Deciding the First-Order Theory of an Algebra of Feature Trees with Updates (Extended Version)

Nicolas Jeannerod, Ralf Treinen

Deciding the First-Order Theory of an Algebra of Feature Trees with Updates (Extended Version)*

Nicolas Jeannerod and Ralf Treinen

Univ. Paris Diderot, Sorbonne Paris Cité, IRIF, UMR 8243, CNRS, Paris, France

Abstract. We investigate a logic of an algebra of trees including the update operation, which expresses that a tree is obtained from an input tree by replacing a particular direct subtree of the input tree, while leaving the rest unchanged. This operation improves on the expressivity of existing logics of tree algebras, in our case of feature trees. These allow for an unbounded number of children of a node in a tree.
We show that the first-order theory of this algebra is decidable *via* a weak quantifier elimination procedure which is allowed to swap existential quantifiers for universal quantifiers. This study is motivated by the logical modeling of transformations on UNIX file system trees expressed in a simple programming language.

1 Introduction

Feature trees are trees where nodes have an unbounded number of children, and where edges from nodes to their children carry names such that no node has two different outgoing edges with the same name. Hence, the names on the edges can be used to select the different children of a node. Feature trees have been used in constraint-based formalisms in the field of computational linguistics (e.g. [Smo92]) and constrained logic programming [AKPS94, ST94]. This work is motivated by a different application of feature trees: they are a quite accurate model of UNIX file system trees. The most important abstraction in viewing a file structure as a tree is that we ignore multiple hard links to files. Our mid-term goal is to derive, using symbolic execution techniques, from a shell script a logical formula that describes the semantics of this script as a relation between the initial file tree and the one that results from execution of the script.

Feature tree logics have at their core basic constraints like $x[f]y$, expressing that y is a subtree of x accessible from the root of x via feature f, and $x[f]\uparrow$, expressing that the tree x does not have a feature f at its root node. This is already sufficient to describe some tree languages that are useful in our context. For instance, the script consisting of the single command `mkdir /home/john`, which creates a directory `john` under the directory `home`, succeeds on a tree if the

* This work has been partially supported by the ANR project CoLiS, contract number ANR-15-CE25-0001.

tree satisfies the formula $\exists d.(r[\texttt{home}]d \wedge d[\texttt{john}] \uparrow)$, which expresses that `home` is a subdirectory of the root, which does itself *not* have a subdirectory `john`. (We ignore here the difference between directories and regular files, as well as file permissions.)

Update Constraints. In order to describe the effect of executing the above script we need more expressivity. A first idea is to introduce an update constraint $y \doteq x[f \mapsto z]$, expressing that the tree y is obtained from the tree x by setting its child f to z (creating the child when it does not exist). Using this, the semantics of our `mkdir` command could be described by

$$\exists d, d', e.\, (in[\texttt{home}]d \wedge d[\texttt{john}] \uparrow \wedge out \doteq in[\texttt{home} \mapsto d'] \wedge d' \doteq d[\texttt{john} \mapsto e] \wedge e[\emptyset])$$

Here, $e[\emptyset]$ expresses that e is an empty directory. Note that this formula, by virtue of the update constraint, expresses that any existing directories under `home` are not touched.

Programming constructs translate to combinations of logical formula. For instance, if $\phi_p(in, out)$, resp. $\phi_q(in, out)$ describe the semantics of script fragments p and q, then their composition is described by $\exists t.(\phi_p(in, t) \wedge \phi_q(t, out))$. The reality of our use case is more complex than that due to the hairy handling of error conditions in shell scripts [JMT17], and is up to future work.

Formulas with more complex quantification structure occur when we express interesting properties of scripts. For instance, p and q are equivalent if

$$\forall in, out.\, (\phi_p(in, out) \leftrightarrow \phi_q(in, out))$$

Debian requires in its policy [Deb17] so-called maintainer scripts to be idempotent, which can be expressed for a script p as

$$\forall in, out.\, (\phi_p(in, out) \leftrightarrow \exists t(\phi_p(in, t) \wedge \phi_p(t, out))$$

Since we are interested in verifying these kinds of properties on scripts we need a logic of feature trees including update constraints, and which enjoys a decidable first-order logic.

Related Work. The first decidability result of a full first-order theory of *Herbrand trees* (i.e, based on equations $x = f(x_1, \ldots, x_n)$) is due to Malc'ev [Mal71], this result has later been extended by [Mah88, CL89]. A first decidability result for the first-order theory of *feature trees* was given for the logic FT [AKPS94], which comprises the predicates $x[f]y$ and $x[f] \uparrow$, by [BS95]. This was later extended to the logic CFT [ST94], which in addition to FT has an *arity constraint* $x[F]$ for any finite set F of feature symbols, expressing that the root of x has precisely the features F, in [Bac95,BT98]. Note that in these logics one can only quantify over trees, not over feature symbols. The generalization to a two-sorted logic which allows for quantification over features is undecidable [Tre93], but decidability can be recovered if one restrains the use of feature variables to talk about existence of features only [Tre97]. All these decidable logics of trees have a non-elementary lower bound [Vor96]. The case of a feature logic with update constraints was open up to now.

Choosing the Right Predicates. The difficulty in solving update constraints stems from the fact that an update constraint involves three trees: the original tree, the final tree and the sub-tree that changes.

There are no symmetries between these three arguments, and a conjunction of several update constraints may become quite complex. The key in solving that problem is to work on a more elementary constraint system which is based on the classical $x[f]y$, and the new *similarity constraint* $x \mathrel{\dot{\sim}}_f y$. The latter constraint expresses that x and y have the same children with the same names, except for the name f where they may differ. This system has the same expressive power as update constraints since on the one hand $z \doteq x[f \mapsto y]$ is equivalent to $x \mathrel{\dot{\sim}}_f z \wedge z[f]y$, and on the other hand $x \mathrel{\dot{\sim}}_f y$ is equivalent to $\exists z, v.(z \doteq x[f \mapsto v] \wedge z \doteq y[f \mapsto v])$. In order to simplify these constraints one needs in fact the generalization $x \mathrel{\dot{\sim}}_F y$ where F is a finite set of features. For each set of features F, similarities $\sim_F$ are equivalence relations, which is very useful when designing simplification rules, and these relations have useful properties: $(x \mathrel{\dot{\sim}}_F y \wedge x \mathrel{\dot{\sim}}_G y) \leftrightarrow x \mathrel{\dot{\sim}}_{F \cap G} y$, and $(x \mathrel{\dot{\sim}}_F y \wedge y \mathrel{\dot{\sim}}_G z) \rightarrow x \mathrel{\dot{\sim}}_{F \cup G} z$.

Eliminating Quantifiers. Our theory of feature trees does not have the property of quantifier elimination in the strict sense [Hod93]. This is already the case without the update (or similarity) constraints, as we can see in the following example: $\exists x.(y[f]x \wedge x[g] \uparrow)$. This formula means that there is a local variable x such that y points to x through the feature f, and that x have no feature g. The problem here is that that formula contains an information about the global variable y. This situation is not unusual when designing decision procedures. There are basically two possible remedies: the first one is to extend the logical language by new predicates which express properties which otherwise would need existential quantifiers to express. This method is well-known from Presburger arithmetic, it was also used in [BS95, Bac95].

However, in the case of feature tree logics, the needed extension of the language is substantial and requires the introduction of *path constraints*. For instance, the above formula would be equivalent to the path constraint $y[f][g] \uparrow$ stating that the variable y has a feature f pointing towards a tree where there is no feature g. Unfortunately, this extension entails the need of quite complex simplification rules for these new predicates.

The alternative solution is to our knowledge due to [Mal71] and consists in exploiting the fact that certain predicates of the logic behave like functions. This solution was also used in [CL89] for Herbrand trees. When switching to feature trees this solution becomes quite elegant [Tre97], the above formula would be replaced by $\neg y[f] \uparrow \wedge \forall x.(y[f]x \rightarrow x[g] \uparrow)$ stating that y has a feature f ($\neg y[f] \uparrow$) and that for each variable x such that y points towards x via f (in fact, there is only one), x has no feature g. The price is that existential quantifiers are not completely eliminated but swapped for universal ones. This is, however, sufficient, since one can now apply this transformation to a formula in prenex normal form, and successively reduce the number of quantifier eliminations.

Structure of this paper. We summarize some notions from logic that will be used in the rest of the paper in Section 2. Our model of trees as well as the syntax and semantics of our logic are defined formally in Section 3. The quantifier elimination procedure in given in Section 4. We conclude in Section 5. Proofs are only sketched, full proofs are to be found in the companion technical report [JT18].

2 Preliminaries

We assume logical conjunction and disjunction to be associative and commutative, and equality to be symmetric. For instance, we identify the formula $x \doteq y \wedge (x[f]\uparrow \vee x[g]z)$ with $(x[g]z \vee x[f]\uparrow) \wedge y \doteq x$.

The set of free variables of a formula ϕ is written $\mathcal{V}(\phi)$. We write $\phi\{x \mapsto y\}$ for the formula obtained by replacing in ϕ all free occurrences of x by y. We write $\tilde{\exists}\phi$ for the existential closure $\exists\mathcal{V}(\phi).\phi$, and similarly $\tilde{\forall}\phi$ for $\forall\mathcal{V}(\phi).\phi$.

A *conjunctive clause with existential quantifiers*, or in short *clause*, is either $\bot$, or a finite set of literals prefixed by a string of existential quantifiers. Note that such a clause may still contain free variables, that is we do *not* require all its variables to be quantified. If $\exists\bar{X}.(l_1 \wedge \ldots \wedge l_n)$ is such a clause, then we can partition its set of literals $c = g_c \cup l_c$ such that g_c contains all the literals of c that contain no variable of X, and l_c the set of literals of c that contain at least one variable of X. We have the following logical equivalence:

$$\models (\exists X.c) \leftrightarrow (g_c \wedge \exists X.l_c)$$

We call (g_c, l_c) the *decomposition* of $\exists X.c$. g_c is the *global part* and l_c the *local part* of c, X is the set of *local variables* and $\mathcal{V}(\exists X.c) \supseteq \mathcal{V}(g_c)$ the set of *global variables*.

A *disjunctive normal form* (dnf) is a finite set of clauses, all of which are different from $\bot$.

A formula is in *prenex normal form* (pnf) if it is of the form $Q_1x_1 \ldots Q_nx_n.\phi$ where ϕ is quantifier-free, and where the Q_i are existential or universal quantifiers. If all Q_i are $\exists$ (resp. $\forall$) then the formula is called a Σ_1-formula (resp. Π_1-formula).

$A \rightsquigarrow B$ denotes the set of partial function from the set A to the set B with a finite domain. The domain of a partial function f is written $\mathtt{dom}(f)$. The complement of a set is written X^c. We write $X \setminus Y$ for $\{x \in X \mid x \notin Y\}$.

3 A logic for an algebra of trees with similarities

3.1 Feature Trees

In addition to what has been said in the introduction, our model of feature trees also has information attached to the *nodes* of the trees. In our application to UNIX filesystems, these could be records containing the usual file attributes like various timestamps and access permission bits, owner and group, and so on. This

work abstracts from the details of the information attached to tree nodes: we take the definition of node decorations, and the pertaining logic as a parameter. We only require that the logic of file node decorations has a quantifier elimination procedure, and that the logic contains a binary predicate $x \not\cong y$ expressing the *disequality* of two information items.

We assume given a set $\mathcal{D}$ of *decorations*, and an infinite set $\mathcal{F}$ of *features*. The letters f, g, h will always denote features.

The set $\mathcal{FT}$ of *feature trees* is inductively defined as

$$\mathcal{FT} = \mathcal{D} \times (\mathcal{F} \rightsquigarrow \mathcal{FT})$$

Here, the case of a partial function with empty domain serves as base case of the induction. Hence, this amounts to saying that a feature tree is a finite unordered tree where nodes are labeled by decorations, and edges are labeled by features. Each node in a feature tree has a finite number of outgoing edges, and all outgoing edges of a node carry different names. We write $\underline{t}$ for the decoration of the root node of t and we write $\overline{t}$ for its mapping at the root, i.e. $t = (\underline{t}, \overline{t})$. Our notion of equality on trees is *structural equality*, i.e. $t = s$ iff $\underline{t} = \underline{s}$ and $\overline{t} = \overline{s}$, that is $\mathtt{dom}(\overline{t}) = \mathtt{dom}(\overline{s})$ and $\overline{t}(f) = \overline{s}(f)$ for every $f \in \mathtt{dom}(\overline{t})$. Examples of feature trees are given in Figure 1.

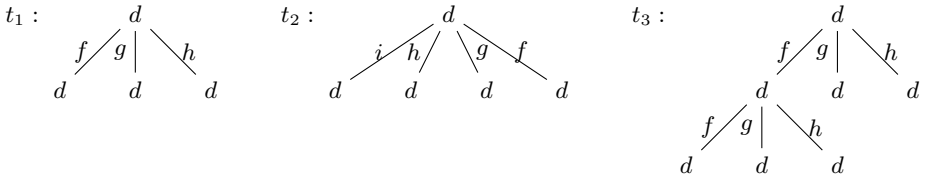

Fig. 1. Examples of Feature Trees. $d \in \mathcal{D}$ is some arbitrary node decoration.

For the reasons explained in the introduction, our logical language does not contain $y \doteq z[x \mapsto f]$ but the simpler $x \dot{\sim}_F y$ for any *finite* set $F \subseteq \mathcal{F}$. If $F \subseteq \mathcal{F}$ then we say that t *is similar to* s *outside* F, written $t \sim_F s$, if for all $f \in F^c = \mathcal{F} \setminus F$ we have that

- either $f \notin \mathtt{dom}(\overline{t}) \cup \mathtt{dom}(\overline{s})$
- or $f \in \mathtt{dom}(\overline{t}) \cap \mathtt{dom}(\overline{s})$, and $\overline{t}(f) = \overline{s}(f)$.

In other words, t and s are similar outside F if they have precisely the same children except maybe for the features in F.

3.2 Constraints and their interpretation

We assume given a set D of predicate symbols for decorations, and an interpretation $\mathcal{D}$ for D with universe $\mathcal{D}$. We also assume that we have a quantifier elimination procedure for D.

$x \doteq y$	Equality	$A(x_1 \dots x_n)$	Decoration predicate ($A \in D$)
$x[f]y$	Feature f from x to y	$x[f]\uparrow$	Absence of feature f from x
$x[F]$	Fence constraint (F finite)	$x \mathrel{\dot\sim_F} y$	Similarity outside F (F finite)

Fig. 2. Atomic constraints

The set of predicate symbols (or atomic constraints) of our logic is given in Figure 2. We will use following syntactic sugar: $x \neq y$ for $\neg(x \doteq y)$ and $x \not\sim_F y$ for $\neg(x \mathrel{\dot\sim_F} y)$. The universe of our model $\mathcal{FT}$ is the set $\mathcal{FT}$. The predicate symbols are interpreted as show in Figure **??** where ρ is a valuation from the free variables of the right-hand-side formula to the trees in $\mathcal{FT}$. Similarly to equality, we consider similarity predicates to be symmetric, that is we identify $x \mathrel{\dot\sim_F} y$ with $y \mathrel{\dot\sim_F} x$.

$$
\begin{array}{ll}
\mathcal{FT}, \rho \models x \doteq y & \text{iff } \rho(x) = \rho(y) \\
\mathcal{FT}, \rho \models x[f]y & \text{iff } f \in \mathtt{dom}(\overline{\rho(x)}) \text{ and } \overline{\rho(x)}(f) = \rho(y) \\
\mathcal{FT}, \rho \models x[f]\uparrow & \text{iff } f \notin \mathtt{dom}(\overline{\rho(x)}) \\
\mathcal{FT}, \rho \models x[F] & \text{iff } \mathtt{dom}(\overline{x}) \subseteq F \\
\mathcal{FT}, \rho \models x \mathrel{\dot\sim_F} y & \text{iff } \rho(x) \sim_F \rho(y) \\
\mathcal{FT}, \rho \models A(x_1, \dots, x_n) & \text{iff } \mathcal{D}, (\lambda x_i.\underline{\rho(x_i)}) \models A(x_1, \dots, x_n)
\end{array}
$$

Fig. 3. Interpretation of predicate symbols

Example 1. Let ρ be the valuation $[x \to t_1, y \to t_2, z \to t_3]$ for the trees defined in Figure 1. The following formulas are satisfied in $\mathcal{FT}, \rho$:

$$z[f]x, \quad x[i]\uparrow, \quad x[\{f, g, h, i\}], \quad x \mathrel{\dot\sim_{\{i\}}} y$$

Note the difference between our *fence* constraint, which states an upper bound on the root features of a tree, and the *arity* constraint of [ST94, Bac95] which states a precise set of root features of a tree. Both are equivalent, since one can express a fence F as a disjunction of all the arities that are subsets of F. Reciprocally, in our logic, we can express that x has arity F as $x[F] \wedge \bigwedge_{f \in F} \neg x[f]\uparrow$.

4 Quantifier Elimination

4.1 Clashing Clauses

We say that a clause c that is not $\bot$ *clashes* if one of the patterns of Figure 3 matches (modulo associativity and commutativity of $\wedge$) a sub-clause $c' \subseteq c$. C-Cycle is a clash since our model allows for finite feature trees only, the other clash cases should be obvious.

Lemma 1. *If a clause c clashes then $\mathcal{FT} \models (c \to \bot)$.*

C-Cycle	$x_1[f_1]x_2 \wedge \ldots \wedge x_n[f_n]x_1$	$(n \geq 1)$
C-Unsat-Decs	δ	(δ is a D-clause where $\mathcal{D} \models \neg\tilde{\exists}\delta$)
C-Feat-Abs	$x[f]y \wedge x[f]\uparrow$	
C-Feat-Fence	$x[f]y \wedge x[F]$	$(f \notin F)$
C-Neq-Refl	$x \neq x$	
C-NSim-Refl	$x \not\sim_F x$	

Fig. 4. Clash patterns

4.2 Positive Clauses with Local Variables

S-Eq	$\exists X, x.(x \doteq y \wedge c)$	$\Rightarrow$	$\exists X.c\{x \mapsto y\}$	$(x \neq y)$
S-Feats	$\exists X, z.(x[f]y \wedge x[f]z \wedge c)$	$\Rightarrow$	$\exists X.(x[f]y \wedge c\{z \mapsto y\})$	
			$(y \neq z$, and if $z \in \mathcal{V}_o$ then $y \in \mathcal{V}_o)$	
S-Feats-Global	$\exists X, x.(x[f]y \wedge x[f]z \wedge c)$	$\Rightarrow$	$\exists X, x.(x[f]y \wedge y \doteq z \wedge c)$	$(y, z \notin X)$
S-Sims	$x \dot{\sim}_F y \wedge x \dot{\sim}_G y \wedge c$	$\Rightarrow$	$x \dot{\sim}_{F \cap G} y \wedge c$	
P-Feat	$x \dot{\sim}_F y \wedge x[f]z \wedge c$	$\Rightarrow$	$x \dot{\sim}_F y \wedge x[f]z \wedge y[f]z \wedge c$	$(f \notin F)$
P-Abs	$x \dot{\sim}_F y \wedge x[f]\uparrow \wedge c$	$\Rightarrow$	$x \dot{\sim}_F y \wedge x[f]\uparrow \wedge y[f]\uparrow \wedge c$	$(f \notin F)$
P-Fence	$x \dot{\sim}_F y \wedge x[G] \wedge c$	$\Rightarrow$	$x \dot{\sim}_F y \wedge x[G] \wedge y[F \cup G] \wedge c$	
P-Sim	$x \dot{\sim}_F y \wedge x \dot{\sim}_G z \wedge c$	$\Rightarrow$	$x \dot{\sim}_F y \wedge x \dot{\sim}_G z \wedge y \dot{\sim}_{F \cup G} z \wedge c$	
			(if $\bigcap_{y \dot{\sim}_H z} H \not\subseteq F \cup G$)	

Fig. 5. Transformation rules for the positive case. Existential quantifiers are only written were relevant. Rule S-Feats is parameterized by a set $\mathcal{V}_o$ of variables.

As a preparation for the general case we first consider only one single clause $\exists X.(a_1 \wedge \ldots \wedge a_n)$ containing only positive atoms, prefixed by some existential quantifiers.

In this subsection and the following, we will use transformation rules as the ones in Figure 4. These rules describe transformations that map a clause to a formula (in this subsection the resulting formula is even a clause, but that will no longer be the case in the next subsection). We say that such a rule $left \Rightarrow right$ applies to a clause c if:

1. The pattern $left$ matches the complete clause c modulo associativity and commutativity of conjunction.
2. The side conditions of the rule, if any, are met.
3. The transformation yields a formula which is *different* from c.

If c is a clause and r a transformation rule then we write $r(c)$ for the formula obtained by applying r to c.

Each of the rules of Figure 4 describes an equivalence transformation in the model $\mathcal{FT}$. Equation elimination (S-Eq) is even a logical equivalence. S-Feats

implements the fact that features are functional. This rule is parameterized by a set $\mathcal{V}_o$ of variables which will in our procedure be the set of variables (local or global) of the input clause. The variable replacement is $\mathcal{V}_o$-oriented in the sense that we never replace a variable in $\mathcal{V}_o$ by a variable outside $\mathcal{V}_o$. S-Feat-Global is similar to S-Feat for the case that y and z are both global variables. S-Sims allows us to contract multiple similarities between the same pair of variables into one. P-Feats, P-Abs and P-Fence propagate constraints along a similarity, taking into account the index of the similarity. Finally, P-Sim is a kind of transitivity of similarity, where we take care not to add a similarity which is subsumed by already existing similarities.

The propagations play two important roles in that system. First, they move information, possibly leading to a clash. This is the case in the following example where a fence moves through similarities to clash with a feature constraint:

$$\begin{array}{ll} & x[f]v \wedge x \dot{\sim}_{\{g\}} y \wedge y \dot{\sim}_{\{h\}} z \wedge z[\varnothing] \\ \text{P-Fence} & x[f]v \wedge x \dot{\sim}_{\{g\}} y \wedge y[\{h\}] \wedge y \dot{\sim}_{\{h\}} z \wedge z[\varnothing] \\ \text{P-Fence} & x[f]v \wedge x[\{g,h\}]x \dot{\sim}_{\{g\}} y \wedge y[\{h\}] \wedge y \dot{\sim}_{\{h\}} z \wedge z[\varnothing] \end{array}$$

Second, they take information from local variables and move it to global variables. This mechanism is at the key of the elimination of existential quantifications, the idea being that once all the propagations took place, all interesting information is now explicit in the global part, and we can hence drop the local part.

$$\begin{array}{ll} & y[h]\uparrow \wedge \exists z.\, \big(x \dot{\sim}_{\{f\}} z \wedge z \dot{\sim}_{\{g\}} y\big) \\ \text{P-Abs} & y[h]\uparrow \wedge \exists z.\, \big(x \dot{\sim}_{\{f\}} z \wedge z[h]\uparrow \wedge z \dot{\sim}_{\{g\}} y\big) \\ \text{P-Sim} & y[h]\uparrow \wedge x \dot{\sim}_{\{f,g\}} y \wedge \exists z.\, \big(x \dot{\sim}_{\{f\}} z \wedge z[h]\uparrow \wedge z \dot{\sim}_{\{g\}} y\big) \\ \text{P-Abs} & x[h]\uparrow \wedge y[h]\uparrow \wedge x \dot{\sim}_{\{f,g\}} y \wedge \exists z.\, \big(x \dot{\sim}_{\{f\}} z \wedge z[h]\uparrow \wedge z \dot{\sim}_{\{g\}} y\big) \end{array}$$

The following function computes a normal form with respect to the rules of Figure 4:

```
function normalize-positive(c: positive clause)
  V_o := V(c_1) where c = ∃X.c_1
  while c does not clash and some rule r of Figure 4 applies to c
    c := r(c)
  return(c)
```

Lemma 2. *For any positive clause c, the function* `normalize-positive` *terminates and yields a positive clause that is equivalent in $\mathcal{FT}$ to c.*

Lemma 3. *Let* `normalize-positive` *return a clause $\exists X.c$ that does not clash and (g_c, l_c) be its decomposition. If c contains no atom $x[f]y$ with $x \notin X$ and $y \in X$ then*

$$\mathcal{FT} \models \tilde{\forall}(\exists X.c \leftrightarrow g_c)$$

Both lemmas are in fact special cases of the forthcoming Lemmas 5 and 6 of Section 4.3.

Lemma 3 can serve for quantifier elimination in the positive case, at least when there is no feature constraint from a global variable to a local one. We will see in Section 4.4 what can be done if this is not the case.

4.3 General Clauses with Local Variables

In the case of clauses containing both positive and negative literals we have to consider transformation rules that introduce negations or disjunctions. However, our rules will continue to always take a single clause as input. As a consequence, we have to transform the result obtained by a transformation into disjunctive normal form. We assume given a function `dnf` that takes a formula without universal quantifiers, and which contains only positive occurrences of existential quantifiers and returns an equivalent dnf that does not contain any clashing clauses. This can be achieved by using a standard dnf transformation and then purging all clashing clauses, or alternatively by applying the clash rules on the fly.

Syntactic sugar. In the transformation rules to be presented below we will use several abbreviations that allow us to write the rules more concisely. First we have

$$x\langle F\rangle \;:=\; \bigvee_{f\in F} \exists z.x[f]z$$

where $F \subset \mathcal{F}$ is a finite set. This formula states that x has *at least one* feature in the set F, it can be seen as a dual to the fence constraint $x[F]$ which states that x has *at most* the features in the set F. Note that $x\langle F\rangle$ introduces a disjunction, so introducing such a formula requires the result to be put into dnf of the result.

The formula $x \not\doteq_f y$ states that x and y differ at feature f, that is either one of them has f and the other one has not, or both have children at f which are different. The formula $x \not\doteq_F y$ generalizes this to a finite set $F \subset \mathcal{F}$, stating that x and y differ at at least one of the features in F.

$$\begin{aligned} x \not\doteq_f y := {} & \exists z'.(x[f]\uparrow \wedge y[f]z') \vee \exists z.(x[f]z \wedge y[f]\uparrow) \\ & \vee \exists z, z'.(x[f]z \wedge y[f]z' \wedge (z \not\cong z' \vee z \not\sim_\varnothing z')) \\ x \not\doteq_F y := {} & \bigvee_{f\in F} x \not\doteq_f y \end{aligned}$$

These formulas introduce disjunctions. They also introduce negated similarities at some newly created children of x and y, so we have to take care in the termination proof when these formulas are introduced by a transformation.

New rules. Figure 5 extends the previously defined set of rules by adding several replacement rules and two enlargement rules. First, we have R-NEq, R-NFeat and R-NAbs which respectively eliminate occurrences of the negated constraints $x \not\doteq y$, $\neg x[f]y$ and $\neg x[f]\uparrow$. Since no other rule introduces any of these negated constraints we now have these out of the way.

$$
\begin{array}{lrcll}
\textsc{r-neq} & x \not\approx y \wedge c & \Rightarrow & (x \not\cong y \vee x \not\sim_{\varnothing} y) \wedge c & \\
\textsc{r-nfeat} & \neg x[f]y \wedge c & \Rightarrow & (x[f]\uparrow \vee \exists z.(x[f]z \wedge (y \not\cong z \vee y \not\sim_{\varnothing} z))) \wedge c & \\
\textsc{r-nabs} & \neg x[f]\uparrow \wedge c & \Rightarrow & \exists z.x[f]z \wedge c & \\
\textsc{r-nfence-fence} & x[F] \wedge \neg x[G] \wedge c & \Rightarrow & x[F] \wedge x\langle F \setminus G\rangle \wedge c & \\
\textsc{r-nsim-sim} & x \dot{\sim}_F y \wedge x \not\sim_G y \wedge c & \Rightarrow & x \dot{\sim}_F y \wedge x \not\approx_{F\setminus G} y \wedge c & \\
\textsc{r-nsim-fence} & x[F] \wedge x \not\sim_G y \wedge c & \Rightarrow & x[F] \wedge \big(\neg y[F \cup G] \vee x \not\approx_{F\setminus G} y\big) \wedge c & \\
\textsc{e-nfence} & x \dot{\sim}_F y \wedge \neg x[G] \wedge c & \Rightarrow & x \dot{\sim}_F y \wedge (\neg x[F \cup G] \vee x\langle F \setminus G\rangle) \wedge c & (F \not\subseteq G) \\
\textsc{e-nsim} & x \dot{\sim}_F y \wedge x \not\sim_G z \wedge c & \Rightarrow & x \dot{\sim}_F y \wedge \big(x \not\sim_{F\cup G} z \vee x \not\approx_{F\setminus G} z\big) \wedge c & (F \not\subseteq G)
\end{array}
$$

Fig. 6. Replacement and Enlargement rules for the general case.

Then we have three rules that combine a positive with a negative constraint. R-NFence-Fence applies to the case where we have both a positive fence F and a negated fence G for x. We simplify this by keeping the positive fence F, and replacing the negative fence by saying that x must have a feature that is in F (since that is all it can have), but not in G. Similarly, R-NSim-Sim applies when we have between x and y both a positive similarity except in F, and a negated similarity except in G. We simplify this by keeping the positive similarity, and replacing the negated similarity by stating that x and y differ at a feature that is in F (since these are the only features where they may differ) but not in G. Finally, R-NSim-fence applies when we have a fence F for x, and a negated similarity with y except in G. Note that $G^c = ((F \cup G)^c \cup (F \setminus G)$. Hence, the negated similarity is equivalent to saying that either y has a feature outside $F \cup G$, which is the only possibility to have a difference with x outside $F \cup G$ since x has already fence F, or the difference is in the finite set $F \setminus G$.

Finally, we have the two enlargement rules E-NFence and E-NSim. Their sole purpose is to ensure (by enlarging the negated fence or the index of a negated similarity) that the rules in Figure 6 can be applied when we have a similarity in conjunction with a negated fence or a negated similarity. The correctness proof of these rules is in fact similar to the three previous rules. In fact, the similarity between x and y is not needed for the correctness of these two rules and serves only for the termination proof since the requirement of a context $x \dot{\sim}_F y$ excludes arbitrary enlargments.

$$
\begin{array}{lrcll}
\textsc{p-nfence} & x \dot{\sim}_F y \wedge \neg x[G] \wedge c & \Rightarrow & x \dot{\sim}_F y \wedge \neg x[G] \wedge \neg y[G] \wedge c & (F \subseteq G) \\
\textsc{p-nsim} & x \dot{\sim}_F y \wedge x \not\sim_G z \wedge c & \Rightarrow & x \dot{\sim}_F y \wedge x \not\sim_G z \wedge y \not\sim_G z \wedge c & (F \subseteq G)
\end{array}
$$

Fig. 7. Propagation rules for the general case.

The two rules in Figure 6 may propagate a negated fence or a negated similarity through a similarity. In fact, if x and y coincide outside F and $F \subseteq G$, then x and y also coincide outside G. Hence, if x has a feature outside G then

so does y (P-NFence), and if x differs from z at some feature outside G then so does y (P-NSim).

We define the set of rules R_1 as the union of all the transformation rules of Figures 4 and 5, and R_2 as the union of all transformation rules of Figure 6.

```
function normalize(c: clause)
  d := {c}
  V_o := V(c_1) where c = ∃X.c_1
  while exists c ∈ d to which some rule r ∈ R_1 applies
    d := (d \ {c}) ∪ dnf(r(c))
  while exists c ∈ d to which r ∈ R_2 applies
    d := (d \ {c}) ∪ {r(c)}
  return(d)
```

The function `normalize` normalizes first by rule set R_1, and then by rule set R_2. This decomposition is necessary to ensure termination. It also makes sense since application of rules R_2 conserves normal forms with respect to R_1.

Lemma 4. *The output of* ***normalize*** *is a dnf where each conjunction is in normal form for* $R_1 \cup R_2$.

Proof (sketch). We have to prove that the application of one of the rules in R_2 to a normal form with respect to R_1 does not produce a redex for any of the rules in R_1. Assume, for instance that the application of P-NFence to c introduces a redex of R-NFence-Fence. This means that c must contain a positive fence for y. c must be in normal form in particular with respect to P-Fence, which means that x must have a fence constraint in c, which yields a contradiction since then c is not in normal form with respect to R-NFence-Fence. The other cases are similar (see [JT18]).

Lemma 5. *The function* ***normalize****, when applied to a clause* c*, terminates and yields a dnf* d *such that* $\mathcal{FT} \models \tilde{\forall}(c \leftrightarrow d)$.

Proof (sketch). Equivalence of c and d follows from the fact that each transformation rule is an equivalence in $\mathcal{FT}$. Termination is shown by defining a well-founded order on clauses such that each rules transforms a clause into a set of stricter smaller clauses. The termination order on dnf formulas is the multiset extension [DM79] of this order.

This order is a lexicographic order over twelve different measures that decrease with the applications of the rules. We can for instance handle the rules R-NEq, R-NFeat and R-NAbs first by saying that they decrease the number of negated equalities, feature constraints or absences. Since nothing introduces those literals, this is already a good start.

The first main difficulty in finding that order comes from the fact that all the propagation rules are trying to saturate the clause. A good measure that decreases with them is then the set of all possible atoms that are not in the formula. For P-Feat, for instance: $\{(x[f]y) \mid x, y \in \mathcal{V}(c); f \in \mathcal{F}(c); (x[f]y) \notin c\}$. That would make a good measure if $\mathcal{V}(c)$ could not increase with the application

of other rules such as R-NSim-Fence. We have thus to handle these other rules first, which leads us to another main difficulty.

The second main difficulty comes from the negated similarities. Indeed, while all other literals may only move "horizontally" following the similarities, negated similarities may "descend" in the constraint, creating variables and feature constraints if needed. It is really non trivial to see when it will stop, and in particular find a bound on the number of variables introduced.

Let us consider the following example constraint and one of its reduction paths (that is, the reduction may create several branches in the dnf, and we take only the one we are interested in):

$$x_0[f]x_1 \wedge x_1[f]y_0 \wedge \wedge x_0[\{f\}] \wedge x_1[\{f\}] \wedge x_0 \not\sim_\varnothing y_0$$
By R-NSim-Fence:
$$\exists y_1, z_1 . x_0[f]x_1 \wedge x_1[f]y_0 \wedge x_0[\{f\}] \wedge x_1[\{f\}] \wedge x_0[f]z_1 \wedge y_0[f]y_1 \wedge z_1 \not\sim_\varnothing y_1$$
By S-Feats:
$$\exists y_1 . x_0[f]x_1 \wedge x_1[f]y_0 \wedge y_0[f]y_1 \wedge x_0[\{f\}] \wedge x_1[\{f\}] \wedge x_1 \not\sim_\varnothing y_1$$

In two rules, we created a new variable y_1, and removed a negated similarity just to put it again somewhere else. Note in particular that R-NSim-Fence can still apply, because x_1 has now a fence and a negative similarity. In fact, if, instead of two, we take a number n of variables x_i, we can extend that example into one that always doubles the number of variables.

The key to our solution to this problem is that rules that make negative similarities descend, thus introducing feature constraints and new variables, need some "fuel", which is the presence of positive fences or similarities. We define the *original variables* as the variables that were in the clause at the begining of `normalize`. Then, we show that

1. the number of original variables cannot grow;
2. there are never feature constraints from non-original variables towards original ones;
3. the positive fences and similarities can only be present on original variables.

It remains the problem that negative similarities can descend. At some point, they will necessarily go too deep and leave the area where the original variables may live. By doing so, they loose the positive fences and similarities that they need to keep descending, and the process stops.

The full proof, including the lemmas corresponding to the points (1), (2) and (3), the definition of the measures and the technical details can be found in [JT18].

Lemma 6. *Let the function* `transform` *return a dnf which contains a clause* $\exists X.c$*. Let* (g_c, l_c) *be the decomposition of* c*. If* c *contains no atom* $x[f]y$ *with* $x \notin X$ *and* $y \in X$ *then*

$$\mathcal{FT} \models \tilde{\forall}(\exists X.c \leftrightarrow g_c)$$

Proof (sketch). Recall that any clause in a dnf is clash-free. We only have to show that $\mathcal{FT} \models \tilde{\forall} g_c \rightarrow \exists X.l_c$. Let $\mathcal{FT}, \alpha_0 \models g_c$. Let $d \subseteq c$ contain all the D literals of c. Due to clash pattern Decs there exists a valuation β such that $\mathcal{D}, \beta \models d$.

We define a relation on X as follows: $x \sqsubset_c y$ iff $y[f]x \in c$ for some $f \in \mathcal{F}$. Due to clash pattern C-Cycle, its transitive closure is a strict partial order which is embedded in some strict total order $\sqsubset$. Hence, we have that if $y[f]x \in c$ then we either have that $x \sqsubset y$, or that $x \notin X$. We now define $\rho_{Y \cup X}$ by induction on X following $\sqsubset$ ensuring that we keep satisfied all the literals containing variables where our valuation is defined. The base case of the induction is trivial (ρ_Y). In the induction case, we take $x \in X$, we define $Z = \{z \in Y \cup X \mid z \sqsubset x\}$ and we assume that we have ρ_Z already defined such that it satisfies all the literals about variables in Z.

We are going to extend it by defining $\rho_{Z \cup \{x\}}$. We define the following partial map m_x for x:

$$m_x = \{(f, \rho_Z(y)) \mid (x[f]y) \in c\}$$

Note that this defines a partial function, because there cannot be f, y and y' such that $(x[f]y) \in c$, $(x[f]y') \in c$ and $y \neq y'$ due to clash pattern S-Feats. Consider now the set of all the variables that are smaller than x and that are in a similarity relation with x:

$$down(x) = \{y \mid y \sqsubset x, (x \dot{\sim}_H y) \in c \text{ for some } H\}$$

We will now define m'_x using m_x and $\rho_Z(y)$ for all the $y \in down(x)$. There are three cases depending on whether $down(x)$ is empty, and depending on whether there is a fence constraint for x.

1. If $down(x) = \emptyset$ and there is some fence constraint $(x[F]) \in c$, then we define $m'_x = m_x$.
2. If $down(x) = \emptyset$ and there is no fence constraint $(x[F]) \in c$, then we choose a fresh feature h_x which does not occur in c (not even in a fence or similarity), does not occur in $\mathtt{dom}(\overline{\rho_Y(y)})$, for any $y \in Y$, is different from h_z, for any $z \sqsubset x$. (Recall that the mapping of a feature tree is required to have a finite domain.) Let $d \in \mathcal{D}$ be some arbitrary node decoration. We define $m'_x = m_x \cup \{(h_x, (d, \emptyset))\}$, that is m' is obtained from m by adding an edge labeled h_x going to the empty tree.
3. If $down(x) \neq \emptyset$ then we define

$$m'_x = m_x \cup \bigcup_{\substack{z \sqsubset x \\ (x \dot{\sim}_H z) \in c}} \overline{\rho_Z(z)} \restriction_{H^c}$$

where, when $\overline{\rho_Z(z)} \restriction_{H^c}$ is the restriction of $\overline{\rho_Z(z)}$ to the complement of H in $\mathcal{F}$. Of course one has to show in this last case that m'_x is indeed well-defined as a function.

Finally, one can show by induction on the order $\sqsubset$ that $\mathcal{FT}, \rho_{Z \cup \{x\}} \models l$ for every $l \in c$ such that $z \sqsubset x$ for every $z \in \mathcal{V}(c)$. The details can be found in [JT18].

We call a clause *normalized* when it is an element of a dnf returned as result of function `normalize`.

4.4 Quantifier Elimination

In order to eliminate a block of existential quantifiers from a clause we apply iteratively the following rule:

$$\textsc{Feat-Fun} \quad \exists X, x.(y[f]x \wedge c) \quad \Rightarrow \quad \neg y[f]\uparrow \wedge \forall x.\,(y[f]x \rightarrow \exists X.c) \qquad (y \notin X, y \neq x)$$

This rule follows the idea of [Mal71], and was already applied to feature constraints in [Tre97]. The rule obviously describes an equivalence in $\mathcal{FT}$ due to the fact that features are functional, observing that $\exists X, x.(y[f]x \wedge c)$ is logically equivalent to $\exists x.(y[f]x \wedge \exists X.c)$ when $y \notin X$.

```
recursive function switch(c: normalized clause)
  if ∃X, x.(y[f]x ∧ c') matches c and y ∉ X:
    (¬y[f]↑ ∧∀x.(y[f]x → switch(∃X.c')))
  else:
    let (g_c, l_c) = decomposition(c) in g_c
```

Example 2. When given the following formula

$$\exists v, w.(y[f]v \wedge v[f]w \wedge w[f]z \wedge w[\{f, g\}] \wedge y \dot{\sim}_{\varnothing} z$$

the function `switch` returns

$$\neg y[f]\uparrow \wedge \forall v.(y[f]v \rightarrow (\neg v[f]\uparrow \wedge \forall w.(v[f]w \rightarrow (w[f]z \wedge y \dot{\sim}_{\varnothing} z))))$$

Lemma 7. *Given a normalized clause c, `switch(c)` terminates and yields a formula ψ such that*

1. $\mathcal{FT} \models \tilde{\forall}(c \leftrightarrow \psi)$;
2. $\mathcal{V}(\psi) \subseteq \mathcal{V}(c)$;
3. *ψ contains no existential quantifiers and only positive occurrences of universal quantifiers;*
4. *If $\mathcal{V}(c) = \emptyset$ then ψ is quantifier-free.*

We can now write a function that transforms a Σ_1 formula into an equivalent Π_1 formula. For this we assume given a function `pnf` that transforms any formula into its prenex normal form.

```
function solve(p: Σ_1 formula)
  let ∃X.q = p where p is quantifier-free
  d := dnf(q)
  dt := ⋁_{c∈d} transform(∃X.c)
  u := ⋁_{c∈dt} switch(c)
  pnf(u)
```

Finally, the function `decide` takes a formula in prenex normal form and returns an equivalent (in $\mathcal{FT}$) formula without any quantifiers. If Q is a string of quantifiers, then $\overline{Q}$ is the string of quantifiers obtained from Q by changing $\exists$ into $\forall$ and vice-versa. For instance, $\overline{\exists x \forall y \exists z} = \forall x \exists y \forall z$.

```
recursive function decide(p: pnf)
  if p is quantifier-free then p
  else if p is Q.∃X.q
       where q quantifier-free, Q does not end on ∃
       then decide(Q.solve(∃X.q))
  else if p is Q.∀X.q
       where q quantifier-free, Q does not end on ∀
       then ¬(decide(Q̄.solve(∃X.¬q)))
```

Theorem 1. *Given a formula p in prenex normal form, decide(p) terminates and yields a formula q such that*

- $\mathcal{FT} \models \tilde{\forall}(p \leftrightarrow q)$
- $\mathcal{V}(q) \subseteq \mathcal{V}(p)$
- *q is a Π_1 formula, and quantifier-free in case $\mathcal{V}(p) = \emptyset$*

Proof (sketch). Termination follows from the fact that at each call to `decide`, the number of quantifier *alternations* in the pnf decreases.

If we apply `decide` to a closed formula, we hence obtain an equivalent (in $\mathcal{FT}$) formula that contains no free variables and no quantifiers, that is a boolean combinaton of True and False.

Corollary 1. *The first order theory of $\mathcal{FT}$ is decidable.*

5 Conclusion

We have presented a quantifier elimination procedure for a first-order theory of feature trees with similarity constraints. Since update constraints can be expressed by similarity and feature constraints, this implies in particular that the first-order theory of feature trees with update constraints is decidable.

Our model of feature trees is in several respects an abstraction of UNIX file systems [Bac86]. First, real file systems make a distinction between different kinds of files (directories, regular files, various kinds of device files). This distinction is omitted here just for the sake of presentation. More importantly, real file systems are not really trees as they allow for multiple paths from the root to regular files (which must be sinks), and they provide for symbolic links. Extending the model by any of these seems to lead to undecidability of the full first-order theory, so if we want to include them in the model we will have to look for smaller fragments which are still decidable in such an extension, and sufficient for our application to the symbolic execution of scripts.

Acknowledgments. The idea of investigating update constraints on feature trees originates from discussions with Gert Smolka a long time ago. We would like to thank the members of the CoLiS project for numerous discussions on tree constraints and their use in modeling tree operations, in particular Yann Régis-Gianas, Claude Marché, Kim Nguyen, Joachim Niehren, Sylvain Salvati, and Mihaela Sighireanu.

References

[AKPS94] Hassan Aït-Kaci, Andreas Podelski, and Gert Smolka. A feature-based constraint system for logic programming with entailment. *Theoretical Computer Science*, 122(1–2):263–283, January 1994.

[Bac86] Maurice Bach. *The Design of the UNIX Operating System*. Prentice-Hall, 1986.

[Bac95] Rolf Backofen. A complete axiomatization of a theory with feature and arity constraints. *Journal of Logic Programming*, 24(1&2):37–71, July/August 1995.

[BS95] Rolf Backofen and Gert Smolka. A complete and recursive feature theory. *Theoretical Computer Science*, 146(1–2):243–268, July 1995.

[BT98] Rolf Backofen and Ralf Treinen. How to win a game with features. *Information and Computation*, 142(1):76–101, April 1998.

[CL89] Hubert Comon and Pierre Lescanne. Equational problems and disunification. *Journal of Symbolic Computation*, 7:371–425, 1989.

[Deb17] Debian Policy Mailing List. *Debian Policy Manual, version 4.1.3*. Debian, December 2017. `https://www.debian.org/doc/debian-policy/`.

[DM79] Nachum Dershowitz and Zohar Manna. Proving termination with multiset orderings. *Communication of the ACM*, 22(8):465–476, 1979.

[Hod93] Wilfried Hodges. *Model Theory*, volume 42 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1993.

[JMT17] Nicolas Jeannerod, Claude Marché, and Ralf Treinen. A formally verified interpreter for a shell-like programming language. In Andrei Paskevich and Thomas Wies, editors, *Verified Software. Theories, Tools, and Experiments*, volume 10712 of *LNCS*, pages 1–18, Heidelberg, Germany, July 2017. Springer.

[JT18] Nicolas Jeannerod and Ralf Treinen. Deciding the first-order theory of an algebra of feature trees with updates (extended version), January 2018. `https://www.irif.fr/~treinen/publi/update-constraints.pdf`.

[Mah88] Michael J. Maher. Complete axiomatizations of the algebras of finite, rational and infinite trees. In *LICS*, pages 348–357, Edinburgh, Scotland, UK, July 1988. IEEE.

[Mal71] Anatoliĭ Ivanovič Malc'ev. Axiomatizable classes of locally free algebras of various type. In III Benjamin Franklin Wells, editor, *The Metamathematics of Algebraic Systems: Collected Papers 1936–1967*, chapter 23, pages 262–281. North Holland, 1971.

[Smo92] Gert Smolka. Feature constraint logics for unification grammars. *Journal of Logic Programming*, 12:51–87, 1992.

[ST94] Gert Smolka and Ralf Treinen. Records for logic programming. *Journal of Logic Programming*, 18(3):229–258, April 1994.

[Tre93] Ralf Treinen. Feature constraints with first-class features. In Andrzej M. Borzyszkowski and Stefan Sokołowski, editors, *Mathematical Foundations of Computer Science*, volume 711 of *LNCS*, pages 734–743. Springer, August/September 1993.

[Tre97] Ralf Treinen. Feature trees over arbitrary structures. In Patrick Blackburn and Maarten de Rijke, editors, *Specifying Syntactic Structures*, chapter 7, pages 185–211. CSLI Publications and FoLLI, 1997.

[Vor96] Sergei Vorobyov. An improved lower bound for the elementary theories of trees. In M. A. McRobbie and J. K. Slaney, editors, *CADE'96*, volume 1104 of *LNCS*, pages 275–287, New Brunswick, NJ, July/August 1996. Springer.

A Proof of Lemma 5 (termination)

A.1 Introduction and example

To show the termination, we have to find a measure that decreases strictly at each step of the `while` loop of the function `normalize` of Section 4.3. When trying to find such an order one encounters quickly two problems:

- Some of the rules create new variables.
- Negated similarities cannot only be propagated horizontally (i.e., along similarities $x \stackrel{\cdot}{\sim}_F y$) by rule P-NSim, but also descend vertically (i.e., along feature constraints $x[f]y$), creating new feature constraints by R-NSim-Fence, R-NSim-Sim and E-NSim.

Taken together, these two properties of the rule system pose a serious problem for termination, since we have to assure that pushing downwards of negated similarities terminates even when new variables and feature constraints may be created.

The idea of the proof is to show that the descent of negated similarities is in fact bounded. Let us first define a notion of depth of the variables in a clash-free constraint:

$$d_c(y) = \mathtt{max}\{1 + d_c(x) \mid (x[f]y) \in c\}$$

This definition of depth is valid because there are no cycles in the feature constraints in c. We want to show that, although the negated similarities are descending, they cannot go too far, and that the process stops.

One might hope to find a bound on the depths of the variables in a clause. However, it is not clear how this can be achieved. Here is an example where we create variables and almost double the depth of the clause:

$$\left(\bigwedge_{0 \leq i < n} x_i[f]x_{i+1} \wedge x_i[\{f\}]\right) \wedge x_0 \not\sim_\varnothing x_n$$

Initially, all variables are at depth $\leq n$. There is an occurrence of pattern $x_0[\{f\}] \wedge x_0 \not\sim_\varnothing x_n$. We can thus apply R-NSim-Fence, and we get a set of clauses that contains the following:

$$\exists y_1, z_1. \left(\bigwedge_{0 \leq i < n} x_i[f]x_{i+1} \wedge x_i[\{f\}]\right) \wedge x_0[f]z_1 \wedge x_n[f]y_1 \wedge z_1 \not\sim_\varnothing y_1$$

We have the pattern $\exists z_1.x_0[f]x_1 \wedge x_0[f]z_1$ and we can thus apply S-Eq which gives us:

$$\exists y_1. \left(\bigwedge_{0 \leq i < n} x_i[f]x_{i+1} \wedge x_i[\{f\}]\right) \wedge x_n[f]y_1 \wedge x_1 \not\sim_\varnothing y_1$$

In two rule applications, we created a variable y_1 of depth $n+1$, and we reproduced the pattern consisting of a fence and a negated similarity that was on x_0 and x_n to x_1 and y_1. We can keep doing this, and we obtain in the end:

$$\exists (y_i)_{1\leq i\leq n}. \left(\bigwedge_{0\leq i<n} x_i[f]x_{i+1} \wedge x_i[\{f\}]\right) \wedge x_n[f]y_1 \wedge \left(\bigwedge_{1\leq i<n} y_i[f]y_{i+1}\right) \wedge x_n \not\sim_{\varnothing} y_n$$

where the newly introduced variable y_n is at depth $2n$.

This shows that it is non trivial to have a bound on the depth of a clause, and on the number of variables in it. Luckily, there is a variant of this argument which we can prove. In fact, to descend and possibly create features, the negated similarities need "fuel", that fuel being the fences and positive similarities that are necessary to trigger R-NSim-Fence, R-NSim-Sim or E-NSim. We can show that this fuel can only be present on variables that were originaly present in the clause. And since their number cannot grow and their depth is bounded, we get what we want.

A.2 Technical definitions

Let $\exists X_o.c_o$ be a clash-free clause. This will be the input clause to which we apply the function `normalize`. Define the original variables as:

$$\mathcal{V}_o = \mathcal{V}(c_o)$$

Note that $\mathcal{V}_o$ contains not only the free variables of the clause but also the variables of X_o that are present in c_o. In the following, we take S-Feats to be $\mathcal{V}_o$-oriented. Also, we will only consider clauses that are descendants of this clause c_o, that is, they are inhabitants at some point of the set d in the function `normalize`(c_o). In particular, they are all clash-free.

Let us define the depth of a variable in a clause c by:

Definition 1.

$$d_c(y) = \mathit{max}\{1 + d_c(x) \mid (x[f]y) \in c\}$$

This definition is valid because c is clash-free, and in particular without cycles. We have two important properties about d_c:

Proposition 1. *For any clause $\exists X.c$, and any variable in $\mathcal{V}(c)$, we have that $d_c(x) < \mathit{card}(\mathcal{V}(c))$.*

Proof. This is an immediate consequence of the fact that c is clash-free: no variable may appear twice on a path leading to x.

Proposition 2. *The depth of a variable cannot decrease. That is, for any clause $\exists X_1.c_1$ that transforms into $\exists X_2.c_2$ and any variable x in these two clauses, $d_{c_1}(x) \leq d_{c_2}(x)$.*

Proof. Consider clauses c_1 and c_2 and a variable x in both of them. There is a path in c_1 of length $d_{c_1}(x)$ leading to x by definition of the depth. We will show that this path is also present in c_2, although it may be slightly changed.

Our system contains several rules that add feature constraint leading to *new* variables, these rule do not modify any existing paths. There is only one rule that may change the existing feature constraints: S-Feats. However, it only renames one variable into an other, which means that all the features of our path are still present in the new clause. Also note that since $x \in \mathcal{V}(x_2)$, x cannot be the variable that is renamed. As a consequence, the path of length $d_{c_1}(x)$ in c_1 to x still leads to x in c_2, that is $d_{c_2}(x) \geq d_{c_1}(c)$.

We will now show several properties on the variables of the clauses $\exists X.c$ that descend from c_o.

Proposition 3. *For all $\exists X.c$ and $x \in \mathcal{V}(c) \setminus \mathcal{V}_o$, x is local in c, that is $x \in X$.*

Proof. The proof is induction on the rule sequence that was applied to obtain $\exists X.c$ from c_0. By inspecting the rules one verifies easily that any variable that is introduced by any rule also is existentially quantified.

We now have three important properties that state that positive equalities, similarities and fences cannot escape the set $\mathcal{V}_o$ of original variables. This means in particular that, although the negated similarities may descend, only the original variables have what is needed to trigger the rules R-NSim-Fence, R-NSim-Sim and E-NSim.

Proposition 4. *If $(x \doteq y) \in c$, then $x, y \in \mathcal{V}_o$.*

Proof. By induction on the rule sequence that led us there from c_o to c. This is trivially verified in c_o by definition of $\mathcal{V}_o$. The only rule that may introduce equalities is S-Feats-Global, that only adds equalities between global hence original (Prop. 3) variables.

Proposition 5. *If $(x \dot{\sim}_F y) \in c$, then $x, y \in \mathcal{V}_o$.*

Proof. By induction on the rule sequence that led us there from c_o to c. The only rules that may modify similarities are

1. S-Sims which only changes the index of an already existing similarity between the same variables,
2. P-Sim which creates a new similarity between two variables that each are already in a similarity relation,
3. S-Eq which may rename the variables in an existing similarity. However, it is only renaming original variables into original ones (Prop. 4).
4. S-Feats which may rename the variables in an existing similarity. However, if a renaming of say z into y introduced a similarity for variable $y \notin \mathcal{V}_o$ then we would have already a similarity for z, and $z \notin \mathcal{V}_o$ by the side condition of (S-Feats), which a is contradiction to the induction hypothesis.

Proposition 6. *If* $(x[F]) \in c$*, then* $x \in \mathcal{V}_o$*.*

Proof. By induction on the rule sequence that led us there from c_o to c. The only rules that may modify fences are

1. P-Fence which creates a fence constraint for a variable that appears in a similarity constraint and hence is, by Proposition 5, a variable of $\mathcal{V}_o$,
2. S-Eq which may rename the variable in an existing fence constraint. However, it only renames an original variable into an other original one.
3. S-Feats which may rename the variable in an existing fence constraint. However, if a renaming of say z into y introduced a fence constraint for variable $y \notin \mathcal{V}_o$ then we would have already a fence constraint for z, and $z \notin \mathcal{V}_o$ by the side condition of (S-Feats), which a is contradiction to the induction hypothesis.

We now want to prove that the depth of original variables is bounded by $\texttt{card}(\mathcal{V}_o)$. However, new variables may have been introduced, and in order to prove our bound we have to assure that the introduction of new variables cannot increase the depth of original variables. The two following propositions will allow us to conclude that only original variables may occur on paths leading to original variables, which is sufficient to show that the depth of original variables cannot increase.

Let us first define the notion of fathers of a variable in a clause:

Definition 2.

$$\mathit{fathers}_c(y) = \{x \mid (x[f]y) \in c \text{ for some } f\}$$

Proposition 7. *Let* $y \in \mathcal{V}(c) \backslash \mathcal{V}_o$*. Then either* $\mathit{fathers}_c(y) \subseteq \mathcal{V}_o$*, or* $\mathit{fathers}_c(y)$ *is a singleton. In that second case, the only father is not in* $\mathcal{V}_o$*.*

Proof. This proposition is shown by induction on the rule applications from c_o that led to our constraint. The proposition is obviously true for c_o as there are no non-original variables in it. Let us now consider a step in the transformation, that is a rule that transformed a descendant $\exists X_1.c_1$ into $\exists X_2.c_2$. Assume that the proposition holds for c_1. There are several cases depending on the rule that constitutes our step.

- S-Eq transforms $c_1 = \exists x.(x \doteq y \wedge c)$ into $c_2 = c\{x \mapsto y\}$. From the side-condition of the rule and Prop. 4, we know that $x \neq y$ and that they are both original variables.
  The only variables whose fathers may have changed are y and the variables that had x for a father in c_1. y is not interesting for us here as it is an original variable.
  Say we have v non-original such that $x \in \texttt{fathers}_{c_1}(v)$. Then $\texttt{fathers}_{c_2}(v) = \texttt{fathers}_{c_1}(v) \backslash \{x\} \cup \{y\}$. Since x is original, by induction hypothesis, $\texttt{fathers}_{c_1}(v) \subseteq \mathcal{V}_o$. And since y is original, $\texttt{fathers}_{c_2}(v) \subseteq \mathcal{V}_o$.

- S-Feats transforms $c_1 = \exists z.x[f]y \wedge x[f]z \wedge c$ into $c_2 = x[f]y \wedge c\{z \mapsto y\}$. The only variables whose fathers may have changed are y and the variables that had z for a father in c_1. Let us handle these two subcases:
  - If y is original, then there is nothing to prove. Let us assume it is not. Then, from the side-condition of S-Feats, we can tell that z is not original either. From the clash-freeness of c_1, we can deduce that y and z are not in $\texttt{fathers}_{c_1}(y)$ nor $\texttt{fathers}_{c_1}(z)$. Then $\texttt{fathers}_{c_2}(y) = \texttt{fathers}_{c_1}(y) \cup \texttt{fathers}_{c_1}(z)$. Since y and z are both non-original, the induction hypothesis applies, and $\texttt{fathers}_{c_1}(y)$ and $\texttt{fathers}_{c_1}(z)$ are either both included in $\mathcal{V}_o$ or both singletons containing only x. In both cases, the invariant holds.
  - Let v be a non-original variable such that $z \in \texttt{fathers}_{c_1}(v)$. Then $\texttt{fathers}_{c_2}(v) = \texttt{fathers}_{c_1}(v) \setminus \{z\} \cup \{y\}$. From the induction hypothesis, we can say that either $\texttt{fathers}_{c_1}(v)$ is a singleton containing only z, in which case $\texttt{fathers}_{c_2}(v)$ is a singleton too and the invariant holds; or $\texttt{fathers}_{c_1}(v) \subseteq \mathcal{V}_o$. In that second case, we understand that z is original. From S-Feats's side-condition, we deduce that y has to be original too, and the invariant holds.
- P-Feat only adds one feature constraint $y[f]z$ in a constraint where there are $x[f]z$ and $x \tilde{\sim}_F y$, and thus only changes $\texttt{fathers}(z)$ with $\texttt{fathers}_{c_2}(z) = \texttt{fathers}_{c_1}(z) \cup \{y\}$. Since we have a similarity between them, x and y are originals (Proposition 5). But x is z's father in c_1, which means that $\texttt{fathers}_{c_1}(z) \subseteq \mathcal{V}_o$. Adding a new original father does not break the invariant: $\texttt{fathers}_{c_2}(z) \subseteq \mathcal{V}_o$.
- R-NFeat and R-NAbs both create one variable, but with only one father.
- R-NFence-Fence and E-NFence use the shortcut $x\langle F\rangle$ that contains a lot of variable introductions, but every time with only one father.
- R-NSim-Fence, R-NSim-Sim and E-NSim use the shortcut $x \not\tilde{\sim}_F y$ that contains a lot of variable introductions, but every time with only one father.
- S-Sims, P-Abs, P-Fence, P-Sim, R-NEq, P-NFence and P-NSim do not introduce variables nor change the fathers of existing ones.

Now that we have proven this rather technical lemma, we can use it to prove the following, which is more interesting and leads directly to what we want:

Proposition 8. *If $y \in \mathcal{V}_o \cap \mathcal{V}(c)$, then $\texttt{fathers}_c(y) \subseteq \mathcal{V}_o$. In other words, there is no feature constraint from a non-original variable to an original variable.*

Proof. This proposition is shown by induction on the transformation from c_o that led to our clause. The proposition is obviously true for c_o as there are no non-original variables. Let us now consider a step in the transformation, that is a rule that transformed a descendant $\exists X_1.c_1$ into $\exists X_2.c_2$. Assume that the proposition holds for c_1. There are several cases depending on the rule that constitutes our step.

- S-Eq is non applicable because there are no equalities involving local variables in the constraints c_o and thus in c_1.

- S-Feats transforms $c_1 = \exists z.x[f]y \wedge xfz \wedge c$ into $c_2 = x[f]y \wedge c\{z \mapsto y\}$. From the side-condition and the clash-freeness of c_1, we know that $x \neq y$, $y \neq z$ and $x \neq z$. There are four sub-cases depending on whether y and z are originals.
  - if y and z are originals, we can freely replace z by y in the feature constraints without breaking the proposition for c_2.
  - if y and z are non-originals, there is by induction hypothesis no feature constraint from any of them to an original variable in c_1, so this is still the case in c_2.
  - if y is original and z is non-original, then we obtain from the induction hypothesis that $\texttt{fathers}_{c_1}(y) \subseteq \mathcal{V}_o$. Since $x \in \texttt{fathers}_{c_1}(y)$, we have $x \in \mathcal{V}_o$. Since $x \in \texttt{fathers}_{c_1}(z)$ and $x \in \mathcal{V}_o$, we obtain by Proposition 7 that $\texttt{fathers}_{c_1}(z) \subseteq \mathcal{V}_o$. We conclude by $\texttt{fathers}_{c_2}(y) = \texttt{fathers}_{c_1}(y) \cup \texttt{fathers}_{c_1}(z)$.
  - if y is non-original and z is original, then we are in contradiction with the side condition: this case cannot happen.
- P-Feat adds one feature constraint $y[f]z$ to a constraint where there are $x[f]z$ and $x \dot{\sim}_F y$. y is original by Proposition 5. Adding a feature constraint $y[f]z$ where $y \in \mathcal{V}_o$ cannot invalidate the invariant.
- R-NFeat and R-NAbs both create one feature constraint but also introduce the necessary variable. We are thus sure that the proposition holds, as the introduced variable is non-original.
- R-NFence-Fence and E-Fence use the shortcut $x\langle F\rangle$ that contains a lot of feature constraints, but every time with a freshly created (thus non-original) variable.
- R-NSim-Fence, R-NSim-Sim and E-NSim use the shortcut $x \not\dot{\sim}_F y$ that contains a lot of feature constraints, but every time with a freshly created (thus non-original) variable.
- S-Sims, P-Abs, P-Fence, P-Sim, R-NEq, P-NFence and P-NSim do not introduce nor remove feature constraints.

As a consequence, no path leading to a original variable can contain a non-original variable. This means that the directed acyclic graph of the father relation has all the original variables in its top area, that is at a depth that is bounded by $\texttt{card}(\mathcal{V}_o)$. Combined with the fact that the fences and similarities do not leave this area, this will provide us with the weapons that we need to terminate our proof.

Proposition 9. *$d_c(x) < \texttt{card}(\mathcal{V}_o)$ if $x \in \mathcal{V}(c) \cap \mathcal{V}_o$.*

We now define the depth of atoms: it is the minimum of the depths of the variables involved. In particular, $d_c(x \not\sim_F y) = \texttt{min}(d_c(x), d_c(y))$.

Proposition 10. *$d_c(x \not\sim_F y) \leq \texttt{card}(\mathcal{V}_o)$*

Proof. We prove this property by induction on the transformation that led from c_0 to c. It is obviously true for c_0 because there are no non-original variables in

c_0. Now, assume that the last step of the transformation leads from $\exists X_1.c_1$ to $\exists X_2.c_2$.

By induction hypothesis, we have for each negated similarity $x \not\sim_F y$ in c_1 only two possible cases: either $d_{c_1}(x \not\sim_F y) < \texttt{card}(\mathcal{V}_o)$ or $d_{c_1}(x \not\sim_F y) = \texttt{card}(\mathcal{V}_o)$.

1. If $(x \not\sim_F y) \in c_1$ with $d_{c_1}(x \not\sim_F y) = \texttt{card}(\mathcal{V}_o)$ then $d_{c_1}(x), d_{c_1}(y) \geq \texttt{card}(\mathcal{V}_o)$, hence $x, y \notin \mathcal{V}_o$ by Proposition 9. By Proposition 5 there is no similarity contraint for x or y in c_1, and by Proposition 6 there is no fence constraint for x or y in c_1. Hence, no rule producing a new negated similarity can apply to $x \not\sim_F y$.
2. If $(x \not\sim_F y) \in c_1$ with $d_{c_1}(x \not\sim_F y) < \texttt{card}(\mathcal{V}_o)$ then the negated similarity may either travel through a similarity with P-NSim, in which case it still touches an original variable and its depth stays smaller than $\texttt{card}(\mathcal{V}_o)$, or it can be rewritten with our rules R-NSim-Fence, R-NSim-Sim or E-NSim, in which case it reaches a higher depth. However, each of these rules can produced a negated similarity with depth at most $d_{c_1}(x \not\sim_F y) + 1$. Since $d_{c_1}(x \not\sim_F y) < \texttt{card}(\mathcal{V}_o)$, each of the newly produced negated similarities has a depth in c_1 which is $\leq \texttt{card}(\mathcal{V}_o)$.

A.3 A decreasing measure

		S-Eq	S-Feats	S-Feats-Global	S-Sim	P-Feat	P-Abs	P-Fence	P-Sim	R-NEq	R-NFeat	R-NAbs	R-NFence-Fence	R-NSim-Fence	R-NSim-Sim	E-NFence	E-NSim
1	Number of neg. eq.	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
2	Number of neg. feat. constr.	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
3	Number of neg. abs.	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
4	Missing fences	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
5	Combined sims.	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
6	Depth of neg. sims.	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\leq$	$\leq$	$\cdot$	$\cdot$	$>$	$>$	$\cdot$	$\cdot$ | $>$
7	Missing feats. in neg. sims.	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\leq$	$\leq$	$\cdot$	$\cdot$	$\geq$	$\geq$	$\cdot$	$>$ | $\geq$
8	Missing feats. in neg. fences	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$>$	$\leq$	$\cdot$	$>$	$\cdot$
9	Missing equalities	$\geq$	$\geq$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$<$	$<$	$\leq$	$\leq$	$\leq$	$\leq$	$\leq$
10	Missing feature constraints	$\geq$	$\geq$	$<$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$<$	$<$	$\leq$	$\leq$	$\leq$	$\leq$	$\leq$
11	Missing absences	$\geq$	$\geq$	$\cdot$	$\cdot$	$\cdot$	$>$	$\cdot$	$\cdot$	$\cdot$	$<$	$<$	$\leq$	$\leq$	$\leq$	$\leq$	$\leq$
12	Number of literals	$>$	$>$	$\cdot$	$>$	$<$	$<$	$<$	$<$	$\cdot$	$\leq$	$\cdot$	$\cdot$	$\leq$	$<$	$\cdot$	$\leq$

1. Nothing introduces negative equalities, ever. The only rule acting on them is R-NEq that removes them.
2. Nothing introduces negative feature constraint. The only rule acting on them is R-NFeat that removes them.
3. Nothing introduces negative absences. The only rule acting on them is R-NAbs that removes them.

4. "Missing fences"

$$\{(x, F) \mid x \in \mathcal{V}(c) \cap \mathcal{V}_o;\ \text{no } x[F] \text{ in } c\}$$

This is the set of all the fences that could exist but that do not. This set clearly diminishes when we apply P-Fence. It is not touched by the rules that introduce variables, because those only introduce non original variables. It cannot increase when applying S-Eq or S-Feats, although those may remove may remove existing fences. However, when they do, this is because they removed the original variable to which the fence was attached.

5. "Combined similarities"

$$(x, y) \in (\mathcal{V}(c) \cap \mathcal{V}_o)^2 \mapsto \bigcap_{(x \tilde{\sim}_F y) \in c} F$$

We consider that a function is strictly smaller than an other if its domain is strictly included in the other's domain, or if their domain are equal, all values are smaller and at least one value is strictly smaller.
This function decreases with P-Sim but is left constant with S-Sim (it has been crafted precisely for that). Once again, if S-Eq or S-Feats were to remove a similarity and thus potentially change the values of these intersections, then an original variable would have disappeared, strictly decreasing the domain.

6. "Depth of the negated similarities" of c

$$\{\{\texttt{card}(\mathcal{V}_o) - d_c(x \not\sim_F y) \mid (x \not\sim_F y) \in c\}\}$$

We consider as an order for these multisets the lifting of the order in natural numbers: we allow ourselves to add a finite numbers to the multiset as long as we remove a number that is strictly greater than all of them. This is a well founded order if it is the lifting of a well founded order. And this is the case, because Prop. 10 tells us that all the numbers in it are non-negative.
This multiset is increased by R-NEq and R-NFeat because they add a new negative similarity (It would be increased by P-NSim too). It decreases in R-NSim-Fence and R-NSim-Sim because although we potentially add a negated similarity, we remove one the is higher (and whose depth is thus smaller). The case of E-NSim is a bit particular, because or measure stays constant in the first part of the disjunction while it decreases in the second part. The next measure will take care of that.

7. Missing features in negative similarities

$$\{\{\mathcal{F}(c) - F \mid (x \not\sim_F y) \in c\}\}$$

By definition, all the sets of the negated similarities are included in $\mathcal{F}(c)$.

8. Missing features in negative fences

$$\{\{\mathcal{F}(c) - F \mid (\neg x[F]) \in c\}\}$$

This multiset decreases with the rules R-NFence-Fence because this one removes an element, and with E-NFence because it either removes a negative fence or replace it by a larger one.

9. Missing equalities
$$\{(x, y) \mid (x \doteq y) \notin c\}$$
This set decreases with S-Feats-Global. It increases a lot with all the rules adding variables, but they are no threat to us since we have already taken care of them.
10. Missing feature constraints
$$\{(x, f, y) \mid (x[f]y) \notin c\}$$
This set decreases with P-Feat.
11. Missing absences
$$\{(x, f) \mid (x[f] \uparrow) \notin c\}$$
This set decreases with P-Abs.
12. Number of literals
This number decreases with S-Eq and S-Feats.

B Proof of Lemma 6 (elimination)

B.1 Normal form

Lemma 8. *The output of* `normalize` *is a dnf where each conjunction is in normal form for R.*

We have to prove that P-NFence and P-NSim, when provided with a clause that is in normal form for the rules in R_1, conserves the normal form.

Let us take a constraint in normal form for R_1 and show that the application of one of the two rules in R_2 keeps it in normal form.

We will start with P-NFence. Clearly, since P-NFence only creates a new negative fence, the only rules it could trigger are R-NFence-Fence and E-NFence. Let us then take $c_1 = x \mathrel{\dot{\sim}_F} y \wedge \neg x[G] \wedge c$ that rewrites into $c_2 = x \mathrel{\dot{\sim}_F} y \wedge \neg x[G] \wedge \neg y[G] \wedge c$.

- R-NFence-Fence: c_1 is in normal form for R-NFence-Fence by hypothesis, so c cannot contain a positive fence for x. c_1 is also in normal form for P-Fence, so c cannot contain a positive fence for y either. Hence, c_2 is in normal form for R-NFence-Fence.
- E-NFence: c_1 is in normal form for P-Sim and S-Sim, so for each similarity $y \mathrel{\dot{\sim}_H} y'$ in c there is an I such that $x \mathrel{\dot{\sim}_I} y'$ is in c. For the same reason, we have that $H \subseteq F \cup I$. Since c_1 is in normal form for E-NFence, $F \subseteq G$ and $I \subseteq G$. Thus $H \subseteq G$, and c_2 is in normal form for E-NFence.

Let us continue with P-NSim. Since it only creates a negative similarity, the only rules it could trigger are R-NSim-Fence, R-NSim-Sim and E-NSim. Let us then take $c_1 = x \mathrel{\dot{\sim}_F} y \wedge x \not\sim_G z \wedge c$ that rewrites into $c_2 = x \mathrel{\dot{\sim}_F} y \wedge x \not\sim_G z \wedge y \not\sim_G z \wedge c$.

- R-NSim-Fence: c_1 is in normal form for R-NSim-Fence by hypothesis, so c cannot contain a positive fence for x. c_1 is also in normal form for P-Fence, so c cannot contain a positive fence for y either. Hence, c_2 is in normal form for R-NSim-Fence.

- R-NSim-Sim: c_1 is in normal form for R-NSim-Sim by hypothesis, so c cannot contain a positive similarity for (x, z). c_1 is also in normal form for P-Sim, so c cannot contain a positive fence for (y, z) either. Hence, c_2 is in normal form for R-NSim-Sim.
- E-NSim: c_1 is in normal form for P-Sim and S-Sim, so for each similarity $y \dot{\sim}_H y'$ in c there is an I such that $x \dot{\sim}_I y'$ is in c. For the same reason, we have that $H \subseteq F \cup I$. Since c_1 is in normal form for E-NSim, $F \subseteq G$ and $I \subseteq G$. Thus $H \subseteq G$, and c_2 is in normal form for E-NSim.

Thus, in the following, we can consider that the output of `normalize` is in normal form for all the rules.

B.2 Introduction

Let $\exists X.c$ be a clause, that is an element of a dnf returned by the function `normalize`. Since $c = g_c \wedge l_c$ where $\mathcal{V}(g_c) \cap X = \emptyset$, we have to show that

$$\mathcal{FT} \models \tilde{\forall}(\exists X.(g_c \wedge l_c) \leftrightarrow g_c)$$

The only non-trivial implication to show is

$$\mathcal{FT} \models \tilde{\forall}(g_c \rightarrow \exists X.l_c)$$

Let $Y = \mathcal{V}(g_c)$, and assume a valuation $\rho_Y : Y \rightarrow \mathcal{FT}$ such that $\mathcal{FT}, \rho_Y \models g_c$. We will show that we can extend it to $\rho_{Y \cup X} : Y \cup X \rightarrow \mathcal{FT}$ such that $\mathcal{FT}, \rho_{Y \cup X} \models l_c$.

Note that l_c cannot contain a conjunction $x[f]y \wedge x[f]z$. Otherwise, we obtain a contradiction:

- If $y, z \notin X$ then (S-Feats-Global) applies.
- If $y \in X$ and $z \notin X$, then $z \in \mathcal{V}_o$ (by Prop. 3 in App. A), hence (S-Feats) applies. The same reasoning applies when $y \notin X$ and $z \in X$.
- If $y, z \in X$ then (S-Feats) applies since we can choose the replacement such that the side condition is satisfied.

Recall that any clause in a dnf is clash-free. Let $d \subseteq c$ contain all the D literals of c. Due to clash pattern Decs there exists a valuation β such that $\mathcal{D}, \beta \models d$.

In order to construct the extension of ρ_Y to X, choose a strict total order $\sqsubset$ over $Y \cup X$ such that

1. $y \sqsubset x$ whenever $(x[f]y) \in c$,
2. $y \sqsubset x$ whenever $y \in Y$ and $x \in X$.

This is possible because there are no feature cycles in c, due to clash pattern C-Cycle. Hence we can start with the partial order defined by $y \sqsubset x$ iff $x[f]y \in c$, and complete it into a total order by taking care to range all global variables before the local variables, which is possible due to the hypothesis of Lemma 6 that there is no feature from a global variable to a local variable.

B.3 Construction

We now define $\rho_{Y \cup X}$ by induction on X following $\sqsubset$ ensuring that we keep satisfied all the literals containing variables where our valuation is defined. The base case of the induction is ρ_Y, which is already defined and satisfies its literals by hypothesis. In the induction case, we take $x \in X$, we define the set of variables that are smaller than x, $Z = \{z \in Y \cup X \mid z \sqsubset x\}$ and we assume that we have ρ_Z already defined such that it satisfies all the literals about variables in Z.

We are going to extend it by defining $\rho_{Z \cup \{x\}}$. We define the following partial map m_x for x:

$$m_x = \{(f, \rho_Z(y)) \mid (x[f]y) \in c\}$$

Note that this defines a partial function, because there cannot be f, y and y' such that $(x[f]y) \in c$, $(x[f]y') \in c$ and $y \neq y'$ due to clash pattern S-Feats. Consider now the set of all the variables that are smaller than x and that are in a similarity relation with x:

$$down(x) = \{y \mid y \sqsubset x, (x \mathrel{\dot{\sim}}_H y) \in c \text{ for some } H\}$$

We will now define m'_x using m_x and $\rho_Z(y)$ for all the $y \in down(x)$. There are three cases depending on whether $down(x)$ is empty, and depending on whether there is a fence constraint for x.

1. If $down(x) = \emptyset$ and there is some fence constraint $(x[F]) \in c$, then we define $m'_x = m_x$.
2. If $down(x) = \emptyset$ and there is no fence constraint $(x[F]) \in c$, then we choose a fresh feature h_x which
   - does not occur in c (not even in a fence or similarity),
   - does not occur in $\mathtt{dom}(\overline{\rho_Y(y)})$, for any $y \in Y$,
   - is different from h_z, for any $z \sqsubset x$.

   This is possible since the mapping of a feature tree is required to have a finite domain, and since we have an infinite supply of feature symbols. Hence, the set
   $$\mathcal{F} \setminus \bigcup_{y \in Y} \mathtt{dom}(\overline{\rho_Y(y)})$$
   is infinite. Let $d \in \mathcal{D}$ be some arbitrary node decoration. We define $m'_x = m_x \cup \{(h_x, (d, \emptyset))\}$, that is m' is obtained from m by adding an edge labeled h_x going to the empty tree. This still defines a function because h_x is different from all the features encountered so far.
3. If $down(x) \neq \emptyset$ then we define
   $$m'_x = m_x \cup \bigcup_{\substack{z \sqsubset x \\ (x \mathrel{\dot{\sim}}_H z) \in c}} \overline{\rho_Z(z)} \restriction_{H^c}$$
   where, when $\overline{\rho_Z(z)} \restriction_{H^c}$ is the restriction of $\overline{\rho_Z(z)}$ to the complement of H in $\mathcal{F}$. This union is not disjoint, so we have to show that m'_x is well-defined as a function.

(a) Assume that $f \in \mathtt{dom}(m_x)$ and $f \in \mathtt{dom}(\overline{\rho_Z(z)} \restriction_{H^c})$, with $z \sqsubset x$ and $(x \mathrel{\dot\sim}_H z) \in c$. By definition of m_x, there must be a $(x[f]y) \in c$, with $y \sqsubset x$ and $m_x(f) = \rho_Z(y)$.
Since $f \in \mathtt{dom}(\overline{\rho_Z(z)} \restriction_{H^c})$, we have that $f \notin H$. By rule P-Feat, It must be the case that $(z[f]y) \in c$. Since $y, z \sqsubset x$, we obtain by induction hypothesis that $\mathcal{FT}, \rho_Z \models z[f]y$, that is

$$\overline{\rho_Z(z)} \restriction_{H^c} (f) = \overline{\rho_Z(z)}(f) = \rho_Z(y)$$

(b) Assume that $f \in \mathtt{dom}(\overline{\rho_Z(z)} \restriction_{H^c})$ and $f \in \mathtt{dom}(\overline{\rho_Z(z')} \restriction_{H'^c})$, with $z, z' \sqsubset x$, $(x \mathrel{\dot\sim}_H z) \in c$, $(x \mathrel{\dot\sim}_{H'} z') \in c$, and $z \neq z'$ or $H \neq H'$.
By rule P-Sim, there is some $I \subseteq H \cup H'$ such that $(z \mathrel{\dot\sim}_I z') \in c$. Since $f \notin H, H'$, we also have that $f \notin I$. Since $z, z' \sqsubset x$, we have by induction hypothesis that $\mathcal{FT}, \rho_Z \models z \mathrel{\dot\sim}_I z'$. Since $f \notin I$, this means that

$$\overline{\rho_Z(z)} \restriction_{H^c} (f) = \overline{\rho_Z(z)}(f) \restriction_{I^c} (f) = \overline{\rho_Z(z')}(f) \restriction_{I^c} (f) = \overline{\rho_Z(z')} \restriction_{H'^c} (f)$$

Finally, we define

$$\rho_{Z \cup \{x\}}(z) = \begin{cases} (\beta(x), m'_x) & \text{if } z = x \\ \rho(z) & \text{if } z \sqsubset x \end{cases}$$

B.4 Verification

Let us now show that that the invariant is satisfied, that is that $\mathcal{FT}, \rho_{Z \cup \{x\}} \models l$ for every $l \in c$ such that $z \sqsubset x$ for every $z \in \mathcal{V}(c)$. By induction hypothesis, we may restrict ourselves to the case where $x \in \mathcal{V}(c)$. We distinguish the different possible forms of a literal c:

$x \doteq y$ By rule S-Eq, this is only possible when $x = y$ or when x and y are global. The first case is always trivially satisfied, the second is satisfied by ρ_Y by hypothesis.

$x \not\doteq y$ Eliminated by the system (R-NEq).

$x[f]y$ This is immediate considering the way m'_x was defined.

$\neg x[f]y$ Eliminated by the system (R-NFeat).

$x[f] \uparrow$ We distinguish the three cases in the construction of m'_x:

1. There cannot be a literal $(x[f]y) \in c$, thanks to rule (C-Feat-Abs). Thus, m_x is not defined for f.
2. In addition to case (1), note that $f \neq h_x$ due to the way h_x was chosen. Thus, m'_x is not defined for f.
3. For all $z \sqsubset x$ with $(x \mathrel{\dot\sim}_H z) \in c$, if $f \notin H$, then $(z[f] \uparrow) \in c$ (P-Abs). Since this last atom is satisfied by induction hypothesis, $f \notin \mathtt{dom}(\overline{\rho_Z(z)})$. That, combined with the point (1) gives us: $f \notin \mathtt{dom}(\overline{\rho_{Z \cup \{x\}}(x)})$.

$\neg x[f] \uparrow$ Eliminated by the system (R-NAbs).

$x[F]$ We distinguish the three cases in the construction of $\rho(x)$:

1. There cannot be a literal $(x[f]y) \in c$ with $f \notin F$ for x (C-Feat-Fence). Thus, $\mathtt{dom}(m_x) \subseteq F$.

2. Does not apply.
3. In addition to case (1), for all $z \sqsubset x$ with $(x \mathrel{\dot\sim}_H z) \in c$, we have that $(z[H \cup F]) \in c$ (P-FENCE. Since it is satisfied by ρ_Z by induction hypothesis, $\mathtt{dom}(\overline{\rho_Z(z)}) \subseteq H \cup F$, and hence $\mathtt{dom}(\overline{\rho_Z(z)} \restriction_{H^c}) \subseteq F$. This, together with the reasoning of case (1), give us $\mathtt{dom}(\overline{\rho_{Z\cup\{x\}}(x)}) \subseteq F$.

$\neg x[F]$ We distinguish the three cases in the construction of $\rho(x)$.

1. This rule does not apply: By rule R-NFENCE-FENCE, we cannot have both a positive and negative fence constraint for the same variable.
2. In this case, $\mathtt{dom}(\overline{\rho_{Z\cup\{x\}}(x)})$ contains the fresh feature $h_x \notin F$.
3. In this case there is a variable $z \sqsubset x$, such that $(x \mathrel{\dot\sim}_H z) \in c$. By E-NFENCE we must have that $H \subseteq F$ (Note that this rule generates several alternatives: one containing an enlarged negative fence, and the other ones where the negative fence is replaced by an absence constraint. Hence, in presence of the negative fence, we must be in the first alternative). Then, by P-NFENCE, $(\neg z[F]) \in c$. By induction hypothesis, $\mathcal{FT}, \rho_Z \models \neg z[F]$, that is there is $f \in \mathtt{dom}(\overline{\rho_Z(z)})$ with $f \notin F$, hence $f \notin H$. By construction of m'_x this means that $f \in \mathtt{dom}(\overline{\rho_{Z\cup\{x\}}(x)})$, that is $\mathcal{FT}, \rho_{Z\cup\{x\}} \models \neg x[F]$.

$x \mathrel{\dot\sim}_F y$ Satisfied by construction.

$x \not\mathrel{\dot\sim}_F y$ We distinguish the three cases in the construction of $\rho(x)$.

1. This case does not apply, since by R-NSIM-FENCE, x cannot have both a negated similarity and a fence constraint.
2. By construction, there is a fresh feature $h_x \in \mathtt{dom}(\overline{\rho_{Z\cup\{x\}}(x)})$ Since $y \neq x$ (C-NSIM-REFL) and since there is no similarity between x and y (R-NSIM-SIM), we have that $h_x \notin \mathtt{dom}(\overline{\rho_Z(y)})$. Since $h_x \notin F$, this means that $\mathcal{FT}, \rho_{Z\cup\{x\}} \models x \not\mathrel{\dot\sim}_F y$.
3. In this case there is variable $z \sqsubset x$ and H such that $(x \mathrel{\dot\sim}_H z) \in c$. As we have seen in the previous case, this means that

$$\mathcal{FT}, \rho_{Z\cup\{x\}} \models x \mathrel{\dot\sim}_H z \tag{1}$$

By (E-NSIM) we must have that $H \subseteq F$ (by the same reasoning as above, we must be in the first of the alternatives introduced by this rule). Then, by rule (P-NSIM), $H \subseteq F$ and $(z \not\mathrel{\dot\sim}_F y) \in c$. By induction hypothesis,

$$\mathcal{FT}, \rho_Z \models z \not\mathrel{\dot\sim}_F y \tag{2}$$

Since $H \subseteq F$, it follows from (1) that

$$\mathcal{FT}, \rho_{Z\cup\{x\}} \models x \mathrel{\dot\sim}_F z \tag{3}$$

Finally, we conclude from (3) and (2) that

$$\mathcal{FT}, \rho_{Z\cup\{x\}} \models x \not\mathrel{\dot\sim}_F y \tag{4}$$

C Optimisation

We can add a few simplification rules that do not affect the outcome of the algorithm nor its properties. See Fig. 7.

S-Eq-Global	$x \doteq y \wedge c$	$\Rightarrow$	$x \doteq y \wedge c\{x \mapsto y\}$	$(x, y \in \mathcal{V}(c))$
S-Eq-Refl	$x \doteq x \wedge c$	$\Rightarrow$	c	
S-Sim-Refl	$x \mathrel{\dot{\sim}}_F x \wedge c$	$\Rightarrow$	c	
S-Fences	$x[F] \wedge x[G] \wedge c$	$\Rightarrow$	$x[F \cap G] \wedge c$	
S-Fence-Abs	$x[F] \wedge x[f]\uparrow \wedge c$	$\Rightarrow$	$x[F \setminus \{f\}] \wedge c$	
S-NFence-Abs	$\neg x[F] \wedge x[f]\uparrow \wedge c$	$\Rightarrow$	$\neg x[F \cup \{f\}] \wedge c$	
S-NFence-Feat	$\neg x[F] \wedge x[f]y \wedge c$	$\Rightarrow$	$x[f]y \wedge c$	$(f \notin F)$
S-NFences	$\neg x[F] \wedge \neg x[G] \wedge c$	$\Rightarrow$	$\neg x[G] \wedge c$	$(F \subseteq G)$
S-NSims	$x \not\sim_F y \wedge x \not\sim_G y \wedge c$	$\Rightarrow$	$x \not\sim_G y \wedge c$	$(F \subseteq G)$

Fig. 8. Optional simplification rules.