \wedge \mu x. \langle \triangle \rangle (a \wedge \mu x'. \langle \triangle \rangle \top \vee \langle \triangleleft \rangle x') \vee \langle \triangleleft \rangle x$$

The counting part requires a more sophisticated translation in the logic. This is because it makes implicit that "e" nodes (whose existence is simply tested for counting purposes) must be children of selected "b" nodes. The translation of the full aforementioned XPath expression is as follows:

$$\vartheta \wedge @n \wedge \langle (\triangle \mid \triangleleft)^*, (\triangledown \mid \triangleright)^* \rangle_{>3} \eta$$

where $@n$ is a new fresh nominal used to mark a "b" node which is filtered by the qualifier and the formula η describes the counted "h" nodes:

$$\eta = h \wedge \mu x. \langle \triangle \rangle (e \wedge \mu x'. \langle \triangle \rangle @n \vee \langle \triangleleft \rangle x') \vee \langle \triangleleft \rangle x \vee \langle \triangle \rangle x$$

Intuitively, the general idea behind the translation is to first translate the leading path, use a fresh nominal for marking a node which is filtered, then find at least "3" instances of "h" nodes from which we can reach back the marked node via the inverse path of the counting formula.

Since trails make it possible to navigate but not to test properties (like existence of labels), we test for labels in the counted formula η and we use a general navigation $(\triangle \mid \triangleleft)^*, (\triangledown \mid \triangleright)^*$ to look for counted nodes everywhere in the tree. Introducing the nominal is necessary to bind the context properly (without loss of information). Indeed, the XPath expression e_1 makes implicit that a "e" node must be a child of a "b" node selected by the outer path. Using a nominal, we restore this property by connecting the counted nodes to the initial single context node.

Lemma 3.1. *The translation of Core XPath expressions with counting constraints into the logic is linear.*

It is proven by structural induction in a similar manner to [10] (in which the translation is proven for expressions without counting constraints). For counting formulas, the use of nominals and the general (constant-size) counting trail make it possible to avoid duplication of trails so that the translation remains linear.

We can now address several decision problems such as equivalence, containment, and emptiness of XPath expressions. These decision problems are reduced to test satisfiability for the logic (in the manner of [10]). We present in Section 4 a satisfiability testing algorithm with a single exponential complexity with respect to the formula size.

In [10], it was show the logic is also able to capture XML schema languages. This allows to test the XPath decision problems in the presence of XML types. We now show our logic can also succinctly express cardinality constraints on XML types.

3.2 Regular Tree Languages with Cardinality Constraints

Regular tree grammars capture most of the schemas in use today [23]. The logic can express all regular tree languages (it is easy to prove that regular expression types in the manner of e.g., [14] can be linearly translated into the logic: see [10]).

In practice, schema languages often provide shorthands for expressing cardinality constraints on node occurrences. XML Schema notably offers two attributes *minOccurs* and *maxOccurs* for this purpose. For instance, the following XML schema definition:

```
<xsd:element name="a">
  <xsd:complexType>
    <xsd:sequence>
     <xsd:element name="b" minOccurs="4" maxOccurs="9"/>
    </xsd:sequence>
  </xsd:complexType>
</xsd:element>
```

is a notation that restricts the number of occurrences of "b" nodes to be at least 4 and at most 9, as children of "a" nodes. The goal here is to have a succinct notation for expressing regular languages which could otherwise be exponentially large if written with usual regular expression operators. The above regular requirement can be translated as the formula:

$$\phi \wedge \langle \triangledown \rangle (\langle \triangleright^\star \rangle_{>3} b \wedge \langle \triangleright^\star \rangle_{\leq 9} b)$$

where ϕ corresponds to the regular tree type $a[b^*]$ as follows:

$$\begin{aligned} \phi = \quad & (a \wedge (\neg\langle \triangledown \rangle \top \vee \langle \triangledown \rangle \psi)) \wedge \neg\langle \triangleright \rangle \top \\ \psi = \quad & \mu x.\, (b \wedge \neg\langle \triangledown \rangle \top \wedge \neg\langle \triangleright \rangle \top) \vee (b \wedge \neg\langle \triangledown \rangle \top \wedge \langle \triangleright \rangle x) \end{aligned}$$

This example only involves counting over children nodes. The logic allows counting through more general trails, and in particular arbitrarily deep trails. Trails corresponding to the XPath axes "preceding, ancestor, following" can be used to constrain the context of a schema. The "descendant" trail can be used to specify additional constraints over the subtree defined by a given schema. For instance, suppose we want to forbid webpages containing nested anchors "a" (whose interpretation makes no sense for web browsers). We can build the logical formula f which is the conjunction of a considered schema for webpages (e.g. XHTML) with the formula a/descendant::a in XPath notation. Nested anchors are forbidden by the considered schema iff f is unsatisfiable.

As another example, suppose we want paragraph nodes ("p" nodes) not to be nested inside more than 3 unordered lists ("ul" nodes), regardless of the schema defining the context. One may check for the unsatisfiability of the following formula:

$$p \wedge \langle (\triangle | \triangleleft)^\star, \triangle \rangle_{>3} ul$$

4 Satisfiability Algorithm

We present a tableau-based algorithm for checking satisfiability of formulas. Given a formula, the algorithm seeks to build a tree containing a node selected by the formula. We show that our algorithm is correct and complete: a satisfying tree is found if and only if the formula is satisfiable. We also show that the time complexity of our algorithm is exponential in the size of the formula.

4.1 Overview

The algorithm operates in two stages.

First, a formula ϕ is decomposed into a set of subformulas, called the *lean*. The lean gathers all subformulas that are useful for determining the truth status of the initial formula, while eliminating redundancies. For instance, conjunctions and disjunctions are eliminated at this stage. More precisely, the lean (defined in 4.2) mainly gathers atomic propositions and modal subformulas. From the lean, one may gather a finite number of formulas, called a ϕ-node, which may

be satisfied at a given node of a tree. Trees of ϕ–nodes represent the exhaustive search universe in which the algorithm is looking for a satisfying tree.

The second stage of the algorithm consists in the building of sets of such trees in a bottom-up manner, ensuring consistency at each step. Initially, all possible leaves (i.e., ϕ–node that do not require children nodes) are considered. During further steps, the algorithm considers every possible ϕ–node that can be connected with a tree of the previous steps, checking for consistency. For instance, if a formula at a ϕ–node n involve a forward modality $\langle\triangledown\rangle\phi'$, then ϕ' must be verified at the first child of n. Reciprocally, due to converse modalities, a ϕ–node may impose restrictions on its possible parent nodes. The new trees that are built may involve converse modalities, which will be satisfied during further steps of the algorithm. To ensure the algorithm terminates, a bound on the number of times each ϕ–node may occur in the tree is given.

Finally, the algorithm terminates whenever:

- either a tree that satisfies the initial formula has been found, and its root does not contain any pending (unproven) backward modality; or
- every tree has been considered (the exploration of the whole search universe is complete): the formula is unsatisfiable.

4.2 Preliminaries

To track where counting formulas are satisfied, we annotate each one with a fresh *counting proposition* c, yielding formulas of the form $\langle\alpha\rangle^{c}_{\#k}\phi$. To define the notions of lean and ϕ–nodes, we need to extract navigating formulas from counting formulas (Figure 9).

We now define the *Fisher-Ladner* relation to extract subformulas. In the following, i ranges over 1 and 2.

$$R^{fl}(\phi_1 \wedge \phi_2, \phi_i), \qquad R^{fl}(\phi_1 \vee \phi_2, \phi_i),$$
$$R^{fl}(\mu x.\phi, \phi[^{\mu x.\phi}/_x]), \qquad R^{fl}(\langle\alpha\rangle^{c}_{\#k}\psi, nav(\langle\alpha\rangle^{c}_{\#k}\psi)),$$
$$R^{fl}(\langle m\rangle\phi, \phi).$$

The *Fisher-Ladner* closure of a formula ϕ, written $FL(\phi)$, is the set defined as follow.

$$\begin{array}{lcl} FL(\phi)_0 & = & \{\phi\}, \\ FL(\phi)_{i+1} & = & FL(\phi)_i \cup \{\phi' \mid R^{fl}(\phi'', \phi'), \phi'' \in FL(\phi)_i\}, \\ FL(\phi) & = & FL(\phi)_k, \end{array}$$

where k is the smallest integer s.t. $FL(\phi)_k = FL(\phi)_{k+1}$. Note that this set is finite since only one expansion of a fixpoint formula is required in order to produce all its subformulas in the closure.

The *lean* of a formula ϕ is a set of formulas containing navigating formulas of the form $\langle m\rangle\top$, every navigating formulas of the form $\langle m\rangle\psi$ (i.e., that do

$$
\begin{array}{rr}
nav(x) = x & nav(p) = p \\
nav(\top) = \top & nav(c) = c \\
nav(\neg p) = \neg p & nav(\neg\langle m\rangle\top) = \neg\langle m\rangle\top
\end{array}
$$

$$
\begin{aligned}
nav(\phi_1 \wedge \phi_2) &= nav(\phi_1) \wedge nav(\phi_2) \\
nav(\phi_1 \vee \phi_2) &= nav(\phi_1) \vee nav(\phi_2) \\
nav(\langle m\rangle\phi) &= \langle m\rangle nav(\phi) \\
nav(\mu x.\psi) &= \mu x.nav(\psi) \\
nav(\langle\alpha\rangle^c_{>k}\psi) &= nav((\alpha), \psi \wedge c) \\
nav(\langle\alpha\rangle^c_{\leq k}\psi) &= nav((\alpha), (\psi \wedge c) \vee (\neg\psi \wedge \neg c)) \\
nav((\epsilon), \psi) &= \psi \\
nav((m), \psi) &= \langle m\rangle\psi \\
nav((\alpha_1, \alpha_2), \psi) &= nav((\alpha_1), nav((\alpha_2), \psi)) \\
nav((\alpha_1 \mid \alpha_2), \psi) &= nav((\alpha_1), \psi) \vee nav((\alpha_2), \psi) \\
nav((\alpha^\star), \psi) &= \mu x.nav(\psi) \vee nav((\alpha), x)
\end{aligned}
$$

Figure 9: Navigation extraction from counting formulas

not contain counting formulas) from $FL(\phi)$, every proposition occurring in ϕ, written P_ϕ, every counting proposition, written C, and an extra proposition that does not occur in ϕ used to represent other names, written $p_{\overline{\phi}}$.

$$lean(\phi) = \{\langle m\rangle\top\} \cup \{\langle m\rangle\psi \in FL(\phi)\} \cup P_\phi \cup C \cup \{p_{\overline{\phi}}\}$$

A *ϕ–node* , written n^ϕ, is a subset of $lean(\phi)$, such that:

- exactly one proposition from $P_\phi \cup \{p_{\overline{\phi}}\}$ is present;
- when $\langle m\rangle\psi$ is present, then $\langle m\rangle\top$ is present; and
- both $\langle\triangle\rangle\top$ and $\langle\triangleleft\rangle\top$ cannot be present at the same time.

The set of ϕ–nodes is defined as N^ϕ.

Intuitively, a node n^ϕ corresponds to a formula.

$$n^\phi = \bigwedge_{\psi\in n^\phi} \psi \wedge \bigwedge_{\psi\in lean(\phi)\setminus n^\phi} \neg\psi$$

When the formula ϕ under consideration is fixed, we often omit the superscript.

A *ϕtree* is either the empty tree $\varnothing$, or a triple $(n^\phi, \Gamma_1, \Gamma_2)$ where Γ_1 and Γ_2 are ϕtrees. When clear from the context, we usually refer to ϕtrees simply as trees.

$$\frac{}{n \vdash^{\phi} \top} \qquad \frac{\psi \in n}{n \vdash^{\phi} \psi} \qquad \frac{\psi \notin n}{n \vdash^{\phi} \neg\psi} \qquad \frac{n \vdash^{\phi} \psi_1 \qquad n \vdash^{\phi} \psi_2}{n \vdash^{\phi} \psi_1 \wedge \psi_2} \qquad \frac{n \vdash^{\phi} \psi_1}{n \vdash^{\phi} \psi_1 \vee \psi_2}$$

$$\frac{n \vdash^{\phi} \psi_2}{n \vdash^{\phi} \psi_1 \vee \psi_2} \qquad \frac{n \vdash^{\phi} \psi\{^{\mu x.\psi}/_{x}\}}{n \vdash^{\phi} \mu x.\psi}$$

Figure 10: Local entailment relation: between nodes and formulas

We now turn to the definition of consistency of a ϕtree. To this end, we define an entailment relation between a node and a formula in Figure 10.

Two nodes n_1 and n_2 are consistent under modality $m \in \{\triangledown, \triangleright\}$, written $R^{\phi}(n_1, m) = n_2$, iff

$$\forall \langle m \rangle \psi \in lean(\phi), \langle m \rangle \psi \in n_1 \iff n_2 \vdash^{\phi} \psi$$
$$\forall \langle \overline{m} \rangle \psi \in lean(\phi), \langle \overline{m} \rangle \psi \in n_2 \iff n_1 \vdash^{\phi} \psi$$

Consistency is checked each time a node is added to the tree, ensuring that forward modalities of the node are indeed satisfied by the nodes below, and that pending backward modalities of the node below are consistent with the added node. Note that counting formulas are not considered at this point, as they are globally verified in the next step.

Upon generation of a finished tree, i.e., a tree with no pending backward modality, one may check whether a node of this tree satisfies ϕ. To this end, we first define forward navigation in a ϕtree Γ. Given a path consisting of forward modalities ρ, $\Gamma(\rho)$ is the node at that path. It is undefined if there is no such node.

$$(n, \Gamma_1, \Gamma_2)(\epsilon) = n$$
$$(n, \Gamma_1, \Gamma_2)(\triangledown\rho) = \Gamma_1(\rho)$$
$$(n, \Gamma_1, \Gamma_2)(\triangleright\rho) = \Gamma_2(\rho)$$

We also allow extending the path with backward modalities if they match the last modality of the path.

$$(n, \Gamma_1, \Gamma_2)(\rho \triangledown \triangle) = (n, \Gamma_1, \Gamma_2)(\rho)$$
$$(n, \Gamma_1, \Gamma_2)(\rho \triangleright \triangleleft) = (n, \Gamma_1, \Gamma_2)(\rho)$$

Now, we are able to define an entailment relation along paths in ϕtrees in Figure 11. This relation extends local entailment relation (Figure 10) with checks for counting formulas. Note that the case for fixpoints is contained in the case for formulas with no counting subformula. In the "less than" case, we need to make sure that every node reachable through the trail is taken into account, either as counted if it satisfies ψ, or not counted otherwise (in this case, $\neg\psi$ denotes the negation normal form).

$$\frac{\phi' \text{ does not contain counting formulas} \qquad \Gamma(\rho) \vdash^{\phi} \phi'}{\rho \vdash^{\phi}_{\Gamma} \phi'} \qquad \frac{\rho \vdash^{\phi}_{\Gamma} \phi_1 \qquad \rho \vdash^{\phi}_{\Gamma} \phi_2}{\rho \vdash^{\phi}_{\Gamma} \phi_1 \wedge \phi_2}$$

$$\frac{\rho \vdash^{\phi}_{\Gamma} \phi_1}{\rho \vdash^{\phi}_{\Gamma} \phi_1 \vee \phi_2} \qquad \frac{\rho \vdash^{\phi}_{\Gamma} \phi_2}{\rho \vdash^{\phi}_{\Gamma} \phi_1 \vee \phi_2} \qquad \frac{\rho m \vdash^{\phi}_{\Gamma} \phi'}{\rho \vdash^{\phi}_{\Gamma} \langle m \rangle \phi'}$$

$$\frac{|\{n',\ \rho' \in \alpha \wedge \Gamma(\rho\rho') = n' \wedge n' \vdash^{\phi} \psi \wedge c\}| > k}{\rho \vdash^{\phi}_{\Gamma} \langle \alpha \rangle^{c}_{>k} \psi}$$

$$\frac{\begin{array}{c}|\{n',\ \rho' \in \alpha \wedge \Gamma(\rho\rho') = n' \wedge n' \vdash^{\phi} \psi \wedge c\}| \leq k \\ \forall \rho' \in \alpha, \Gamma(\rho\rho') \vdash^{\phi} (\psi \wedge c) \vee (\neg\psi \wedge \neg c)\end{array}}{\rho \vdash^{\phi}_{\Gamma} \langle \alpha \rangle^{c}_{\leq k} \psi}$$

Figure 11: Global entailment relation (incl. counting formulas)

We conclude these preliminaries by introducing some final notations. The *root* of a ϕtree is defined as follows.

$$\begin{aligned} root(\varnothing) &= \varnothing \\ root((n, \Gamma_1, \Gamma_2)) &= n \end{aligned}$$

A ϕtree Γ *satisfies* a formula ϕ, written $\Gamma \vdash \phi$, if neither $\langle \triangle \rangle \top$ nor $\langle \triangleleft \rangle \top$ occur in $root(\Gamma)$, and if there is a path ρ such that $\rho \vdash^{\phi}_{\Gamma} \phi$. A set of trees ST *satisfies* a formula ϕ, written $ST \vdash \phi$, when there is a tree $\Gamma \in ST$ such that $\Gamma \vdash \phi$.

4.3 The Algorithm

We are now ready to present the algorithm, which is parameterized by $K(\phi)$ (defined in Figure 12), the maximum number of occurrences of a given node in a path from the root of the tree to a leaf. The algorithm builds consistent candidate trees from the bottom up, and checks at each step if one of the built tree satisfies the formula, returning 1 if it is the case. As the set of nodes from which to build the trees is finite, it eventually stops and returns 0 if no satisfying tree has been found.

To bound the size of the trees that are built, we restrict the number of identical nodes on a path from the root to any leaf by $K(\phi) + 2$, defined in Figure 12, using `nmax` defined as follows.

$$\begin{aligned} \texttt{nmax}(n, \Gamma_1, \Gamma_2) &= \texttt{max}(\texttt{nmax}(n, \Gamma_1), \texttt{nmax}(n, \Gamma_2)) \\ \texttt{nmax}(n, (n, \Gamma_1, \Gamma_2)) &= 1 + \texttt{nmax}(n, \Gamma_1, \Gamma_2) \\ \texttt{nmax}(n, (n', \Gamma_1, \Gamma_2)) &= \texttt{nmax}(n, \Gamma_1, \Gamma_2) \quad \text{if } n \neq n' \\ \texttt{nmax}(n, \varnothing) &= 0 \end{aligned}$$

Algorithm 1 Check Satisfiability of ϕ

$ST \leftarrow \varnothing$
repeat
  $AUX \leftarrow \{(n, \Gamma_1, \Gamma_2) \mid$ {we extend the trees}
    $\mathtt{nmax}(n, \Gamma_1, \Gamma_2) \leq K(\phi) + 2$ {with an available node}
    for i in $\triangledown, \triangleright$ {and each child is either}
    $\Gamma_i = \varnothing$ and $\langle i \rangle \top \notin n$ {an empty tree}
    or $\Gamma_i \in ST$ {or a previously built tree}
      $\langle \overline{i} \rangle \top \in root(\Gamma_i)$ {with pending backward modalities}
      $R^{\phi}(n, i) = root(\Gamma_i)\}$ {checking consistency}
  if $AUX \subseteq ST$ **then**
    return 0 {No new tree was built}
  end if
  $ST \leftarrow ST \cup AUX$
until $ST \vdash \phi$
return 1

$$K(p) = K(\neg p) = K(\neg\langle m \rangle \top) = K(\top) = K(\mu x.\psi) = 0$$
$$K(\phi_1 \wedge \phi_2) = K(\phi_1 \vee \phi_2) = K(\phi_1) + K(\phi_2)$$
$$K(\langle m \rangle \phi) = K(\phi)$$
$$K(\langle \alpha \rangle_{\#k} \psi) = k + 1$$

Figure 12: Occurrences bound

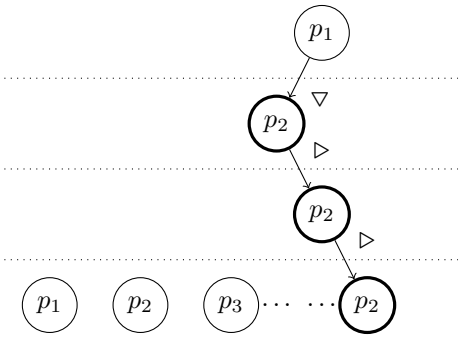

Figure 13: Checking $\phi = p_1 \wedge \langle \bigtriangledown \rangle \langle \triangleright^\star \rangle_{>2} p_2$

Consider for instance the formula $\phi = p_1 \wedge \langle \bigtriangledown \rangle \langle \triangleright^\star \rangle_{>2} p_2$. The computed lean is as follows, where $\psi = \mu x.(p_2 \wedge c) \vee \langle \triangleright \rangle x$.

$$\{p_1, p_2, p_3, c, \langle \bigtriangledown \rangle \top, \langle \triangleright \rangle \top, \langle \bigtriangleup \rangle \top, \langle \triangleleft \rangle \top, \langle \bigtriangledown \rangle \psi, \langle \triangleright \rangle \psi\}$$

Names other than p_1 and p_2 are represented by p_3; c identifies counted nodes. Computing the bound on nodes, we get $K(\phi) = 3$.

After the first step, ST consists of the trees $(\{p_i\}, \varnothing, \varnothing)$, $(\{p_i, c\}, \varnothing, \varnothing)$, $(\{p_i, \langle \overline{j} \rangle \top\}, \varnothing, \varnothing)$, and $(\{p_i, c, \langle \overline{j} \rangle \top\}, \varnothing, \varnothing)$ with $i \in \{1, 2, 3\}$ and $j \in \{\bigtriangledown, \triangleright\}$. At this point the three finished trees in ST are tested and found not to satisfy ϕ.

After the second iteration many trees are created, but the one of interest is the following.

$$T_0 = (\{p_2, c, \langle \triangleright \rangle \top, \langle \triangleleft \rangle \top, \langle \triangleright \rangle \psi\}, \varnothing, (\{p_2, c, \langle \triangleleft \rangle \top\}, \varnothing, \varnothing))$$

The third iteration yields the following tree.

$$T_1 = (\{p_2, c, \langle \triangleright \rangle \top, \langle \bigtriangleup \rangle \top, \langle \triangleright \rangle \psi\}, \varnothing, T_0)$$

We can conclude by the fourth iteration when we find the tree $(\{p_1, \langle \bigtriangledown \rangle \psi, \langle \bigtriangledown \rangle \top\}, T_1, \varnothing)$, which is found to satisfy ϕ at path ϵ. As the nodes at every step are different, the limit is not reached. Figure 13 depicts a graphical representation of the example where counted nodes (containing c) are drawn as thick circles.

4.4 Termination

Proving termination of the algorithm is straightforward, as only a finite number of trees may be built and the algorithm stops as soon as it cannot build a new tree.

4.5 Soundness

If the algorithm terminates with a candidate, we show that the initial formula is satisfiable. Let Γ and ρ be the ϕtree and path such that $\rho \vdash^{\phi}_{\Gamma} \phi$. We build

a tree from Γ and show that the interpretation of ϕ for this tree includes the node at path ρ.

We write $T(\Gamma)$ for the tree (N, R, L) defined as follows. We first rewrite Γ such that each node n is replaced by the path to reach it (i.e, nodes are identified by their path).

$$\begin{aligned} path(n, \Gamma_1, \Gamma_2) &\to (\epsilon, path(\triangledown, \Gamma_1), path(\triangleright, \Gamma_2)) \\ path(\rho, (n, \Gamma_1, \Gamma_2)) &\to (\rho, path(\rho\triangledown, \Gamma_1), path(\rho\triangleright, \Gamma_2)) \\ path(\rho, \varnothing) &\to \varnothing \end{aligned}$$

We then define:

- $N = nodes(path(\Gamma))$;
- for every $(\rho, \Gamma_1, \Gamma_2)$ in $path(\Gamma)$ and $i = \triangledown, \triangleright$, if $\Gamma_i \neq \varnothing$ then $R(\rho, i) = \rho i$ and $R(\rho i, \bar{i}) = \rho$; and
- for all $\rho \in N$ if $p \in \Gamma(\rho)$ then $L(\rho) = p$.

Lemma 4.1. *Let ψ a subformula of ϕ with no counting formula. If $\Gamma(\rho) \vdash^{\phi} \psi$ then we have $\rho \in [\![\psi]\!]_{\varnothing}^{T(\Gamma)}$.*

Proof. We proceed by induction on the lexical ordering of the number of unfolding of ψ that are required for $T(\Gamma)$ as defined by Lemma 2.1, and of the size of the formula.

The base cases are $\top$, atomic or counting propositions, and negated forms. These are immediate by definition of $[\![\psi]\!]_{\varnothing}^{T(\Gamma)}$. The cases for disjunction and conjunction are immediate by induction (the formula is smaller). The case for fixpoints is also immediate by induction, as the number of unfoldings required decreases, and as $[\![\mu x.\psi]\!]_{\varnothing}^{T(\Gamma)} = [\![\psi\{^{\mu x.\psi}/_{x}\}]\!]_{\varnothing}^{T(\Gamma)}$.

The last case is the presence of a modality $\langle m\rangle\psi$ from the ϕnode $\Gamma(\rho)$. In this case we rely on the fact that the nodes $\Gamma(\rho m)$ and $\Gamma(\rho)$ are consistent to derive $\Gamma(\rho m) \vdash^{\phi} \psi$. We then conclude by induction as the formula is smaller. □

Theorem 4.2 (Soundness)**.** *If $\rho \vdash_{\Gamma}^{\phi} \phi$ then $\rho \in [\![\phi]\!]_{\varnothing}^{T(\Gamma)}$*

Proof. The proof proceeds by induction on the derivation of $\rho \vdash_{\Gamma}^{\phi} \phi$. Most cases are immediate (or rely on Lemma 4.1). For the "greater than" counting case, we rely on the $k+1$ selected nodes that have to satisfy $\psi \wedge c$ thus ψ. In addition, in the "less than" case, every node that is not counted has to satisfy $\neg\psi \wedge \neg c$, so in particular $\neg\psi$. In both cases we conclude by induction. □

4.6 Completeness

Our proof proceeds in two step. We build a ϕtree that satisfies the formula, then we show it is actually built by the algorithm. As the proof is quite complex, we devote some space to detail it.

Assume that formula ϕ is satisfiable by a tree T. We consider the smallest such tree (i.e., the tree with the fewest number of nodes) and fix $n^\star$, a node witnessing satisfiability.

We now build a ϕtree homomorphic to T, called the lean labeled version of ϕ, written $\Gamma(T, \phi)$. To this end, we start by annotating counted nodes along with their corresponding counting proposition, yielding a new tree T_c. Starting from $n^\star$ and by induction on ϕ, we proceed as follows. For formulas with no counting subformula, including recursion, we stop. For conjunction and disjunction of formulas, we recursively annotate according to both subformulas. For modalities, we recursively annotate from the node under the modality. For $\langle\alpha\rangle^c_{\leq k}\psi$, we annotate every selected node with the counting proposition corresponding to the formula. For $\langle\alpha\rangle^c_{>k}\psi$, we annotate exactly $k+1$ selected nodes.

We now extend the semantics of formulas to take into account counting propositions and annotated nodes, written $[\![\cdot]\!]^{T_c}_V$. The definition is identical to Figure 4, with one addition and two changes. The addition is for counting propositions, which we define as $n \in [\![c]\!]^{T_c}_V$ iff n is annotated by c. The two changes are for counting propositions, which we define as follows, where we select only nodes that are annotated.

$$[\![\langle\alpha\rangle_{\leq k}\phi']\!]^{T_c}_V = \{n, |\{n' \in [\![\phi']\!]^{T_c}_V \cap [\![c]\!]^{T_c}_V, n \xrightarrow{\alpha} n'\}| \leq k\}$$
$$[\![\langle\alpha\rangle_{>k}\phi']\!]^{T_c}_V = \{n, |\{n' \in [\![\phi']\!]^{T_c}_V \cap [\![c]\!]^{T_c}_V, n \xrightarrow{\alpha} n'\}| > k\}$$

We show that this modification of the semantics does no change the satisfiability of the formula.

Lemma 4.3. *We have* $n^\star \in [\![\phi]\!]^{T_c}_\varnothing$.

Proof. We proceed by recursion on the derivation $n^\star \in [\![\phi]\!]^T_\varnothing$. The cases where no counting formula is involved, thus including fixpoints, are immediate, as the selected nodes are identical. The disjunction, conjunction, and modality cases are also immediate by induction. The interesting cases are the counting formulas.

For $\langle\alpha\rangle^c_{>k}\psi$, as there are exactly $k+1$ nodes annotated, the property is true by induction. For $\langle\alpha\rangle^c_{\leq k}\psi$, we rely on the fact that every counted node is annotated. We conclude by remarking that ψ does not contain a counting formula, thus we have $[\![\psi]\!]^{T_c}_V = [\![\psi]\!]^T_V$ and $[\![\neg\psi]\!]^{T_c}_V = [\![\neg\psi]\!]^T_V$. □

To every node n, we associate n^ϕ, the largest subset of formulas of the lean selecting the node.

$$n^\phi = \{\phi_0 \mid n \in [\![\phi_0]\!]^T_\varnothing, \phi_0 \in \mathit{lean}(\phi)\}$$

This is a ϕ-node as it contains one and exactly one proposition, and if it includes a modal formula $\langle m\rangle\psi$, then it also includes $\langle m\rangle\top$. The tree $\Gamma(T, \phi)$ is then built homomorphically to T.

In the remainder of this section, we write Γ for $\Gamma(T, \phi)$. We now check that Γ is consistent, starting with local consistency.

$$\frac{\psi \in lean(\phi)}{\psi \dot{\in} lean(\phi)} \qquad \frac{\psi_1 \dot{\in} lean(\phi) \qquad \psi_2 \dot{\in} lean(\phi)}{\psi_1 \wedge \psi_2 \dot{\in} lean(\phi)}$$

$$\frac{\psi_1 \dot{\in} lean(\phi) \qquad \psi_2 \dot{\in} lean(\phi)}{\psi_1 \vee \psi_2 \dot{\in} lean(\phi)} \qquad \frac{}{\top \dot{\in} lean(\phi)} \qquad \frac{\psi \in (P_\phi \cup \langle m \rangle \top \cup C)}{\neg\psi \dot{\in} lean(\phi)}$$

Figure 14: Formula induced by a lean

In the following, we say a formula ψ is induced by the lean of ϕ, written $\psi \dot{\in} lean(\phi)$, if it consists of the boolean combination of subformulas from the lean as defined in Figure 14.

Lemma 4.4. *Let $\langle m \rangle \psi$ be a formula in $lean(\phi)$, and let ψ' be ψ after unfolding its fixpoint formulas not under modalities. We have $\psi' \dot{\in} lean(\phi)$.*

Proof. By definition of the lean and of the $\dot{\in}$ relation. □

Lemma 4.5. *Let ψ be a formula induced by $lean(\phi)$. We have $n \in [\![\psi]\!]^{T_c}_{\varnothing}$ if and only if $n^\phi \vdash^\phi \psi$.*

Proof. We proceed by induction on ψ. The base cases (the formula is in the ϕ-node or is a negation of a lean formula not in the ϕ-node) hold by definition of n^ϕ. The inductive cases are straightforward as these formulas only contain fixpoints under modalities. □

Lemma 4.6. *Let n_1 and n_2 such that $R(n_1, m) = n_2$ with $m \in \{\triangledown, \triangleright\}$. We have $R^\phi(n_1^\phi, m) = n_2^\phi$.*

Proof. Let $\langle m \rangle \psi$ be a formula in $lean(\phi)$. We show that $\langle m \rangle \psi \in n_1^\phi \iff n_2^\phi \vdash^\phi \psi$. We have $\langle m \rangle \psi \in n_1^\phi$ if and only if $n_1 \in [\![\langle m \rangle \psi]\!]^{T_c}_{\varnothing}$ by definition of n_1^ϕ, which in turn holds if and only if $n_2 = R(n_1, m) \in [\![\psi]\!]^{T_c}_{\varnothing}$. We now consider ψ' which is ψ after unfolding its fixpoint formulas not under modalities. We have $[\![\psi']\!]^{T_c}_{\varnothing} = [\![\psi]\!]^{T_c}_{\varnothing}$ and we conclude by Lemmas 4.4 and 4.5. □

We now turn to global consistency, taking counting formulas into account.

Lemma 4.7. *Let ϕ_s be a subformula of ϕ, and ρ be a path from the root in T such that $T(\rho) \in [\![\phi_s]\!]^{T_c}_{\varnothing}$. We then have $\rho \vdash^\phi_\Gamma \phi_s$.*

Proof. We proceed by induction on ϕ_s.

If ϕ_s does not contain any counting formula, we consider ϕ'_s which is ϕ_s after unfolding its fixpoint formulas not under modalities. We have $[\![\phi'_s]\!]^{T_c}_{\varnothing} = [\![\phi_s]\!]^{T_c}_{\varnothing}$ and $\phi'_s \dot{\in} lean(\phi)$. We conclude by Lemma 4.5.

For most inductive cases, the proof is immediate by induction, as the formula size decreases.

For $\langle\alpha\rangle^c_{>k}\psi$, we have by induction for every counted node $\rho\rho' \vdash^\phi_\Gamma \psi$ and $\rho\rho' \vdash^\phi_\Gamma c$. We conclude by the conjunction rule and by the counting rule of Figure 11.

For $\langle\alpha\rangle^c_{\leq k}\psi$, we proceed as above for the counted nodes. For the nodes that are not counted, we have $\Gamma(\rho\rho') \vdash^\phi \neg\psi$ by Lemma 4.5 (since $\neg\psi \mathrel{\dot\in} lean(\phi)$). We conclude by remarking that the node is not annotated by c, hence $\Gamma(\rho\rho') \vdash^\phi \neg c$. □

We next show that the ϕtree Γ is actually built by the algorithm. The proof follows closely the one from [10], with a crucial exception: we need to make sure there are enough instances of each formula. Indeed, in [10], the algorithm uses a ϕtype (a subset of $lean(\phi)$) at most once on each branch from the root to a leaf of the built tree. This yields a simple condition to stop the algorithm and conclude the formula is unsatisfiable. However, in the presence of counting formulas, a given ϕtype may occur more than once on a branch. To maintain the termination of the algorithm, we bound the number of identical ϕtype that may be needed by $K(\phi)$ as defined in Figure 12. We thus need to check that this bound is sufficient to build a tree for any satisfiable formula.

We recall that ϕ is a satisfiable formula and T is a smallest tree such that ϕ is satisfied, and $n^\star$ is a witness of satisfiability.

We proceed in two steps: first we show that counted nodes (with counted propositions) imply a bound on the number of identical ϕtypes on a branch for a smallest tree. Second, we show that this minimal marking is bound by $K(\phi)$.

In the following, we call counted nodes and node $n^\star$ *annotations*. We define the *projection* of an annotation on a path. Let ρ be a path from the root of the tree to a leaf. An annotation projects on ρ at ρ_1 if $\rho = \rho_1\rho_2$, the annotation is at $\rho_1\rho_m$, and ρ_2 shares no prefix with ρ_m.

Lemma 4.8. *Let Γ' be the annotated tree, ρ a path from the root of the tree to a leaf, n_1 and n_2 two distinct nodes of ρ such that $n_1^\phi = n_2^\phi$. Then either annotations projects both on ρ at n_1 and n_2, or an annotation projects strictly between n_1 and n_2.*

Proof. We proceed by contradiction: we assume there is no annotation that projects between n_1 and n_2 and at most one of them has an annotation that projects on it. Without loss of generality, we assume that n_2 is below n_1 in the tree.

Assume neither n_1 nor n_2 is annotated (through projection). We consider the tree Γ_s where n_2 is "grafted" upon n_1. Formally, let ρ_1 be the path to n_1 and $\rho_1\rho_2$ the path to n_2. We remove every node whose path is of the form $\rho_1\rho_3$ where ρ_2 is not a prefix of ρ_3, and we also remove node n_2. The mapping R' from nodes and modalities to nodes is the same as before for the node that are kept except for n_1, where $R'(n_1, \triangledown) = R(n_2, \triangledown)$ and $R'(n_1, \triangleright) = R(n_2, \triangleright)$. For every path ρ of Γ, let ρ_s be the potentially shorter path if it exists (i.e., if it was not removed when pruning the tree). More precisely, if $\rho' = \rho'_1\rho'_3$ where ρ'_1 is a prefix of ρ_1 and the paths are disjoint from there, then $\Gamma_s(\rho') = \Gamma(\rho')$. If $\rho' = \rho_1\rho_2\rho_3$, then $\Gamma_s(\rho_1\rho_3) = \Gamma(\rho')$.

We now show that Γ_s still satisfies ϕ at $n^\star$, a contradiction since this tree is strictly smaller than Γ.

First, as there was no annotation projected, $n^\star$ is still part of this tree at a path ρ_s. We show that we have $\rho_s \vdash^{\phi}_{\Gamma_s} \phi$ by induction on the derivation $\rho \vdash^{\phi}_{\Gamma} \phi$. Let $\rho' \vdash^{\phi}_{\Gamma} \phi'$ in the derivation, assuming that ρ'_s is defined.

The case where ϕ' does not mention any counting formula is trivial: $\Gamma(\rho') = \Gamma_s(\rho'_s)$ thus local entailment is immediate.

Conjunction and disjunction are also immediate by induction.

We now turn to the modality case, $\langle m\rangle\phi'$ where ϕ' contains a counting formula. If ρ' is neither ρ_1 nor $\rho_1\rho_2$, we deduce from the fact that ρ'_s is defined that $(\rho' m)_s$ is also defined and we conclude by induction. We now assume that ρ' is either ρ_1 or $\rho_1\rho_2$ and find a contradiction. First, remark that $\rho' \vdash^{\phi}_{\Gamma} \langle m\rangle\phi'$ implies that the navigation generated by $\langle m\rangle\phi'$ is in $\Gamma(\rho_1) = \Gamma(\rho_1\rho_2)$. As each syntactic occurrence of a counting formula mentions a distinct counting proposition c, this is possible only if the counting formula is under a fixpoint or under another counting formula, both of which are impossible.

We finally turn to the counting case $\langle\alpha\rangle^{c}_{\#k}\psi$. We say that a path *does not cross over* when this path does not contain n_1 nor n_2. For nodes that are reached using paths that do not cross over, we conclude by induction that they are also counted. We show that the remaining nodes reached through a crossover remain reachable (there cannot be any counted node in the part of the tree that is removed since counted nodes are annotated and there was no annotation in the part removed). Without loss of generality, assume that ρ' is a prefix of ρ_1 (the counting formula is in the "top" part of the tree), and let ρ_n be the path from the counting formula to the counted node (ρ_n is an instance of the trail α). This path is of the shape $\rho'_1\rho_2\rho_c$, with $\rho_1 = \rho'\rho'_1$. We now show that the path $\rho'_1\rho_c$ is an instance of α if and only if ρ_n is an instance of the trail, thus the same node is still reached.

Recall that α is of the shape $\alpha_1, \ldots, \alpha_n, \alpha_{n+1}$ where α_1 to α_n are of the form $\alpha^\star_{r_i}$ and where α_{n+1} does not contain a repeated trail. We say that a prefix ρ_p of a path ρ *stops at* i if there is a suffix ρ_s such that $\rho_p\rho_s$ is still a prefix of ρ, $\rho_p\rho_s \in \alpha_1, \ldots, \alpha_i$, and there is no shorter suffix ρ'_s and j such that $\rho_p\rho'_s \in \alpha_1, \ldots, \alpha_j$. (Intuitively, α_i is the trail being used when matching the end of ρ_p.) If there are several satisfying indices i, we consider the smallest.

We first show that a counting proposition is necessarily mentioned in a formula of n^ϕ_2, by contradiction. Assume no counting proposition is mentioned, yet the counting crossed-over. This can only occur for a "less than" counting formula that reaches n_2 which is not counted (because the formula was false), and if there is no path whose ρ_n is a strict prefix that is an instance of α (otherwise, by definition of the lean and of *nav* (Figure 9), a formula of the form $nav((\alpha'), (\psi \wedge c) \vee (\neg\psi \wedge \neg c))$ would be true and thus would be present, contradicting the assumption that no counting proposition is mentioned). Since $n^\phi_1 = n^\phi_2$, the same is true for n^ϕ_1, a direct contradiction to the fact that n_2 is also reached by the trail. Thus counting propositions are mentioned in n^ϕ_1 and n^ϕ_2.

We next show that there are $i \leq j \leq n$ such that both ρ'_1 stops at i and $\rho'_1\rho_2$ stop at j, i.e., neither i nor j may be $n+1$. Recall that α_{n+1} does not contain a repeated subtrail. Thus every formula of n_2^ϕ mentioning c is of the form $nav((\alpha'), \psi)$, where α' does not contain a repetition. We consider the largest such formula. Since n_1 is before n_2 in the path from the counting node to the counted node, a similar formula with a larger trail or with a repetition must occur in n_1^ϕ, contradicting $n_1^\phi = n_2^\phi$.

Consider next the suffixes ρ_s^1 and ρ_s^2 computed when stating that the paths stop at i and j. These suffixes correspond to the path matching the end of α_i and α_j, respectively (before the next iteration or switching to the next subtrail). They have matching formulas in n_1^ϕ and n_2^ϕ. As the formulas are present in both nodes, then the remainder of the paths ($\rho_2\rho_c$ and ρ_c) are instances of $(\rho_s^1|\rho_s^2)\alpha_i \dots \alpha_{n+1}$, thus $\rho'_1\rho_c$ is an instance of α if and only if ρ_n is.

In the case of "greater than" counting, we conclude immediately by induction as the same nodes are selected (thus there are enough). In the case of "less than", we need to check that no new node is counted in the smaller tree. Assume it is not the case for the formula $\langle\alpha\rangle_{\leq k}\psi$, thus there is a path $\rho_n \in \alpha$ to a node satisfying ψ. As the same node can be reached in Γ, and as we have $\Gamma(\rho'\rho_n) \vdash^\phi \neg\psi$ by induction, we have a contradiction.

This concludes the proof when neither n_1 nor n_2 is annotated. The proof is identical when n_2 is annotated. If n_1 is annotated, we look at the first modality between n_1 and n_2. If it is a $\triangledown$, then we build the smaller tree by doing $R'(n_1, \triangledown) = R(n_2, \triangledown)$ (we remove the $\triangleright$ subtree from n_2 instead of n_1). Symmetrically, if the first modality is a $\triangleright$, we consider $R'(n_1, \triangleright) = R(n_2, \triangleright)$ as smaller tree. The rest of the proof proceeds as above. □

Theorem 4.9 (Completeness)**.** *If ϕ is satisfiable, then a satisfying tree is built.*

Proof. The proof proceeds as in [10], we only need to check there are enough copies of each node to build every path. Let ρ be a path from the root of the tree to the leaves. By Lemma 4.8, there are at most $n+1$ identical nodes in this path, where n is the number of annotations. The number of annotations is $c+1$ where c is the number of counted nodes. We show by an immediate induction on the formula ϕ that c is bound by $K(\phi)$ as defined in Figure 12. We conclude by remarking that $K(\phi) + 2$ is the number of identical nodes we allow in the algorithm. □

4.7 Complexity

We now show that the time complexity of the satisfiability algorithm is exponential in the formula size. This is achieved in two steps: we first show that the lean size is linear in the formula size, then we show that the algorithm has a single exponential complexity with relation to the lean size.

Lemma 4.10. *The lean size is linear in terms of the original formula size.*

Proof Sketch. First note that the size of the lean is the number of elements it contains; the size of each element does not matter.

It was shown in [10] that the size of the lean generated by a non-counting formula is linear with respect to the formula size.

We now describe the case for counting formulas. The lean consists of propositions and of modal subformulas, including the ones generated by the navigation of counting formulas (Figure 9). Moreover, each counting formula adds one fresh counting proposition. In the case of "less than" formulas $\langle\alpha\rangle_{\leq k}\psi$, a duplication occurs due to the consideration of the negated normal form of ψ. Since there is no counting under counting, this duplication and the fact that the negated normal form of a formula is linear in the size of the original formula (Figure 3) result in the lean remaining linear. Another duplication occurs in the case of counting formulas of the form $\langle\alpha_1|\alpha_2\rangle_{\#k}\psi$. This duplication does not double the size of the lean, however, since ψ still occurs only once in the lean, thus the number of elements in the lean induced by $nav((\alpha_1),\psi) \vee nav((\alpha_2),\psi)$ is the same as the sum of the ones in $nav((\alpha_1),\psi)$ and in $nav((\alpha_2),\cdot)$. □

Theorem 4.11. *The satisfiability algorithm for the logic is decidable in time* $2^{O(n)}$*, where* n *is the size of the lean.*

Proof Sketch. The maximum number of considered nodes is the number of distinct tree nodes which is 2^n, the number of subsets of the lean. For a given formula ϕ, the number of occurrences of the same node in the tree is bounded by $K(\phi) \leq k * m$, where k is the greatest constant occurring in the counting formulas and m is the number of counting subformulas of ϕ. Hence the number of steps of the algorithm is bounded by $2^n * k * m$.

At each iteration, the main operation performed by the algorithm is the composition of trees stored in AUX. The cost of each iteration consists in: the different searches needed to form the necessary triples (n, Γ_1, Γ_2), the `nmax` function and R^ϕ. Since the total number of nodes is exponential, and the number of different subtrees too, therefore the maximum number of newly formed trees (triples) at each step has also an exponential bound. The function `nmax` performs a single traversal of the tree which is also exponential. Since the entailment relation involved in the definition of R^ϕ is local, R^ϕ is performed in linear time. Computing the containment $AUX \subseteq ST$ and the union $ST \cup AUX$ are linear operations over sets of exponential size.

The stop condition of the algorithm is checked by the global entailment relation. It involves traversals parametrized by the number of trees, the number of nodes in each tree, the number of traversals for the entailment relation of counting formulas, and $K(\phi)$. Its time complexity is bounded by $(2^n * k * m)^3$.

Hence, the total time complexity of the algorithm is bounded by $(2^n * k * m)^{k'}$, for some constant k'. □

5 Related Work

Counting over trees The notion of Presburger Automata for trees, combining both regular constraints on the children of nodes and numerical constraints given by Presburger formulas, has independently been introduced by Dal Zilio and Lugiez [5] and Seidl et al. [25]. Specifically, Dal Zilio and Lugiez [5] propose a modal logic for unordered trees called Sheaves logic. This logic allows to impose certain arithmetical constraints on children nodes but lacks recursion (i.e., fixpoint operators) and inverse navigation. Dal Zilio and Lugiez consider the satisfiability and the membership problems. Demri and Lugiez [6] showed by means of an automata-free decision procedure that this logic is only PSPACE-complete. Restrictions like p_1 *nodes have no more "children" than* p_2 *nodes*, are succinctly expressible by this approach. Seidl et al. [25] introduce a fixpoint Presburger logic, which, in addition to numerical constraints on children nodes, also supports recursive forward navigation. For example, expressions like *the descendants of* p_1 *nodes have no more "children" than the number of children of descendants of* p_2 *nodes* are efficiently represented. This means that constraints can be imposed on sibling nodes (even if they are deep in the tree) by forward recursive navigation but not on distant nodes which are not siblings.

Compared to the work presented here, neither of the two previous approaches can efficiently support constraints like *there are more than* 5 *ancestors of "p" nodes*.

Furthermore, due to the lack of backward navigation, the works found in [5, 25, 6] are not suited for succinctly capturing XPath expressions. Indeed, it is well-known that expressions with backward modalities are exponentially more succinct than their forward-only counterparts [11, 29].

There is poor hope to push the decidability envelope much further for counting constraints. Indeed, it is known from [16, 6, 27] that the equivalence problem is undecidable for XPath expressions with counting operators of the form:

- $\text{PathExpr}_1[\text{count}(\text{PathExpr}_2) = \text{count}(\text{PathExpr}_3)]$, or
- $\text{PathExpr}_1[\text{position}() = \text{count}(\text{PathExpr}_2)]$.

This is the reason why logical frameworks that allow comparisons between counting operators limit counting by restricting the PathExpr to immediate children nodes [5, 25]. In this paper, we chose a different tradeoff: comparisons are restricted to constants but at the same time comparisons along more general paths are permitted.

Counting over graphs The μ-calculus is a propositional modal logic augmented with least and greatest fixpoint operators [18]. Kupferman, Sattler and Vardi study a μ-calculus with graded modalities where one can express, e.g., that a graph node has at least n successors satisfying a certain property [19]. The modalities are limited in scope since they only count immediate successors of a given node. A similar notion in trees consists in counting immediate children nodes, as performed by the counting formula $\langle\triangledown\rangle\langle\triangleright^*\rangle_{\#k}\phi$, where ϕ

describes the property to be counted. Compared to graded modalities of [19], we consider trees and we can extend the "immediate successor" notion to nodes reachable from regular paths, involving reverse and recursive navigation.

A recent study [2] focuses on extending the μ-calculus with inverse modalities [29], nominals [24], and graded modalities of [19]. If only two of the above constructs are considered, satisfiability of the enriched calculus is EXPTIME-complete [2, 1]. However, if all of the above constructs are considered simultaneously, the calculus becomes undecidable [2]. The present work shows that this undecidability result in the case of graphs does not preclude decidable tree logics combining such features.

XPath-like counting extensions The proposed logic can be the target for the compilation of a few more sophisticated counting features, considered as syntactic sugars (and that may come at the potential extra cost of their translation).

In particular, XPath allows nested counting, as in the expression

$$\text{self::book}[\text{chapter}[\text{section} > 1] > 1,$$

which selects the current "book" node provided it has at least two "chapter" child nodes which in turn must contain at least two "section" nodes each. For a simple set of formulas, formulas that count only on children nodes, such nesting can be translated into ordinary logical formulas. For instance, the logical formulation of the above XPath expression can be captured as follows:

$$\text{book} \wedge \langle \triangledown \rangle \mu x.\, (\text{chapter} \wedge \psi \wedge \langle \triangleright \rangle \mu y.\text{chapter} \wedge \psi \vee \langle \triangleright \rangle y) \vee \langle \triangleright \rangle x$$

where $\psi = \langle \triangledown \rangle \mu x.(\text{section} \wedge \langle \triangleright \rangle \mu y.\text{section} \vee \langle \triangleright \rangle y) \vee \langle \triangleright \rangle x$.

In [21], Marx introduced an "until" operator for extending XPath's expressive power to be complete with respect to first-order logic over trees. This operator is trivially expressible in the present logic, owing to the use of the fixpoint binder. We can even combine counting features with the "until" operator and express properties that go beyond the expressive power of the XPath standard. For instance, the following formula states that "*starting from the current node, and until we reach an ancestor named a, every ancestor has at least 3 children named b*":

$$\mu x.\, (\langle \triangledown \rangle \langle \triangleright^* \rangle_{>2} b \wedge \mu y. \langle \triangle \rangle x \vee \langle \triangleleft \rangle y) \vee a$$

These extensions come at an extra cost, however. It is not difficult to observe (by induction) that, given a formula ϕ with subformulas $\psi_1, ..., \psi_n$ counting only on children nodes, if formulas $\psi_1, ..., \psi_n$ are replaced by their expansions in ϕ, yielding a formula ϕ', then $|lean(\phi')| \leq |lean(\phi)| * k^l$, where k is greatest numerical constraint of the counting subformulas, and l is the greatest level nesting of counting subformulas. As a consequence of Theorem 4.11, the logic extended with nested formulas counting on children nodes and formulas counting on children nodes under the scope of a fixpoint operator can be decided in time $2^{O(n*k^l)}$.

6 Conclusion

We introduced a modal logic of trees equipped with (1) converse modalities, which allow to succinctly express forward and backward navigation, (2) a least fixpoint operator for recursion, and (3) cardinality constraint operators for expressing numerical occurrence constraints on tree nodes satisfying some regular properties. A sound and complete algorithm is presented for testing satisfiability of logical formulas. This result is surprising since the corresponding logic for graphs is undecidable [2].

The decision procedure for the logic is exponential time w.r.t. to the formula size. The logic captures regular tree languages with cardinality restrictions, as well as the navigational fragment of XPath equipped with counting features. Similarly to backward modalities, numerical constraints do not extend the logical expressivity beyond regular tree languages. Nevertheless they enhance the succinctness of the formalism as they provide useful shorthands for otherwise exponentially large formulas.

This exponential gain in succinctness makes it possible to extend static analysis to a larger set of XPath and XML schema features in a more efficient way. We believe the field of application of this logic may go beyond the XML setting. For example, in verification of linked data structures [20, 30, 12] reasoning on tree structures with in-depth cardinality constraints seems a major issue. Our result may help building solvers that are attractive alternatives to those based on non-elementary logics such as SkS [28], like Mona [17].

References

[1] Alessandro Bianco, Fabio Mogavero, and Aniello Murano. Graded computation tree logic. In *LICS*, 2009.

[2] Piero Bonatti, Carsten Lutz, Aniello Murano, and Moshe Vardi. The complexity of enriched μ-calculi. In *ICALP*, 2006.

[3] J. Clark and S. DeRose. XML path language (XPath) version 1.0, November 1999. W3C recommendation.

[4] Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Efficient asymmetric inclusion between regular expression types. In *ICDT*, 2009.

[5] Silvano Dal-Zilio, Denis Lugiez, and Charles Meyssonnier. A logic you can count on. In *POPL*, 2004.

[6] S. Demri and D. Lugiez. Presburger modal logic is PSPACE-Complete. In *IJCAR*, 2006.

[7] Wouter Gelade. Succinctness of regular expressions with interleaving, intersection and counting. In *MFCS*, 2008.

[8] Wouter Gelade, Marc Gyssens, and Wim Martens. Regular expressions with counting: Weak versus strong determinism. In *MFCS*, 2009.

[9] Wouter Gelade, Wim Martens, and Frank Neven. Optimizing schema languages for XML: Numerical constraints and interleaving. *SIAM J. Comput.*, 38(5), 2008.

[10] P. Genevès, N. Layaïda, and A. Schmitt. Efficient static analysis of XML paths and types. In *PLDI*, 2007.

[11] Pierre Genevès and Kristoffer Høgsbro Rose. Compiling XPath for streaming access policy. In *DocEng*, 2005.

[12] Peter Habermehl, Radu Iosif, and Tomás Vojnar. Automata-based verification of programs with tree updates. *Acta Inf.*, 47(1):1–31, 2010.

[13] J.G. Henriksen, J. Jensen, M. Jørgensen, N. Klarlund, B. Paige, T. Rauhe, and A. Sandholm. Mona: Monadic second-order logic in practice. In *TACAS*, 1995.

[14] H. Hosoya, J. Vouillon, and B. C. Pierce. Regular expression types for XML. *ACM Trans. Program. Lang. Syst.*, 27(1):46–90, 2005.

[15] Pekka Kilpelinen and Rauno Tuhkanen. One-unambiguity of regular expressions with numeric occurrence indicators. *Information and Computation*, 205(6):890 – 916, 2007.

[16] F. Klaedtke and H. Rueß. Monadic second-order logics with cardinalities. In *ICALP*, 2003.

[17] Nils Klarlund and Anders Møller. *MONA Version 1.4 User Manual*. BRICS Notes Series NS-01-1, Department of Computer Science, University of Aarhus, January 2001.

[18] D. Kozen. Results on the propositional μ-Calculus. In *ICALP*, 1982.

[19] O. Kupferman, U. Sattler, and M. Y. Vardi. The complexity of the graded μ-calculus. In *CADE*, 2002.

[20] Zohar Manna, Henny B. Sipma, and Ting Zhang. Verifying balanced trees. In *LFCS*, 2007.

[21] Maarten Marx. Conditional XPath. *ACM Trans. Database Syst.*, 30(4), 2005.

[22] Albert R. Meyer and Larry J. Stockmeyer. The equivalence problem for regular expressions with squaring requires exponential space. In *FOCS*, 1972.

[23] M. Murata, D. Lee, M. Mani, and K. Kawaguchi. Taxonomy of XML schema languages using formal language theory. *ACM Trans. Internet Techn.*, 5(4):660–704, 2005.

[24] Ulrike Sattler and Moshe Y. Vardi. The hybrid μ-calculus. In *IJCAR*, 2001.

[25] H. Seidl, T. Schwentick, A. Muscholl, and P. Habermehl. Counting in trees for free. In *ICALP*, 2004.

[26] Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific J. Math.*, 5(2):285–309, 1955.

[27] Balder ten Cate and Maarten Marx. Axiomatizing the logical core of XPath 2.0. *Theor. Comp. Sys.*, 44(4):561–589, 2009.

[28] James W. Thatcher and Jesse B. Wright. Generalized finite automata theory with an application to a decision problem of second-order logic. *Mathematical Systems Theory*, 2(1):57–81, 1968.

[29] M. Y. Vardi. Reasoning about the past with two-way automata. In *ICALP*, 1998.

[30] Karen Zee, Viktor Kuncak, and Martin Rinard. Full functional verification of linked data structures. In *PLDI*, 2008.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399