A Foundation for Flow-Based Program Matching Using Temporal Logic and Model Checking
Julien Brunel, Damien Doligez, René Hansen, Julia Lawall, Gilles Muller

To cite this version:
Julien Brunel, Damien Doligez, René Hansen, Julia Lawall, Gilles Muller. A Foundation for Flow-Based Program Matching Using Temporal Logic and Model Checking. 2008. hal-00297708v2

HAL Id: hal-00297708
https://hal.archives-ouvertes.fr/hal-00297708v2
Submitted on 17 Jul 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Foundation for Flow-Based Program Matching
Using Temporal Logic and Model Checking

Julien Brunel
DIKU, University of Copenhagen,
Denmark
brunel@diku.dk

Damien Doligez
INRIA, Gallium Project,
France
damien.doligez@inria.fr

René Rydhof Hansen
Department of Computer Science
Aalborg University, Denmark
rrh@cs.aau.dk

Julia L. Lawall
DIKU, University of Copenhagen, Denmark
julia@diku.dk

Gilles Muller
École des Mines de Nantes, France
Gilles.Muller@emn.fr

Research report 08/2/INFO, Ecole des Mines de Nantes
July 16, 2008

Reasoning about program control-flow paths is an important functionality of a number of recent program matching languages and associated searching and transformation tools. Temporal logic provides a well-defined means of expressing properties of control-flow paths in programs, and indeed an extension of the temporal logic CTL has been applied to the problem of specifying and verifying the transformations commonly performed by optimizing compilers. Nevertheless, in developing the Coccinelle program transformation tool for performing Linux collateral evolutions in systems code, we have found that existing variants of CTL do not adequately support rules that transform subterms other than the ones matching an entire formula. Being able to transform any of the subterms of a matched term seems essential in the domain targeted by Coccinelle.

In this technical report, we propose an extension to CTL named CTL-VW (CTL with variables and witnesses) that is a suitable basis for the semantics and implementation of the Coccinelle’s program matching language. Our extension to CTL includes existential quantification over program fragments, which allows metavariables in the program matching language to range over different values within different control-flow paths, and a notion of witnesses that record such existential bindings for use in the subsequent program transformation process. We formalize CTL-VW and describe its use in the context of Coccinelle. We then assess the performance of the approach in practice, using a transformation rule that fixes several reference count bugs in Linux code.
1. Introduction

Program matching is the process of searching within the source code of a program for code fragments matching a given pattern, described using some language. Recently, program matching languages that combine descriptions of code fragments with information about the control-flow paths between them have been found useful in specifying rules for program manipulation tasks such as compiler optimizations [16, 17], bug finding [10], refactorings [23], and evolution [20].

In recent work, we have developed the transformation system Coccinelle [20], which provides the language SmPL (Semantic Patch Language) for specifying desired matches and transformations. Coccinelle is targeted towards performing collateral evolutions in systems code. Such evolutions comprise the changes that are needed in client code in response to evolutions in library APIs, and may include modifications such as renaming a function, adding a function argument whose value is somehow context-dependent, and reorganizing a data structure. Collateral evolutions may involve fragments of code that are scattered throughout a function or a file, such as a function call and its corresponding error-handling code, and thus it is often necessary to take into account the control-flow relationships between the relevant code fragments. Beyond collateral evolutions, we have also found Coccinelle useful for finding and fixing bugs in the use of API functions [18], for which control-flow relationships must also be taken into account.

In developing Coccinelle, it was necessary to choose a foundation on which to base the matching process. In the context of specifying compiler optimizations, Lacey and De Moor have observed that the temporal logic CTL can provide a convenient foundation for the semantics of program matching languages that take control-flow paths into account, because it is designed for reasoning about paths [16]. In this setting, a pattern is compiled into a formula of the logic and a program is translated into its control-flow graph. Model checking [7, 12] is then used to match the formula against the control-flow graph, thus identifying where the pattern matches in the program. To support the specification of compiler optimizations, Lacey and De Moor extended CTL first with predicates over free variables, producing the logic CTL-FV [17], and then with predicates over variables that may be existentially quantified, producing a logic that we refer to as CTL-V [15]. The former allows collecting information about the program during the matching process, via variable bindings, and imposing constraints on the allowed subterms when a given variable is matched more than once. The latter enables localizing collected information and imposed constraints to within individual subsets of the possible control-flow paths, such as the different branches of a conditional statement.

The needs of Coccinelle, however, have made apparent some inadequacies of CTL-V as the back end of a program matching language. While the use of existentially quantified variables allows variables to have different values within different control flow paths, the semantics of CTL-V does not provide a means of retrieving the values of such variables for use during subsequent transformation. Furthermore, CTL-V model checking, like standard CTL model checking, only provides information about the state at which the entire formula is satisfied, and not about the results for subformulas that contribute to a successful match. Information about such intermediate results, however, is important when a single formula can describe transformations of subterms, as is needed for Coccinelle.

In this paper, we present a variant of CTL named CTL-VW (CTL with variables and witnesses), that addresses the above needs for program matching. The syntax of CTL-VW is the same as that of CTL-V, extending that of CTL with predicates defined over metavariables that can be existentially quantified over the set of program fragments. The semantics of CTL-VW maintains a collection of witnesses that record the states and bindings that satisfy existentially quantified subformulas. We exploit these witnesses both to record variable bindings and to identify states at which transformation should take place. In practice, we have used CTL-VW as the basis of the implementation of Coccinelle. We have used Coccinelle to implement over 60 collateral evolutions, affecting in total over 5800 files in various recent versions of Linux [20]. We have also used Coccinelle to find over 45 bugs in Linux code.

The contributions of this paper are as follows:

- We present the semantics of CTL-VW. We show that this semantics is a conservative extension of that of CTL-V.
- We present in two steps a model checking algorithm for CTL-VW. In the first step, we extend the CTL model checking algorithm with environments, producing a new model checking algorithm for both CTL-FV and CTL-V. In the second step, we further extend this model checking algorithm with witnesses, for CTL-VW.
- We show that the model checking algorithm for CTL-VW is sound and complete with respect to the semantics of CTL-VW.
- We show how to translate the core of Coccinelle’s program matching language SmPL into CTL-VW. This core is sufficient to express a semantic patch for finding and fixing some bugs in the use of reference counts in Linux code. The resulting corrections have been validated by Linux experts and accepted into the Linux kernel. Our examples show that the CTL-VW based implementation of Coccinelle is efficient enough to be usable on a 1.4GHz laptop.

The rest of the paper is organized as follows. Section 2 briefly presents Coccinelle in terms of a real example that highlights the above-cited requirements. Section 3 describes

1 http://www.emn.fr/x-info/coccinelle/#impact
2. Overview of Coccinelle

Our motivation for this work was to be able to use CTL as a foundation for the semantics and implementation of SmPL, the program matching and transformation language of Coccinelle [20]. The syntax of SmPL is derived from that of a Linux patch file [19], which is a notation familiar to Linux programmers. Unlike a standard patch, however, which is text-based, a semantic patch takes into account the semantics of the matched code, in particular its intraprocedural control-flow. From the point of view of CTL, the relevant features of a semantic patch are that it may describe a complex region of code, specified in terms of program fragments that should be connected by control-flow paths, and that it may be needed to specify transformations at any point within the described region. Furthermore, the region may involve multiple control-flow paths, e.g., due to conditionals, and different control-flow path may involve different computations, each of which may be relevant to the transformation process.

In this section, we present the SmPL language via a short example that illustrates the above issues. We defer a more formal presentation of the code language of SmPL to Section 6, where we show how to define its semantics by translation into the CTL-VW logic developed in the next three sections.

2.1 A simple SmPL sample

Managing reference counts is a common source of errors in C code. In particular, we have observed that Linux error handling code sometimes does not appropriately decrement reference counts for objects acquired in the current function. Figure 1 shows a semantic patch type_ref for correcting such problems involving the function of_find_node_by_type. This function increases a reference count that should subsequently be decremented by calling the function of_node_put. The semantic patch in Figure 1 inserts a call to of_node_put before a return in error handling code (indicated by a return of a negative value), when there was no such call previously, and when it is not possible that the value returned by of_find_node_by_type has been saved in a more permanent manner.

The type_ref semantic patch consists of a single rule, which first declares a collection of metavariables and then defines a transformation specification. The metavariables are designated according to the kind of terms they can match, such as a statement, an identifier, or an expression (line 2). An expression metavariable can be further constrained by its type (line 3). The transformation specification essentially has the form of C code, except that lines to remove are annotated with – in the first column, and lines to add are annotated with + (line 13). A transformation specification can also use dots, “…” (line 7), describing an arbitrary sequence of instructions within a control-flow path. Dots may be modified with a when clause (lines 9-11), indicating a pattern that should not occur anywhere within the matched sequence. Finally, lines 12-19 specify a disjunction of patterns, of the form ( pat1 | ... | patn ).

Full SmPL, as implemented in Coccinelle, provides many features not illustrated by this example, such as multiple rules, subpatterns that match 0 or more times, dots within arbitrary subterms, and the ability to use both universal and existential path quantification [20]. All of these features can be encoded similarly to the basic strategies presented in Section 6.

2.2 Assessment

Our goal is to encode the semantic patch of Figure 1 as a single formula of a CTL-like logic. For this, the main issues are 1) to manage the metavariables nx, C, etc., 2) to allow metavariables to match different terms along different control-flow paths, as e.g., the return value C may be different at each return site when there are multiple returns under conditionals, 3) to record the various bindings of the metavariables within the different control-flow paths, as e.g., the binding of C may be needed to transform the enclosing return, and 4) to record the sites where transformation is needed. For the last point, in our case, the complete formula representing the semantic patch would match at a call to of_find_node_by_type, but transformation is required at the various matches of the subformula representing return –C; . A semantic patch may indeed specify a transformation anywhere within the matched term. Among these requirements, CTL-FV provides only a representation of metavariables, and CTL-V additionaly allows metavariables to match different terms along different control-flow paths. We thus develop the logic
3. Background

In this section, we present the theoretical background for the work in the present paper. First, we briefly review the syntax, semantics, and algorithmic implementation of the well-known Computational Tree Logic (CTL) [7]. Next, we present a variant of CTL called CTL-FV (CTL with free variables) used by Lacey et al. to show the correctness of some classical compiler optimizations [17]. Finally, we present CTL-V (CTL with variables), a CTL variant with existentially quantified variables, and formally define its semantics.

3.1 Computational Tree Logic

Computational Tree Logic is a temporal logic based on the notion of branching time [7]. It has been implemented in a variety of model checkers and used to model check properties ranging from hardware verification [11, 14] to program analysis [22].

In what follows we briefly review the syntax and semantics of CTL without going into detail. For a textbook treatment, see [12]. CTL is a logic for reasoning about states and the paths between them. Its syntax is as follows:

\[ \phi ::= p | \phi \land \phi | \phi \lor \phi | \neg \phi | AX\phi | EX\phi | A[\phi U \psi] | E[\phi U \psi] \]

Intuitively, the formulas \( p \) (propositions), \( \phi \land \phi \), \( \phi \lor \phi \), and \( \neg \phi \) are the same as in propositional logic, and permit reasoning about the properties of a given state. The remaining formulas permit reasoning about paths. \( AX\phi \) holds if all the successors of the current state satisfy \( \phi \), while \( EX\phi \) holds if at least one successor of the current state satisfies \( \phi \). \( A[\phi U \psi] \) expresses that along all paths there is some state where \( \phi \) holds, and requires that \( \psi \) hold at all preceding states. \( E[\phi U \psi] \) is similar, except that there need be only one path with these properties. For example, in Figure 2a, the formula \( A[(f(1) \lor g(2)) U h(1, 2)] \) holds at state 1, because all paths eventually reach a state, i.e., state 3 or 5, where the proposition \( h(1, 2) \) holds, and at all previous states, i.e., states 1, 2, and 4, either \( f(1) \) or \( g(2) \) holds. On the other hand, in Figure 2c, \( AX(g(2)) \) does not hold at state 1, because \( g(2) \) does not hold at one of the neighbors, state 4.

The semantics of CTL is defined over a model:

**Definition 1.** A model is a triple, \((\text{States}, \rightarrow, \text{Label})\), where States is a finite set of states; \( \rightarrow \subseteq \text{States} \times \text{States} \) is the successor relation such that \( \forall s.3r.s \rightarrow r \), i.e., every state has at least one successor; and Label : States → \( \mathcal{P}(\text{Atom}) \) is a labelling function that assigns a set of atomic propositions to each state.

The set of infinite paths starting in state \( s \) is denoted \( \text{Path}(s) \).

For a path \( \pi = s_0 \rightarrow s_1 \rightarrow \cdots \rightarrow s_j \rightarrow \cdots \) the \( j \)th element is denoted by \( \pi[j] \), i.e., \( \pi[j] = s_j \). Note that the set of paths may be infinite. Finally, the next function computes the set of successors of a given state: \( \text{next}(s) = \{s' | s \rightarrow s'\} \).

We now formally define the semantics for CTL as a set of judgements of the form \( M, s \models \phi \), where \( M \) is a model, \( s \in \text{States} \) and \( \phi \) is a CTL formula. For convenience, we elide the model in the judgements:

\[ s \models p \iff p \in \text{Label}(s) \]
\[ s \models \phi \land \psi \iff s \models \phi \land s \models \psi \]
\[ s \models \phi \lor \psi \iff s \models \phi \lor s \models \psi \]
\[ s \models \neg \phi \iff s \not\models \phi \]
\[ s \models AX\phi \iff \forall s' \in \text{next}(s), s' \models \phi \]
\[ s \models EX\phi \iff \exists s' \in \text{next}(s), s' \models \phi \]
\[ s \models A[\phi U \psi] \iff \forall \pi \in \text{Path}(s), \exists i \geq 0. \pi[i] = \phi \land \forall 0 \leq j < i. \pi[j] \models \phi \]
\[ s \models E[\phi U \psi] \iff \exists \pi \in \text{Path}(s), \exists i \geq 0. \pi[i] = \phi \land \forall 0 \leq j < i. \pi[j] \models \phi \]

Figure 3 presents a model checking algorithm, SAT, for CTL. Given a formula and a model, SAT returns the set of states at which the formula is true in the model. The algorithm is based on that presented in [12], but we have reorganized it so that it can be easily extended to the logics CTL-V and CTL-VW that we consider subsequently. In particular, the algorithm accepts predicates parameterized by a list of variables \( \pi \), which in the case of CTL will always be an empty list, and it includes a line for formulas \( \exists x.\phi \), which are not part of CTL. In the latter case, the instantiation of the model checking algorithm for CTL simply fails.

In this algorithm, the SAT function simply serves to recursively traverse the formula, except in the case of a proposition, where it returns each state where the proposition is satisfied (recall that \( \pi \) is empty for CTL). The functions applied to the intermediate SAT results then perform the main calculation of the algorithm. The function \( \text{conjunction} \) combines compatible information from two sets of results, and thus implements conjunction. In the case of CTL, this amounts to intersection. The function \( \text{disj} \) collects all information from two sets of results, and thus implements disjunction. The functions \( \text{Conj} \) and \( \text{Disj} \) fold \( \text{conj} \) and \( \text{disj} \) over a set of results, respectively, and are used to define some of the other functions. The function \( \text{neg} \) computes the complement of a set of results, and in the case of CTL amounts to returning the
The functions difference between the complete set of states and the given
stantiation for CTL

determines whether two results are compatible, which simply
means that they are equal in the case of CTL. The function negone
computes the complement of a single result. The
function shift returns all results in a set \( T \) that are associated
with a state \( s_1 \), but replaces the state information in each
case by another state \( s_2 \); this function makes it convenient to
implement \( \text{pre}_y \) and \( \text{pre}_z \). Finally, the function same makes
it possible to vary the termination condition for the fixpoint
iterations in \( \text{SAT}_{\text{AU}} \) and \( \text{SAT}_{\text{EU}} \).

It is straightforward to show that this algorithm is sound
and complete with respect to the semantics [12]:

**Theorem 1** (Soundness and completeness).
\[ s \models \phi \iff s \in \text{SAT}(\phi) \]

### 3.2 Computational Tree Logic with free variables

While CTL has proven to be very useful in the context of
program analysis [22], Lacey et al. showed that the extension
CTL-FV, supporting not propositions but predicates over free
variables, could be used in the context of program transformation [17]. In particular, CTL-FV has been used to formalize and prove correct a number of classical compiler optimizations [17, 21]. The introduction of variables implies that logic satisfies our first requirement for Coccinelle (Section 2.2).

The syntax of CTL-FV is essentially the same as that of
CTL with the addition of predicates over free variables. We
refer to these variables as **metavariables**:

\[
\phi ::= p(\pi) \mid \phi \land \phi \mid \phi \lor \phi \mid \neg \phi \\
\land \mid \lor \mid \neg \mid \text{AX} \phi \mid \text{EX} \phi \mid \text{A} \phi \mid \text{E} \phi
\]

fv(\phi) is the set of free metavariables of a formula \( \phi \).

The semantics of CTL-FV is almost the same as that of
CTL; indeed, they are based on the same models. The new features are that the semantics of predicates now takes free variables into account and all judgements now carry an
*environment* \( \theta \), mapping metavariables to an arbitrary set of
values, Val:

\[
\text{Env} = \text{MetaVar} \rightarrow \text{Val}
\]

When using CTL-FV to describe program transformations,
Val is the set of subterms of the program to be transformed.
The semantics is defined as follows. Again we elide the model:

\[
\begin{align*}
    s \models p(\pi) & \iff p(\theta(\pi)) \in \text{Label}(s) \\
    s \models \phi_1 \land \phi_2 & \iff s \models \phi_1 \land s \models \phi_2 \\
    s \models \phi_1 \lor \phi_2 & \iff s \models \phi_1 \lor s \models \phi_2 \\
    s \models \neg \phi & \iff s \not\models \phi \\
    s \models \text{AX} \phi & \iff \forall s' \in \text{next}(s), s' \models \phi \\
    s \models \text{EX} \phi & \iff \exists s' \in \text{next}(s), s' \models \phi \\
    s \models \text{A} \phi \mid \text{E} \phi & \iff \forall \pi \in \text{Path}(s), \exists i \geq 0, \pi[i] \models \phi \\
    s \models \text{A} \phi \mid \text{E} \phi & \iff \forall \pi \in \text{Path}(s), \exists i \geq 0, \pi[i] \models \phi
\end{align*}
\]

Model checking for CTL-FV has been done using a standard
model checker by instantiating the formula with respect to all
possible bindings of the metavariables [21].
3.3 Computational Tree Logic with quantified variables

Lacey further extends CTL-FV with the ability to existentially quantify over metavariables [15], producing a logic that we refer to as CTL-V. The syntax of CTL-V is as follows:

\[
\phi::= p(x) | \phi \land \phi | \phi \lor \phi | \neg \phi | \exists x.\phi \\
| AX\phi | EX\phi | A[\phi U \psi] | E[\phi U \psi]
\]

The semantics is the same as that of CTL-V, augmented with:

\[
s \models \exists x.\phi \iff \exists v \in \text{Val}, s \models [\theta[x \mapsto v]]\phi
\]

Introducing the ability to quantify over metavariables allows metavariables to have different values within different control-flow paths (our second requirement for Coccinelle) and thus adds flexibility as compared to CTL-FV. For example, consider the formulas in Figure 4, of which the first is in control-flow paths (our second requirement for Coccinelle).

The added bottom element is used to represent an environment: \(\bot\), producing a logic that we refer as CTL-V. The syntax of CTL-V is as follows:

\[
f(x) \land AX(g(y) \land AX(h(x, y))) \\
f(x) \land AX(\exists y.(g(y) \land AX(h(x, y))))
\]

**Figure 4.** CTL-FV and CTL-V formulas

4. A Model Checking Algorithm for CTL-V

In this section, we present a bottom-up model checking algorithm for CTL-V that is based on the CTL model checking algorithm presented in Figure 3. This algorithm is not sufficient as a foundation for Coccinelle, as it does not record information about the bindings of existentially quantified variables. Nevertheless, it permits to introduce a representation of environments that is common to model checking for both CTL-V and CTL-VW, and the algorithm may be of independent interest in context of the work of Lacey and De Moor [16]. While the CTL model checking algorithm identifies the set of states where a formula is satisfied, the CTL-V model checking algorithm must identify the set of pairs of a state and an environment that satisfy the formula. The algorithm is applicable to both CTL-V and CTL-FV, and even to CTL, although the environment it collects is unnecessary in the latter case.

We first present a representation of environments that allows the CTL-V model checking algorithm to efficiently represent information about bindings, then present the CTL-V model checking algorithm, and finally consider its soundness and completeness with respect to the semantics of Section 3.3.

4.1 Environments for CTL-V model checking

A recurring problem in model checking is how to represent the result of negation; the size of the complement of a set of states depends on the size of the model, which can be very large. For CTL-V, where the result of model checking is a set of state-environment pairs, taking the complement of the result includes taking the complement of the environment, where the size of the result depends on the size of Val. In the case of program matching, Val amounts to the set of subterms of the program. In our experience, the set of bindings that derive from matching a predicate against the source code is quite small, and thus taking the complement in this manner would incur a substantial performance overhead. To address this issue, we use constructive negation [6].

Rather than negating a binding by creating a disjunction of all of the possible bindings for the variable other than the current one, we add the ability to represent negative bindings directly. Thus, we define an extended form of environment, \(\text{Env}^{\pm}\), in which a variable is mapped to either a positive value in \(\text{Val}\) or a set of negative bindings (in \(\mathcal{P}(\text{Val})\)):

\[\text{Env}^{\pm} = (\text{MetaVar} \rightarrow \text{Val} + \mathcal{P}(\text{Val}))\]

The added bottom element is used to represent an environment that contains conflicting information. The domain of \(\bot\) is undefined. The following notation facilitates reasoning about the positive and negative bindings of an environment:

**Definition 2.** Let \(\theta \in \text{Env}^{\pm}\). We define

1. \(\text{dom}^+(\theta) = \{x \in \text{dom}(\theta) \mid \theta(x) \in \text{Val}\}\)
2. \(\text{dom}^-(\theta) = \{x \in \text{dom}(\theta) \mid \theta(x) \in \mathcal{P}(\text{Val})\}\)
3. \(\text{dom}^+(x) = \theta(x)\) iff \(x \in \text{dom}^+(\theta)\)
4. \(\text{dom}^-(x) = \theta(x)\) iff \(x \in \text{dom}^-(\theta)\)

\(\text{Env}^{\pm} = \{\theta \in \text{Env}^{\pm} \mid \text{dom}(\theta) \subseteq \text{dom}^+(\theta)\}\) is the set of the environments that have only positive bindings. \(\text{Env}^{\pm}\) is the same as \(\text{Env}\), and thus is the form of environment accepted by the semantics. \(\text{Env}^+_x = \{\theta \in \text{Env}^{\pm} \mid \text{dom}(\theta) = \text{dom}^+(\theta) \cap \text{fv}(\phi)\}\) is the subset of \(\text{Env}^{\pm}\) restricted to the free variables of the formula \(\phi\).

**Notation 1.** For convenience, we define the following explicit notation for environments. Let \(\{x_1, \ldots, x_n, y_1, \ldots, y_m\} \subseteq \mathcal{P}(\text{Val})\)
MetaVar be a set of pairwise different meta-variables. Then $[x_1, x_2, \ldots, v_1, \ldots, x_n \mapsto v_n, y_1 \not= V_1, \ldots, y_n \not= V_n]$ represents the environment $\theta$ such that $\theta(x_i) = v_i \in \text{Val}$ for every $i \in 1..n$ and $\theta(y_j) = V_j \in \mathcal{P}(\text{Val})$ for every $j \in 1..m$.

**Definition 3** (Conflict). Two environments $\theta_1, \theta_2 \in \text{Env}^\pm$, with $\theta_1 \neq \bot, \theta_2 \neq \bot$, conflict iff $\exists x \in \text{dom}(\theta_1) \cap \text{dom}(\theta_2)$

$$\theta_1^+(x) \neq \theta_2^+(x) \lor \theta_1^-(x) \in \theta_2^-(x) \lor \theta_2^+(x) \in \theta_1^-(x)$$

Otherwise, $\theta_1$ and $\theta_2$ are said to be compatible.

Next, we introduce an ordering on environments. This ordering is needed to relate the environments that are returned by the algorithm, which can contain negative bindings, to the ones that satisfy the formula according to the semantics. Intuitively an environment with more specific information is less than one with less specific information, e.g., $[x = 42, y = 12] \subseteq [x = 42] \subseteq [x \neq 87]$.

**Definition 4** (Environment ordering). Let $\theta_1, \theta_2 \in \text{Env}^\pm$ and define $\theta_1 \preceq \theta_2$ iff $\theta_1 = \bot \lor \text{dom}(\theta_1) \subseteq \text{dom}(\theta_2) \land \forall x \in \text{dom}(\theta_1)\theta_1^+(x) = \theta_2^+(x) \lor \theta_1^-(x) \notin \theta_2^-(x) \lor \theta_2^+(x) \in \theta_1^-(x)$

According to this ordering, the environment with the empty set as its domain is the greatest element $\top$ and $\bot$ is the least element.

To compute the environment that satisfies a conjunction in a given state, the algorithm joins the environments that satisfy each of the conjunct formulas in the same state. The join of two environments is defined as follows:

**Definition 5** (Environment join). Let $\theta_1, \theta_2 \in \text{Env}^\pm$. We define $\theta_1 \sqcap \theta_2$ to be $\bot$ if $\theta_1 = \bot$ or $\theta_2 = \bot$, i.e., $\bot \sqcap \bot = \bot \land \bot = \bot$. For $\theta_1 \neq \bot, \theta_2 \neq \bot$, if $\theta_1$ and $\theta_2$ are compatible, $\theta_1 \sqcap \theta_2$ is defined as follows:

$$(\theta_1 \sqcap \theta_2)(x) = \begin{cases} v & \text{if } \theta_1^+(x) = v \lor \theta_2^+(x) = v \\ V_1 & \text{if } \theta_1^+(x) = \bot \land V_1 \land x \notin \text{dom}(\theta_2) \\ V_2 & \text{if } \theta_2^+(x) = V_1 \land x \notin \text{dom}(\theta_1) \\ V_1 \sqcap V_2 & \text{if } \theta_1^-(x) = V_1 \land x \notin \text{dom}(\theta_1) \\ V_1 \sqcup V_2 & \text{if } \theta_2^-(x) = V_1 \land \theta_2^-(x) = V_2 \end{cases}$$

If $\theta_1$ and $\theta_2$ conflict, then $\theta_1 \sqcap \theta_2 = \bot$.

It is straightforward to show that the join operator computes the greatest lower bound of two environments according to the environment ordering.

Finally, we conclude with the definition of the negation of an environment. This operation can create negative bindings.

**Definition 6** (Negation of an environment). For $\theta \in \text{Env}^\pm$, we define

$$-\theta = \begin{cases} \{\top\} & \text{if } \theta = \bot \\ \{x \not\in \{v\} | v = \theta^+(x)\} \cup \{x \mapsto v | v \in \theta^-(x)\} & \text{otherwise} \end{cases}$$

Notice that $-\theta \in \mathcal{P}(\text{Env}^\pm)$.

### 4.2 A bottom-up CTL-V model checking algorithm

The algorithm SAT for CTL-V takes as arguments a model and a formula and returns result in the form of a set of pairs $(s, \theta) \in \text{State} \times \text{Env}^\pm$. As compared to the definition of SAT in Figure 3, it is only necessary to redefine the element-level functions $\text{inj, negone, existsone, shift, and same}$ to take environments into account. The new definitions are as follows:

$$\text{inj}(s, \theta) = (s, \theta)$$

$$(s_1, \theta_1) \sqcap (s_2, \theta_2), \text{ if } s_1 = s_2 \land \theta_1 \sqcap \theta_2 = \bot$$

$$\text{negone}(s, \theta) = \{ (s', \theta) | s' \in \text{States} \setminus \{s\} \cup \{ (s, \theta') | \theta' \in -\theta \}$$

$$\text{existsone}(x, s, \theta) = (s, \theta - [x \mapsto v]), \text{ if } \theta^+(x) = v \land \theta(x) = V \land \theta(x) \text{, otherwise}$$

$$\text{shift}(s_1, T, s_2) = \{ (s_2, \theta) | (s_1, \theta) \in T \}$$

$$\text{same}(T_1, T_2) = T_1 = T_2$$

The function $\text{inj}$, used to inject the result of matching a predicate into the codomain of SAT, now keeps both the state and the environment argument, as the environment argument is now non-trivial. The function $\sqcap$, used in computing a conjunction of results, extends the CTL definition by taking the join of the associated environments. The function $\text{negone}$, used in computing the negation of a result, similarly extends the definition for CTL by including not only each element of the complement of the given state, paired with the empty environment, but also pairs for the current state combined with each element of the negation of the current environment. These pairs represent the least specific ones that are incompatible with the given state and environment pair. The function $\text{shift}$ selects the pairs in $T$ that have the same state as its first argument and replaces the state in each pair with the state in its third argument, as is needed to implement the operators $\text{AX}$ and $\text{EX}$. Finally, the function $\text{same}$ simply checks whether the results it is given represent equivalent sets, as in the CTL definition.

### 4.3 Examples

As examples of the CTL-V model checking process, we consider again the models of Figure 2 and the formulas of Figure 4. In checking the formula $f(x) \land AX(g(y) \land AX(h(x, y)))$ with respect the model in Figure 2b, the result for the subformula $g(y) \land AX(h(x, y))$ is

$$(\{2, [x \mapsto 1, y \mapsto 2]\}, \{4, [x \mapsto 1, y \mapsto 3]\})$$

Model checking of the enclosing $\text{AX}$, however fails, because although both of the neighbors of state 1 are in the above result, they are associated with conflicting environments. On the other hand, in checking the formula $f(x) \land AX(\exists y . (g(y) \land AX(h(x, y))))$, with respect to the same model, the same result is obtained for $g(y) \land AX(h(x, y))$, but for $\exists y . (g(y) \land$
AX(h(x, y))), the result is
\[(\{2, [x \mapsto 1]\}, \{4, [x \mapsto 1]\})\]
in which the conflicting bindings of y have been dropped. This time, the result of the enclosing AX is \(\{(1, [x \mapsto 1])\}\), which in turn is the result of checking the entire formula.

### 4.4 Soundness and completeness

The model checking algorithm for CTL-V can return environments that contain negative bindings, while the semantics for CTL-V only accepts environments with positive bindings. Thus, the relationship between them is not as direct as in the case of CTL (Theorem 1). Instead, the soundness and completeness theorem for the CTL-V model checking algorithm relates an environment produced by the algorithm, which may contain negative bindings, to the set of more specific environments that contain only positive bindings and are accepted by the semantics:

**THEOREM 2** (Soundness and completeness). Let \(\phi\) be a formula, and \(T \in \mathcal{P}(\text{State} \times \text{Env}_\phi^+).
Let us define \(\eta_\phi(T) = \{(s, \theta) \in \text{State} \times \text{Env}_\phi^+ \mid \exists (s', \theta') \in T. s = s' \land \theta \subseteq \theta'\}.\)
Then, \(\forall s \forall \theta \ (\theta \in \text{Env}_\phi^+ \land s \models_\theta \phi) \iff (s, \theta) \in \eta_\phi(\text{SAT}(\phi))\).

The proof of this theorem has been validated with Coq [3].

### 5. CTL-VW

CTL-V allows variables to have different values in different control-flow paths, but discards their bindings, making it impossible to refer to such variables in a subsequent transformation phase. The goal of CTL-VW is to collect the bindings of such variables, as witnesses, in a way that does not affect the rest of the matching process. This satisfies the third of our requirements for Coccinelle (Section 2.2). We first define witnesses, then give the semantics and model checking algorithm of CTL-VW, and finally sketch the proof of the soundness and completeness of the model checking algorithm.

#### 5.1 Witnesses

A witness is essentially a record of a state, a binding, and the set of witnesses for other bindings that contributed to establishing the binding in the given state. A witness thus has a tree structure, which corresponds to the structure of the nested existential quantifiers in the formula.

**DEFINITION 7** (Witnesses). The set \(\text{Wit}\) of witnesses is defined as:

\[
\text{Wit} = \text{States} \times \text{MetaVar} \times (\text{Val} + \mathcal{P}(\text{Val})) \times \text{WitForest}
\]

A witness forest \(\Omega \in \text{WitForest}\) is a multiset of witnesses, i.e., a pair \((W, f)\), where \(W \subseteq \text{Wit}\) is a set of witnesses, and \(f : W \rightarrow \mathbb{N}^+\) is a function that associates each witness with its multiplicity (a nonnegative integer). We define \(\text{WitForest}^+\) as the set of the witness forests in which all bindings are positive.

The operator \(\sqcup\) produces the join of two multisets. For ease of reading, we sometimes use the set-like notation \(\{\ldots\}\) to enumerate the elements of a multiset. Then, \(\{a, a, b\}\) represents the multiset whose underlying set is \(\{a, b\}\), and the multiplicities of a and b are 2 and 1, respectively.

We use multisets rather than sets for technical reasons. However, in order to specify a termination criterion in the model checking algorithm, we need to reason about the underlying sets of witness forests. We define the binary operator \(\simeq\) as follows:

\[\Omega_1 \simeq \Omega_2 \iff \text{wit}_\text{red}(\Omega_1) = \text{wit}_\text{red}(\Omega_2)\]

\(\text{wit}_\text{red}(\Omega)\) translates \(\Omega\) into the corresponding underlying set.

As for environments, we define an ordering on witnesses, and on witness forests. These orderings are useful to relate the results of the algorithm, which can contain negative bindings, to witness forests that satisfy the formula according to the semantics.

**DEFINITION 8** (Witness ordering). Let \(w = (s, x, a, \Omega)\) and \(w' = (s', x', a', \Omega')\) be two witnesses. We define \(w \sqsubseteq w'\) iff \(s = s' \land x = x' \land [x \mapsto a] \subseteq [x' \mapsto a'] \land \Omega \subseteq \Omega'\).

The ordering for witness forests is defined as follows:

1. \(h : \{(w, i) \mid w \in W \land i \in 1..f(w)\} \rightarrow \{(w', j) \mid w' \in W' \land j \in 1..f'(w')\}\)
2. \(\forall w \in W. \forall i \in 1..f(w). w \sqsubseteq w', \text{ where there is some } j \text{ such that } (w', j) = h(w, i)\)

#### 5.2 A semantics for CTL-VW

The new feature of the semantics of CTL-VW is to define what it means to be a witness forest for a formula \(\phi\), with respect to some state and environment. If \(\phi = \exists x. \psi\), then a witness forest for \(\phi\) should record a binding of \(x\) that satisfies \(\psi\), as well as a witness forest that records information about the bindings associated with any existential quantifiers in \(\psi\). If \(\phi = \phi_1 \lor \phi_2\), then a witness forest for \(\phi\) has to be a witness forest for either \(\phi_1\) or \(\phi_2\), because the information needed to satisfy \(\phi_1\) or \(\phi_2\) also satisfies \(\phi_1 \lor \phi_2\).

A witness forest for \(\phi = \phi_1 \land \phi_2\) has to contain the information that makes \(\phi_1\) true and the information that makes \(\phi_2\) true. Thus, the semantics joins a witness forest for \(\phi_1\) and a witness forest for \(\phi_2\) to obtain a witness forest for \(\phi_1 \land \phi_2\). Similarly, for \(\forall x. \psi\), the semantics collects, for each successor of the current state, a witness forest that makes \(\psi\) true in this successor. In both cases, we collect the bindings of the metavariabes that make the subformulas true at the relevant states. On the other hand, since \(\text{EX}\) is related to a...
disjunction, a witness forest for $\text{EX}\psi$ is a witness forest for $\psi$ in one of the successors of the current state.

If $\phi = A[\phi_1 U \phi_2]$, then we can define a witness forest for $\phi$ by analogy with conjunction and $\text{AX}$. Then, the usual semantics of $\text{AU}$ implies that a witness forest of $\phi$ collects as many witnesses as there are paths starting from the current state. This leads to an infinite number of witness forests, carrying redundant information. In order to collect only witness forests that are pertinent, we reason about a finite unfolding of the model when defining the semantics of $\text{AU}$, rather than reasoning about all the paths. Then, a witness forest of $A[\phi_1 U \phi_2]$ collects a witness forest of $\phi_1$ for every state in the finite unfolding that is not a leaf, and a witness forest of $\phi_2$ for every leaf. A witness forest for $\phi = E[\phi_1 U \phi_2]$ collects, for some path leading to a leaf, a witness forest for $\phi_1$ in every non-leaf state in this path, and a witness forest for $\phi_2$ in the leaf.

**Definition 9** (Finite paths). Given a model and a state $s$, we define $\text{FPPath}(s)$ as the set of the finite paths starting from $s$. Given a finite path $\pi \in \text{FPPath}(s)$, $|\pi|$ is the length of $\pi$, i.e., $|\pi| + 1$ is the number of states in $\pi$. $\text{prefix}(\pi)$ represents the set of the (finite) proper prefixes of $\pi$, and $\pi_{\leq n}$, for $n \leq |\pi|$, represents the prefix of $\pi$ containing the first $n + 1$ states (from $\pi[0]$ to $\pi[n]$).

**Definition 10** (Finite unfolding). Given a state $s$, a finite unfolding from $s$ is a set $\Sigma \subseteq \text{FPPath}(s)$ where every path $\pi \in \Sigma$ satisfies the following constraints:

- $\forall \pi' \in \text{prefix}(\pi), \pi' \notin \Sigma$ (no redundancy)
- $\forall i \in 0..|\pi| - 1. \forall s' \in \text{next}(\pi[i]). \exists \pi'' \in \Sigma. \pi''_{\leq i} = \pi_{\leq i}$ and $s'' = \pi''[i + 1]$ (full branching)

The set of the finite unfoldings starting from $s$ is denoted $\Pi(s)$. The set of the states that appear in some path of $\Sigma$ is denoted $\Sigma$.

Consider the model shown in Figure 5. Then $\{s_0\}$ and $\{s_0 \to s_1, s_0 \to s_2 \to s_2\}$ are two valid finite unfoldings from $s_0$. On the other hand, $\{s_0 \to s_1, s_0 \to s_2, s_0 \to s_2 \to s_2\}$ violates the first constraint of Definition 10, and $\{s_0 \to s_1\}$ does not explore the right branch, and thus does not satisfy the second constraint.

**Definition 11** (Semantics). Given a model $(\text{States}, \to, \text{Label})$, a state $s \in \text{States}$, a formula $\phi$, an environment $\theta \in \text{Env}^+$, and a witness forest $\Omega \in \text{WitForest}^+$, the semantics is defined as the following relation:

$$
s \models_{\theta, \phi} p(\pi) \iff p(\theta(\pi)) \in \text{Label}(s)
$$

$$
s \models_{\theta, \Omega_1, \Omega_2} \phi_1 \land \phi_2 \iff s \models_{\theta, \Omega_1} \phi_1 \land s \models_{\theta, \Omega_2} \phi_2
$$

$$
s \models_{\theta, \Omega_1, \Omega_2} \phi_1 \lor \phi_2 \iff s \models_{\theta, \Omega_1} \phi_1 \lor s \models_{\theta, \Omega_2} \phi_2
$$

$$
s \models_{\theta, \Omega_1, \Omega_2} \lnot \phi \iff \forall \Omega. \Omega \notin \theta_{\phi}
$$

$$
s \models_{\theta, \Omega_1, \Omega_2} \exists x. \phi \iff s \models_{\theta, \Omega_1, \Omega_2} \exists x. \phi
$$

$$
s \models_{\theta, \Omega, \text{AX}} \phi \iff \exists \Omega_{\pi} | s' \in \text{next}(s). s' \models_{\theta, \Omega_{\pi}} \phi
$$

$$
s \models_{\theta, \Omega_1, \Omega_2} A[\phi_1 U \phi_2] \iff \exists \Omega \in \Pi(s). \exists \Omega_{s'} | s' \models_{\theta, \Omega} \phi_1 \land \forall \Omega_{s''} | s'' \models_{\theta, \Omega} \phi_2 \land \forall 0 \leq j < |\pi|, \pi[j] \models_{\theta, \Omega} \phi_1
$$

$$
s \models_{\theta, \Omega} \text{EF} \phi \iff \exists \Omega_{\pi} | s' \in \text{next}(s). s' \models_{\theta, \Omega} \phi
$$

$$
s \models_{\theta, \Omega} \text{EX} \phi \iff \exists s' \in \text{next}(s). s' \models_{\theta, \Omega} \phi
$$

Note that the semantics does not allow any witness to be satisfied under a negation. This facilitates the proof of the soundness and completeness of the model checking algorithm. Indeed, a negation turns the nature of an existential quantification to a universal quantification, for which witnesses would not make any sense. However, one could justify the interest of a witness for an existential quantifier that appears under an even number of negations. We do not consider a special treatment for this case. Because of this constraint on negation, we must include both $\text{AX}$ and $\text{EX}$ in the logic explicitly, because a definition of $\text{e.g.}$, $\text{EX}$ in terms of $\text{AX}$ and negation, would preclude witnesses under $\text{EX}$.

### 5.3 A model checking algorithm for CTL-VW

The algorithm SAT for CTL-VW takes as arguments a model and a formula and returns a set of triples of the form $(s, \theta, \Omega) \in \text{States} \times \text{Env}^+ \times \text{WitForest}$. The new definitions of the element-level functions of Figure 3 are as follows:

$$
inj(s, \theta) = (s, \theta, \emptyset)
$$

$$
(s_1, \theta_1, \Omega_1) \overset{\text{negone}}{\rightarrow} (s_1, \theta_1 \land \theta_2 \land \Omega_1 \land \Omega_2), \text{ if } s_1 = s_2 \land \theta_1 \land \theta_2 \neq \bot
$$

$$
\text{negone}(s, \theta, \Omega) = \{(s', \theta', \emptyset) | s' \in \text{States} - \{s\} \cup \{(s', \theta', \emptyset) | \theta' \in \theta\}
$$

$$
\text{existstone}(x, (s, \theta, \Omega)) = \{(s, \theta - [x \mapsto v], \{(s, x, v, \Omega)\}, \text{if } \theta^+(x) = v
$$

$$
\text{same}(T_1, T_2) = \text{triple_red}(T_1) = \text{triple_red}(T_2)
$$

where $\text{triple_red}(T) = \{(s, \theta, \text{witness}(\Omega)) | (s, \theta, \Omega) \in T\}$

These definitions augment the corresponding definitions for CTL-V to maintain the witnesses. As there are no existential quantifiers under a predicate, the function inj just uses an empty witness forest. The function $\overset{\text{negone}}{\rightarrow}$ takes the join of the provided witness forests, if the provided states and environments are compatible. The function existstone considers the same
cases as for CTL-V, but rather than discarding the binding of the quantified variable converts it to a witness, encapsulating the current witness forest. The function shift is analogous to the CTL-V definition. Finally, the function same, used to identify a fixed point, first converts all witness forests in its arguments to their underlying sets, using wit_red, and then compares the obtained results.

It remains to consider the definition of negone. This function negates the state component, and then the environment component, as in the CTL-V definition, but completely discards the witness component. This strategy corresponds to the semantics of negation presented in Section 5.2, where the subformula should not be satisfiable at the given state and environment for any witness forest.

5.4 Examples
As in Section 4.3, we consider the models of Figure 2 and the formulas of Figure 4. In checking the formula

\[ f(x) \land AX(g(y) \land AX(h(x, y))) \]

with respect to the model in Figure 2b, the result for the subformula \( g(y) \land AX(h(x, y)) \) is

\[ \{(2, [x \mapsto 1, y \mapsto 2], \emptyset), (4, [x \mapsto 1, y \mapsto 3], \emptyset)\} \]

Model checking for the enclosing \( AX \) fails as before, because the environments are incompatible. Similarly, in checking the formula \( f(x) \land AX(\exists y. (g(y) \land AX(h(x, y)))) \), with respect to the same model, the same result is obtained for \( g(y) \land AX(h(x, y)) \), but now for \( \exists y. (g(y) \land AX(h(x, y))) \), the result is

\[ \{(2, [x \mapsto 1], \{(2, y, 2, \emptyset)\}), (4, [x \mapsto 1], \{(4, y, 3, \emptyset)\})\} \]

in which the information about the various bindings of \( y \) is still available. The result of the enclosing \( AX \) is then

\[ \{(1, [x \mapsto 1], \{(2, y, 2, \emptyset), (4, y, 3, \emptyset)\})\} \]

which is in turn the result for the entire formula. This result includes the possible bindings of \( y \) and the states at which those bindings are applicable.

5.5 Soundness and completeness
As for CTL-V, the CTL-VW model checking algorithm may return an environment with negative bindings. We thus extend the \( \eta \) function of Section 4.4 to take witnesses into account.

Let \( \phi \) be a formula, and \( T \in P(States \times Env^\pm \times WitForest) \). Let us define \( \eta_\phi(T) = \{(s, \theta, \Omega) \in States \times Env^\pm \times WitForest^+ \mid \exists (s', \theta', \Omega') \in T. s = s' \land \theta \subseteq \theta' \land \Omega \subseteq \Omega'\} \).

THEOREM 3 (Soundness). \( \forall s, \theta, \Omega. \)

\[ \theta(s, \theta, \Omega) \in \eta_\phi(SAT(\phi)) \implies s \models_{\theta, \Omega} \phi \]

Proof sketch The proof is by structural induction on \( \phi \). We present the case \( \phi = A[\phi_1 U \phi_2] \). We first define a bounded semantics for AU, and a bounded SAT algorithm for AU. \( s \models_{\theta, \Omega} A[\phi_1 U \phi_2] \iff \exists N \text{ unfolding } \phi \text{ whose maximum length is less than or equal to } N \), and such that the semantic definition of AU is satisfied with \( \Sigma. SAT_{AU}(\phi_1, \phi_2) \in N \) corresponds to \( N \) iterations of the algorithm described in SAT_{AU}(\phi_1, \phi_2). The following lemma relates the bounded semantics to the bounded algorithm.

LEMMA 1. Let \( \phi = A[\phi_1 U \phi_2] \). Then, \( \forall s, \theta, \Omega, N \)

\( (s, \theta, \Omega) \in \eta_\phi(SAT_{AU}(\phi_1, \phi_2) \in N) \implies s \models_{\theta, \Omega} A[\phi_1 U \phi_2] \in N \)

We omit the proof of this lemma, also done by induction.

Let \( (s, \theta, \Omega) \in \eta_\phi(SAT_{AU}(\phi_1, \phi_2)) \). Then, \( \exists N. (s, \theta, \Omega) \in \eta_\phi(SAT_{AU}(\phi_1, \phi_2) \in N) \). By Lemma 1, \( s \models_{\theta, \Omega} A[\phi_1 U \phi_2] \in N \), so \( s \models_{\theta, \Omega} A[\phi_1 U \phi_2] \).

The completeness theorem cannot be stated as directly as the soundness theorem. Indeed, if \( \phi \) contains AU (resp. EU), a witness forest accepted by the semantics may contain more witnesses than are generated by SAT_{AU} (resp. SAT_{EU}) in reaching a fixed point. In this case, the witness forest is not returned by the algorithm in its exact form, but the algorithm produces a shorter and equivalent (in the sense of \( \asymp \)) witness forest.

THEOREM 4 (Completeness). \( \forall s, \theta, \Omega. \)

\[ \theta \in Env^\pm s \models_{\theta, \Omega} \phi \implies \exists \Omega'. \Omega \asymp \Omega'. (s, \theta, \Omega') \in \eta_\phi(SAT(\phi)) \]

Proof sketch The proof is done by induction. In the AU case, we use a lemma analogous to Lemma 1.

To conclude, we consider the relationship between CTL-VW and CTL-V. CTL-VW is a conservative extension of CTL-V: the collection of witnesses does not have any impact on the satisfiability of a formula.

PROPERTY 1 \( \forall s, \theta, s \models_{\theta} \phi \iff \exists \Omega \in \text{WitForest}^+, s \models_{\theta, \Omega} \phi \) 

6. Applying CTL-VW in the context of Coccinelle
In this section, we define a core SmPL language, that is sufficient to treat the example presented in Section 2.1 (Figure 1), present its translation into CTL-VW, describe the matching and transformation process, and then present a few benchmarks. Many more examples of the use of SmPL and the associated performance are available in our previous work [18, 20]. The section concludes with a brief example illustrating the benefit of being able to mix existential and universal path quantification, as provided by temporal logics based on CTL.

For the purposes of this presentation, we have slightly simplified the semantics of SmPL, in that our encoding in CTL-VW allows dots ("...") to match any path, while
in SmPL dots represent the shortest path between terms matching the preceding and following patterns. This shortest path constraint can be encoded straightforwardly in CTL-VW. Furthermore, it is possible to drop the shortest path constraint in full SmPL by annotating the dots with \texttt{when} \ any.

### 6.1 Syntax of a simplified SmPL

The syntax of the transformation specification part of a semantic patch is quite complex, because \texttt{and} and \texttt{+} annotations can be freely mixed. In a transformation specification, the combination of the unannotated code and the \texttt{+} code represents the pattern to match against, and the combination of the unannotated code and the \texttt{and} code represents the code to generate. Each must have the structure of valid C code. Expressing these constraints is, however, unrelated to the use of CTL-VW. Instead, we define the grammar of patterns to match against, and assume that the elements of these patterns are implicitly annotated with the transformation to perform. For example, in the \texttt{type} \texttt{ref} semantic patch of Figure 1, the pattern \texttt{return } \neg C; \ would be annotated to indicate that \texttt{of_node_put}(n); should be inserted before it.

In our simplified SmPL, the transformation specification part of a semantic patch has the form of a sequence \( S \), as defined by the following grammar:

\[
\begin{align*}
S & \in \text{Sequences} ::= \ E \ S | \ \varepsilon \\
E & \in \text{Elements} ::= T | D | (S | S) \\
T & \in \text{Terms} ::= \text{Atomic} | \text{if} (exp) \ T \\
D & \in \text{Dots} ::= \ldots | D \ \text{when} \ ! = S
\end{align*}
\]

A Sequence should not contain consecutive dots (\( D \)), but we do not complicate the grammar with this constraint. \texttt{Atomic} is an arbitrary atomic term, such as an assignment or function call, and \texttt{exp} is an arbitrary expression. \texttt{Atomic} and \texttt{exp} may contain metavariables. We consider conditionals with only one branch, because that is all that is required for our example. The treatment of a conditional with two branches is similar.

### 6.2 Quantification of metavariables

The first step in the translation of SmPL to CTL-VW is to introduce existential quantifiers to delimit the scope of each metavariable. For this, we extend the syntax of Sequences, Elements, and Terms to include existential quantifiers, as shown below:

\[
\begin{align*}
S & \in \text{Sequences} ::= \ E \ S | \ \varepsilon | \ \exists x.S \\
E & \in \text{Elements} ::= T | D | (S | S) | \ \exists x.E \\
T & \in \text{Terms} ::= \text{Atomic} | \text{if} (exp) \ T | \ \exists x.T \\
D & \in \text{Dots} ::= \ldots | D \ \text{when} \ ! = S
\end{align*}
\]

The scope of a metavariable is the smallest Sequence, Element, or Term that contains all references to it, including references in \texttt{+} code. Inserting existential quantifiers according to this strategy is straightforward. For the semantic patch in Figure 1, the scope of the metavariable \( n \) extends around the entire semantic patch, but the scope of the other metavariables is only the immediate containing Term. In particular, the scope of the metavariable \( C \) in the modified term \texttt{return } \neg C; \ is local to the pattern itself, i.e., \( \exists C. \texttt{return } \neg C; \) allowing it to match a return of any negative constant within the different control-flow paths. This property is essential, as the matched code may need to return an error code for many reasons, returning a different value in each case (see Figure 8).

### 6.3 Translation to CTL-VW

Figure 6 defines the translation of a pattern into CTL-VW. The entry point of the translation is the function \( C_v \), which takes as arguments a Sequence and a formula describing how to match the remainder of the semantic patch. Initially, a transformation specification \( S \) is translated as \( C_v[S] \).\footnote{We define \( \text{true} \) as an abbreviation for \( p() \lor \neg p() \), for an arbitrary predicate \( p() \).}

The translation contains two sets of rules: 1) the \( C \) rules \( C_v, C_r, \) and \( C_t \), for translating Sequences, Elements, and Terms, respectively, at the top level, and 2) the \( G \) rules \( G_s, G_e, G_t \), for Sequences, Elements, Terms, and Dots, respectively, when these occur under a \texttt{when} clause. The connection between the two sets of rules is made by the rule \( G_d \), which is used in the translation of \texttt{. . . } and processes each of the associated \texttt{when} clauses. The difference between the two sets of rules is only in the use of the argument \( a \), describing the rest of the SmPL code. For the \( C \) rules, \( a \) represents a pattern at the
same level as the pattern being translated, which thus must be matched after the current pattern, while for the $G$ rules, $a$ represents the code that follows the associated dots, and thus only serves to delimit any dots that appear at the end of the when code. In particular, the rule $G_d$ for Terms does not have an argument $o$, because a Term cannot end in dots.

We examine in more detail the $C$ translation rules for dots, disjunctions, atomic patterns, and conditionals. These rules illustrate the main concepts of the translation process.

**Dots** Dots represent a sequence of arbitrary code, possibly constrained by when clauses, along a control-flow path. In the translation, the end of this sequence is indicated by the formula $a$ that describes the rest of the SmPL code. Such a delimited path can be expressed by an “until” path operator, AU or EU. Our simplified language only supports universal quantification over paths, and thus we use AU. $^5$ The left argument of AU is constructed using the rule $G_d$, which creates a formula checking that none of the patterns in the when clauses are matched within the path. The right argument of AU is the term $a$ matching the rest of the semantic patch.

**Disjunction** A disjunction $(S_1 \lor S_2)$ matches $S_1$ if possible, and otherwise $S_2$. The translation reflects the priority of $S_1$ over $S_2$ by encoding $S_2$ as $\neg(C_s[S_1]) \land C_s[S_2] \land a$. This translation duplicates the previous processing of $S_1$. In the implementation, we use instead a “sequential disjunction” operator, that uses the negation of the previously computed result of processing $S_1$, and thus eliminates this duplication. The translation of a disjunction propagates the formula $a$ separately into the translation of each branch. In this way, the rest of the semantic patch is matched at a point starting from the end of the code matching $S_1$ or $S_2$.

**Atomic** An atomic pattern may involve a transformation, as illustrated in line 13 of our type_ref semantic patch (Figure 1). Thus, the matching process must remember information about the position at which each atomic pattern matches. To collect this information, which is our fourth requirement for Coccinelle (Section 2.2), we use witnesses. Specifically, we introduce a new existentially quantified metavariable $_v$ and create a predicate “=" that simply matches this variable to a textual representation of the atomic pattern, including any annotations about the transformation required. For example, the pattern return ¬C; with the annotation that of_node_neo_put(n); should be inserted before it would be translated as follows:

```
return -C; \land \exists_v.v = \lt\ of\ node\_\ neo\_\ put(n);\ return\ -C;\ "
```

The witnesses for $v$ will contain the current state and the binding of $v$ to the textual representation of the atomic pattern, indicating where and how to perform the transformation.

---

$^5$ Coccinelle automatically converts AU to AW, which can be defined using EU and negation, when the source program is found to contain a loop. AW can accept a path that goes around a loop infinitely without satisfying $\phi_2$, but AW is implemented much less efficiently than AU.

---

**Figure 7.** CTL-VW translation of the semantic patch type_ref. The right argument of the equality constraint on $_v$ is elided in each case, for conciseness.

**Conditionals** The translation of a conditional is determined by the representation of a conditional in a Coccinelle control-flow graph. For conciseness, we omit further details. Note, however, that the translation of the header of a conditional introduces a metavariable $_v$, as in the translation of an atomic pattern, thus making it possible to find and transform the matched code.

Figure 7 shows the result of translating the semantic patch type_ref into CTL-VW. The formula uses the “sequential disjunction” operator, denoted as $|$, that was described above. The pattern $n \equiv \text{NULL}$ has been expanded into a variety of ways to make a NULL test, using Coccinelle’s isomorphisms [20]. For type_ref, the size of the formula is comparable to the size of the semantic patch. In general, however, subformulas can be duplicated 1) in the branches of a disjunction, 2) when a when clause ends in dots, and 3) in full SmPL to implement the shortest path constraint, although in the latter case, only the atomic terms immediately preceding and following the dots are duplicated. In practice, however, we have found that it is the complexity of the source code rather than the size of the formula that has an impact on the performance [20].

### 6.4 Matching and transformation

As illustrated by the CTL-VW formula in Figure 7, a formula resulting from the translation always has the following properties: 1) All metavariables are existentially quantified, 2) The existential quantifiers for the introduced metavariables $_v$ recording the positions of the atomic terms are always innermost, and 3) The translation of an atomic term, including the quantifier of its $_v$ variable, is nested within quantifiers for all of its free metavariables. The first point implies that in the result of matching the CTL-VW formula against the source code, the environment component of each triple is always empty. The second and third points imply that the witnesses are trees in which the leaves are the bindings of the $_v$ metavariables, representing the code to transform, and the path from the root of a witness to a given leaf contains the
Based on the result of the matching process, Coccinelle collects the set of sequences of bindings leading from the root of a witness to each of its leaves that contains a transformation. Coccinelle only accepts a set of transformations if the states involved are disjoint and are only reachable from the states associated with the set of triples. These conditions help avoid ambiguity in the transformation process. If the set of transformations is acceptable, code is generated accordingly, in our case adding a call to of_node_put above lines 7 and 11 of Figure 8. Otherwise, Coccinelle aborts.

6.5 Experiments

The type_ref semantic patch shown in Figure 1 transforms four files in the Linux kernel: arch/powerpc/platforms/series/nvram.c, drivers/macintosh/smu.c, drivers/macintosh/therm_pm72.c, and drivers/video/fsl-diu-fb.c.7 Figure 10 summarizes the sizes of these files and the performance of Coccinelle when applying the type_ref semantic patch to them. All experiments were performed on a 1.4GHz uniprocessor Centrino laptop with 635MB of RAM. Execution times are the average of five runs. The result of applying the semantic patch to nvram.c and smu.c has been validated into the Linux kernel. The result for fsl-diu-fb.c has been validated by a Linux developer. We observed that the result of applying the semantic patch to therm_pm72.c is probably a false positive, due to an interprocedural effect that the semantic patch does not take into account. Nevertheless, the code structure in this case is typical, and thus it also serves as a representative example.

The files range in size from 149 to over 2000 lines of C code, including comments and whitespace. In each case there is a single function that is relevant to the semantic patch, and this function ranges in size from 27 to 94 lines. Coccinelle skips over irrelevant functions quickly, essentially only parsing them, and thus the CTL-VW algorithm is only applied to the relevant functions. Because the files are much larger than the relevant functions in this case, the time for parsing dominates the overall running time for most files. We can nevertheless get a sense of the performance of the CTL-VW algorithm in practice, by considering its performance on the relevant functions, in terms of both time and space usage.

We consider two implementations of the model checking algorithm for CTL-VW: an unoptimized one that closely follows the algorithm presented in Section 5.3, and the optimized one that is implemented in the Cocciolle tool. The optimizations affect both the encoding in CTL-VW and the execution of the algorithm, and are summarized as follows:

**Optimizations to the translation** These optimizations seek to reduce the number of witnesses, and thus the amount of information that the algorithm has to propagate.

- Some metavariables, such as \( n_1 \), \( E_1 \), and \( f_1 \) in Figure 1, only occur under a negation. The CTL-VW algorithm...
<table>
<thead>
<tr>
<th>File</th>
<th>total lines</th>
<th>lines in the relevant function</th>
<th>unoptimized total time</th>
<th>unoptimized steps</th>
<th>unoptimized triples</th>
<th>optimized total time</th>
<th>optimized steps</th>
<th>optimized triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>nvmram.c</td>
<td>149</td>
<td>27</td>
<td>0.309</td>
<td>77+29</td>
<td>541</td>
<td>0.306</td>
<td>0.005</td>
<td>67+14</td>
</tr>
<tr>
<td>therm.pm72.c</td>
<td>2279</td>
<td>32</td>
<td>1.466</td>
<td>77+29</td>
<td>678</td>
<td>1.457</td>
<td>0.006</td>
<td>67+12</td>
</tr>
<tr>
<td>fsl-diufb.c</td>
<td>1721</td>
<td>57</td>
<td>1.226</td>
<td>77+47</td>
<td>1093</td>
<td>1.205</td>
<td>0.011</td>
<td>67+27</td>
</tr>
<tr>
<td>smu.c</td>
<td>1323</td>
<td>94</td>
<td>7.707</td>
<td>77+81</td>
<td>2270</td>
<td>0.888</td>
<td>0.043</td>
<td>67+61</td>
</tr>
</tbody>
</table>

**Figure 10.** The performance of applying the semantic patch of Figure 1 to the files that it affects in Linux. Times are in seconds. Steps is the number of CTL operators considered in processing the formula added to the number of steps involved in computing fixpoint iterations for the operator AU. Triples is the sum of the number of triples in the result of processing each CTL operator. The unoptimized and optimized variants both use an implementation of disjunction that does not duplicate the processing of subformulas.

As shown in Figure 10, the combination of these optimizations reduces the CTL-VW processing time by up to over 150 times, due to reduced memory requirements, and the sum of the number of triples manipulated at each step by around 50%. In other work [20], we have created over 60 semantic patches based on collateral evolutions that have taken place in Linux 2.5 and Linux 2.6. In applying these semantic patches to a total of over 5800 relevant Linux files on a 3.4GHz uniprocessor Pentium 4 PC with 1024MB of RAM, the time for applying a semantic patch to a relevant file is rarely more than 0.5 seconds. Thus, our optimizations for CTL-VW model checking provide acceptable performance for interactive use without resorting to efficient but more complex encoding strategies such as BDDs [5].

### 6.6 An extension: Mixing path quantifiers

A advantage of CTL is the ability to mix universal and existential quantification over paths within a single formula. To conclude, we consider how full SmPL can take advantage of this facility. By default, SmPL uses universal quantification when a semantic patch performs transformation, as indicated by $\forall$ and $+$ annotations, and existential quantification when a semantic patch performs only searching (making it instead a semantic match), in which lines are annotated with * to indicate items of interest. These conventions can be overridden, by indicating $\forall all$ or $\exists all$ at the beginning of the rule. But the quantifier used for individual dots can also be controlled locally using $\forall$ or $\exists$.

The semantic match shown in Figure 11 mixes path quantifiers. This semantic match searches for cases where there exists a dereference of the result of the Linux kernel memory allocation function kmalloc without first checking that the result is valid. As the semantic match marks lines using $*$, paths are existentially quantified by default. In lines 7–10, however, we would like to identify a conditional that always aborts if the allocated value is NULL, as subsequent dereferences are known to be safe. Thus, the dots in the body of this conditional are annotated with $\forall$, so there

### 7. Related Work

Lacey and De Moor write CTL-FV formulas to specify compiler optimizations [16]. In this setting, a CTL formula describes properties of the context of a term that allow an optimization such as constant propagation to be applied to the term. Åberg et al. used CTL to specify the transformations required to integrate the run-time system of the Bossa scheduling framework into the Linux kernel source code [1]. In both cases, the state representing code that is affected by the transformation is the one that satisfies the entire formula, so there
is no need for witnesses. Furthermore, the specifications are fairly simple, so it is possible to write CTL code directly. De Moor et al. observed that full CTL was not always necessary for such specifications, and proposed universal regular path queries, which are based on regular expressions [9]. Such queries, however, do not permit mixing universal and existential path quantifiers and also do not permit specifying transformations within formulas.

Other approaches to bug finding that take program control-flow into account include Metal [10] and SDV [2]. Metal is based on specifications expressed as state machines. The expressiveness of state machines and CTL is incomparable. SDV focuses mainly on eliminating false positives by taking possible run-time values into account. Coccinelle does not currently address this issue.

Bohn et al. also propose a variant of CTL with universal and existentially quantified variables [4]. Their goal is to reason about properties involving variables that range over very large or infinite domains, and they provide a syntactic approach to model checking that allows constraints about such variables to be simplified in ad hoc ways. Their semantics is somewhat different than ours in that environments are explicit in the semantics of state operators, but integrated into the labelling function of a specialized version of the initial model in the semantics of path operators. Our use of constructive negation can be viewed as a restricted version of their more general predicates on variable values.

Since SmPL semantic patches can be translated into CTL-VW, they could also be written using CTL-VW directly. Doing so, however, would be very tedious for complex semantic patches, as already illustrated by Figure 7. The difficulty of creating CTL specifications has been recognized in other areas. Corbett et al. propose a high level language for describing desired properties of Java programs, for use with the Bandera model checking framework [8].

Jones and Hansen present a translation of a toy version of SmPL into CTL-V [13]. They concentrate on matching, and do not provide support for transformation, as is enabled by CTL-VW. They implement CTL-V by translation into CTL, at the cost of increasing the size of the formula exponentially in the size of Val.

8. Conclusion

In this paper, we have identified four requirements for a logic that is to serve as the foundation of a control-flow based program matching language, and we have incrementally derived from CTL a logic, CTL-VW, that meets these requirements. In practice, we have found the decision to base the implementation of Coccinelle on an extension of CTL very beneficial, as it naturally separates the specification of the semantics of the program matching language, represented by the translation into the logic, from the implementation, represented by the model checking algorithm. Indeed, in the case of Coccinelle, we iterated many times over the language semantics, but modified the implementation of the model checking algorithm only rarely, to improve performance. Our extensive experiments [20] show that the approach is efficient enough for practical use, on a standard PC.

A limitation of the formalization of CTL-VW presented here is that it does not allow for collecting witnesses under negation. While this feature is not needed in the examples we have considered in practice with Coccinelle, there could be other contexts in which such witnesses would be useful. We want to extend the formalization to accommodate them. Currently, only the proof of soundness and completeness of the CTL-V model checking algorithm has been validated by a proof assistant. We plan to validate the proof for the CTL-VW model checking algorithm as well.

Availability Coccinelle is available from the following URL: http://www.emn.fr/x-info/coccinelle/.

References


Figure 11. A semantic patch that mixes universal and existential path quantification


