Symbolic Learning of Component Interfaces

Dimitra Giannakopoulou[1], Zvonimir Rakamarić[2], and Vishwanath Raman[2]

[1] NASA Ames Research Center, USA
dimitra.giannakopoulou@nasa.gov
[2] Carnegie Mellon University, USA
{zvonimir.rakamaric,vishwa.raman}@gmail.com

Abstract. Given a white-box component $\mathscr{C}$ with specified unsafe states, we address the problem of automatically generating an interface that captures safe orderings of invocations of $\mathscr{C}$'s public methods. Method calls in the generated interface are guarded by constraints on their parameters. Unlike previous work, these constraints are generated automatically through an iterative refinement process. Our technique, named PSYCO (Predicate-based SYmbolic COmpositional reasoning), employs a novel combination of the L* automata learning algorithm with symbolic execution. The generated interfaces are three-valued, capturing whether a sequence of method invocations is safe, unsafe, or its effect on the component state is unresolved by the symbolic execution engine. We have implemented PSYCO as a new prototype tool in the JPF open-source software model checking platform, and we have successfully applied it to several examples.

1 Introduction

Component interfaces summarize aspects of components that are relevant to their clients, thus being at the heart of modular software development and reasoning techniques. Traditionally, component interfaces have been of a purely syntactic form. However, with the advent of component-based distributed development and service-oriented computing, components are no longer just parts of specific closed systems. Rather, modern components are open building blocks that can be reused or connected dynamically, in a variety of environments, in order to form larger systems. As a result, component interfaces must step up to the role of representing component aspects that are relevant to tasks such as dynamic component retrieval and substitution, or functional and non-functional reasoning about systems.

This paper focuses on "temporal" interfaces, which capture ordering relationships between invocations of component methods. For example, an interface may prescribe that closing a file before opening it is undesirable because an exception will be thrown. An ideal interface should precisely represent the component in all its intended usages. In other words, it should include all the good interactions, and exclude all problematic interactions. Previous work presented approaches for computing temporal interfaces using techniques such as predicate abstraction [14] and learning [2, 10, 22].

This work studies a more general problem: automatic generation of precise interfaces for components that include methods with parameters. Whether a method call is problematic or not depends on the actual values passed for its formal parameters. Therefore, we target the generation of interfaces which, in addition to method orderings, also

include method *guards* (i.e., constraints on the parameters of the methods), as illustrated in Fig. 2. We are not aware of any existing approaches that provide a systematic and automated way of introducing guards in temporal interfaces.

Our proposed solution is based on a novel combination of learning with symbolic execution techniques. In particular, we use the L* automata-learning algorithm to automatically generate a component interface expressed as a finite-state automaton over the public methods of the component. L* generates approximations of the component interface by interacting with a *teacher*; the teacher uses symbolic execution to answer queries from L* about the target component, and produces counterexamples when interface approximations are not precise. The generated interfaces are three-valued, capturing whether a sequence of method invocations is safe, unsafe, or its effect on the component state is unresolved by the underlying symbolic execution engine.

Initially, the interface is computed over method names, making no differentiation between input values of parameters, i.e., input values of parameters are unconstrained. The symbolic-execution-based teacher, however, may detect a need for partitioning the space of input parameter values based on constraints computed by the underlying symbolic engine. The alphabet is then refined accordingly, and learning restarts on the refined alphabet. Several learn-and-refine cycles may occur during interface generation.

We have implemented our approach within the JPF (Java Pathfinder) software verification toolset [17]. JPF is an open-source project developed at the NASA Ames Research Center. The presented technique is implemented as a new tool called PSYCO in the JPF project `jpf-psyco`. PSYCO automatically generates rich interfaces for Java classes. To achieve that, it relies on two other projects within JPF: `jpf-learn`, which provides automata learning algorithms, and `jpf-jdart`, which implements dynamic symbolic execution [11] for Java. We have applied PSYCO to learn component interfaces of several realistic examples that could not be handled automatically and precisely using previous approaches.

The main contributions of this paper are as follows:

- A novel combination of learning and symbolic techniques for the generation of rich component interfaces; the approach refines a component's interface alphabet by automatically generating appropriate method guards.
- The use of three-valued automata to precisely record whether a sequence of method invocations is safe, unsafe, or unresolved; subsequent alternative analyses could be employed to explore the unresolved paths.
- Implementation of the approach in a new tool within the JPF software verification toolset and successful application of the tool on several realistic examples.

Related Work. Interface generation for white-box components has been studied extensively in the literature (e.g., [14, 2, 10, 22]). However, as discussed, none of the existing approaches that we are aware of provide a systematic and automated way of refining the interface method invocations using constraints on their parameters.

Automatically creating component models for black-box components is a related area of research. For methods with parameters, abstractions are introduced that map alphabet symbols into sets of concrete argument values. An argument value set represents the corresponding partition, and is used for invoking the component. In the work

```
class PipedOutputStream {
  PipedInputStream sink = null;

  public void connect(
      PipedInputStream snk) {
    if (snk == null) {
      assert false;
    } else if (sink != null ||
        snk.connected) {
      assert false;
    }
    sink = snk;
    snk.connected = true;
  }

  public void write() {
    if (sink == null) {
      assert false;
    } else { ; }
  }

  public void flush() {
    if (sink != null) { ; }
  }

  public void close() {
    if (sink != null) { ; }
  }
}
```

Fig. 1: Motivating example.

by Aarts et al. [1], abstractions are user-defined. Hower et al. [16] discover such abstraction mappings through an automated refinement process. In contrast to that work, availability of the component source code enables us to generate guards that characterize precisely each method partition, making the generated automata more informative. MACE [7] combines black- and white-box techniques since it uses dynamic symbolic execution on binaries of a component, but does so only to discover new concrete input messages that generate new system states. The input alphabet is automatically refined based on a user-provided abstraction of output messages. Again, our work directly leverages symbolic information in order to perform precise alphabet refinement.

Interface generation is also related to assumption generation for compositional verification, where learning-based approaches have been studied for expressing finite-state components as simple LTSs [19, 13, 6, 5]. A different type of alphabet refinement has been developed in that context [9, 4]. These techniques compute smaller alphabets for more efficient compositional reasoning, and alphabets are refined if compositional reasoning is not conclusive with a given alphabet.

2 Motivating Example

Our motivating example is the *PipedOutputStream* class taken from the *java.io* package. Similar to previous work [2, 22], we removed unnecessary details from the example; Fig. 1 gives the simplified code. The example has one private field *sink* of type *PipedInputStream*, and four public methods called *connect*, *write*, *flush*, and *close*. Throwing exceptions is modeled by asserting *false*, denoting an undesirable error state.

The class initializes field *sink* to *null*. Method *connect* takes a parameter *snk* of type *PipedInputStream*, and goes to an error state (i.e., throws an exception) either if *snk* is *null* or if one of the streams has already been connected; otherwise, it connects the input and output streams. Method *write* can be called only if *sink* is not *null*, otherwise an error state is reached. Methods *flush* and *close* have no effect when *sink* is *null*, i.e., they do not throw an exception.

Fig. 2 shows on the right the interface generated with PSYCO for this example. The interface captures the fact that *flush* and *close* can be invoked unconditionally,

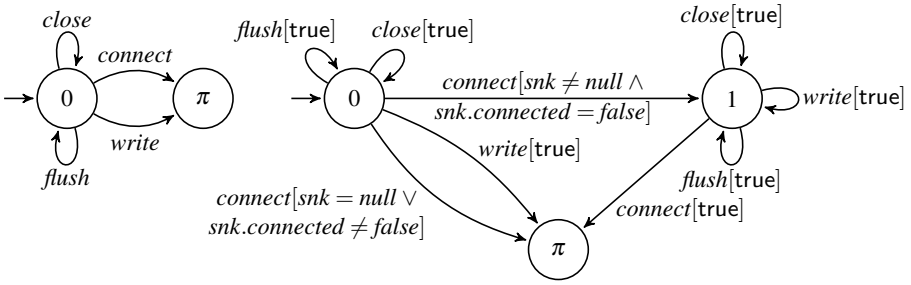

Fig. 2: Interfaces for our motivating example. On the left, there is no support for guards, while on the right, PSYCO is used to generate guards. Initial states are marked with arrows that have no origin; error states are marked with π. Edges are labelled with method names (with guards, when applicable).

whereas *write* can only occur after a successful invocation of *connect*. The guard $snk \neq null \wedge snk.connected = false$, over the parameter *snk* of the method *connect*, captures the condition for a successful connection. Without support for guards in our component interfaces, we would obtain the interface shown on the left. This interface allows only methods that can be invoked unconditionally, i.e., *close* and *flush*. Method *connect* is blocked from the interface since it cannot be called unconditionally. Since *connect* cannot be invoked, *write* is blocked as well. Clearly, the interface on the left, obtained using existing interface generation techniques, precludes several legal sequences of method invocations. In existing approaches, a user is expected to manually define a refinement of the component methods to capture these additional legal behaviors. Our approach performs such a refinement automatically. Therefore, support for automatically generating guards enables PSYCO to generate richer and more precise component interfaces for components that have methods with parameters.

3 Preliminaries

Labeled Transition Systems (LTS). We use deterministic LTSs to express temporal component interfaces. Symbols π and υ denote a special *error* and *unknown* state, respectively. The former models unsafe states and the latter captures the lack of knowledge about whether a state is safe or unsafe. States π and υ have no outgoing transitions.

A deterministic LTS M is a four-tuple $\langle Q, \alpha M, \delta, q_0 \rangle$ where: 1) Q is a finite non-empty set of states, 2) αM is a set of observable actions called the *alphabet* of M, 3) $\delta : (Q \times \alpha M) \mapsto Q$ is a transition function, and 4) $q_0 \in Q$ is the initial state. LTS M is complete if each state except π and υ has outgoing transitions for every action in αM.

A *trace*, also called *execution* or *word*, of an LTS M is a finite sequence of observable actions that label the transitions that M can perform starting from its initial state. A trace is illegal if it leads M to state π, unknown if it leads M to state υ, and legal otherwise. The illegal (resp. unknown, legal) language of M, denoted as $\mathscr{L}_{illegal}(M)$ (resp. $\mathscr{L}_{unknown}(M)$, $\mathscr{L}_{legal}(M)$), is the set of illegal (resp. unknown, legal) traces of M.

Three-Valued Automata Learning with L*. We use an adaptation [6] of the classic L* learning algorithm [3, 20], which learns a three-valued deterministic finite-state automaton (3DFA) over some alphabet Σ. In our setting, learning is based on partitioning

$$
\begin{array}{rcl}
\textit{Component} & ::= & \textbf{class}\ \textit{Ident}\ \{\ \textit{Global}^*\ \textit{Method}^+\ \} \\
\textit{Method} & ::= & \textit{Ident}\ (\textit{Parameters})\ \{\ \textit{Stmt}\ \} \\
\textit{Global} & ::= & \textit{Type}\ \textit{Ident}; \\
\textit{Arguments} & ::= & \textit{Arguments},\ \textit{Expr} \mid \varepsilon \\
\textit{Parameters} & ::= & \textit{Pararameters},\ \textit{Parameter} \mid \varepsilon \\
\textit{Parameter} & ::= & \textit{Type}\ \textit{Ident}
\end{array}
\qquad
\begin{array}{rcl}
\textit{Stmt} & ::= & \textit{Stmt};\ \textit{Stmt} \\
& \mid & \textit{Ident} = \textit{Expr} \\
& \mid & \textbf{assert}\ \textit{Expr} \\
& \mid & \textbf{if}\ \textit{Expr}\ \textbf{then}\ \textit{Stmt}\ \textbf{else}\ \textit{Stmt} \\
& \mid & \textbf{while}\ \textit{Expr}\ \textbf{do}\ \textit{Stmt} \\
& \mid & \textbf{return}\ \textit{Expr}
\end{array}
$$

Fig. 3: Component grammar. *Ident*, *Expr*, and *Type* have the usual meaning.

the words over Σ into three unknown regular languages L_1, L_2, and L_3, with L* using this partition to infer an LTS with three values that is consistent with the partition. To infer an LTS, L* interacts with a teacher that answers two types of questions. The first type is a *membership query* that takes as input a string $\sigma \in \Sigma^*$ and answers *true* if $\sigma \in L_1$, *false* if $\sigma \in L_2$, and *unknown* otherwise. The second type is an *equivalence query* or *conjecture*, i.e., given a candidate LTS M whether or not the following holds: $\mathscr{L}_{legal}(M) = L_1$, $\mathscr{L}_{illegal}(M) = L_2$, and $\mathscr{L}_{unknown}(M) = L_3$. If the above conditions hold of the candidate M, then the teacher answers *true*, at which point L* has achieved its goal and returns M. Otherwise, the teacher returns a counterexample, which is a string σ that invalidates one of the above conditions. The counterexample is used by L* to drive a new round of membership queries in order to produce a new, refined, candidate. Each candidate M that L* constructs is smallest, meaning that any other LTS consistent with the information provided to L* up to that stage has at least as many states as M. Given a correct teacher, L* is guaranteed to terminate with a minimal (in terms of numbers of states) LTS for L_1, L_2, and L_3.

Symbolic Execution. Symbolic execution is a static program analysis technique for systematically exploring a large number of program execution paths [18]. It uses symbolic values as program inputs in place of concrete (actual) values. The resulting output values are then statically computed as symbolic expressions (i.e., constraints), over symbolic input values and constants, using a specified set of operators. A symbolic execution tree, or constraints tree, characterizes all program execution paths explored during symbolic execution. Each node in the tree represents a symbolic state of the program, and each edge represents a transition between two states. A symbolic state consists of a unique program location identifier, symbolic expressions for the program variables currently in scope, and a path condition defining conditions (i.e., constraints) that have to be satisfied in order for the execution path to this state to be taken. The path condition describing the current path through the program is maintained during symbolic execution by collecting constraints when conditional statements are encountered. Path conditions are checked for satisfiability using a constraint solver to establish whether the corresponding execution path is feasible.

4 Components and Interfaces

Components and Methods. A *component* is defined by the grammar in Fig. 3. A component $\mathscr{C}$ has a set of global variables representing internal state and a set of one or more

methods. We assume methods have no recursion. (Note that this is a common assumption since handling recursion in symbolic techniques is a well-known issue orthogonal to this work.) For simplicity of exposition, we assume all methods are public and have unique names, and all method calls are inlined. We also assume the usual statement semantics. Let Ids be the set of component method identifiers (i.e., names), Stmts the set of all component statements, and Prms the set of all input parameters of component methods. We define the signature Sig_m of a method m as a pair $\langle Id_m, P_m \rangle \in \mathsf{Ids} \times 2^{\mathsf{Prms}}$; we write $Id_m(P_m)$ for the signature Sig_m of the method m. A *method* m is then defined as a pair $\langle Sig_m, s_m \rangle$ where $s_m \in \mathsf{Stmts}$ is its top-level statement.

Let $\mathscr{M}$ be the set of methods in a component $\mathscr{C}$ and G be the set of its global variables. For every method $m \in \mathscr{M}$, each parameter $p \in P_m$ takes values from a domain D_p based on its type; similarly for global variables. Given a method $m \in \mathscr{M}$, an execution $\theta \in \mathsf{Stmts}^*$ of m is a finite sequence of visited statements $s_1 s_2 \ldots s_n$ where s_1 is the top-level method statement s_m. The set $\Theta_m \in 2^{\mathsf{Stmts}^*}$ is the set of all unique executions of m. We assume that each execution $\theta \in \Theta_m$ of a method visits a bounded number of statements (i.e., $|\theta|$ is bounded), and also that the number of unique executions is bounded (i.e., $|\Theta_m|$ is bounded); in other words, the methods have no unbounded loops. (Similar to recursion, this is a common assumption and a well-known issue orthogonal to this work.) A *valuation* over P_m, denoted $[P_m]$, is a function that assigns to each parameter $p \in P_m$ a value in D_p. We denote a valuation over variables in G with $[G]$. We take $[G_i]$ as the valuation representing the initial values of all global variables. Given valuations $[P_m]$ and $[G]$, we assume that the execution of m visits exactly the same sequence of statements; in other words, the methods are deterministic.

Symbolic Expressions. We interpret all the parameters of methods symbolically, and use the name of each parameter as its symbolic name; with an abuse of notation, we take Prms to also denote the set of symbolic names. A *symbolic expression* e is defined as follows:

$$e ::= C \mid p \mid (e \circ e),$$

where C is a constant, $p \in \mathsf{Prms}$ a parameter, and $\circ \in \{+, -, *, /, \%\}$ an arithmetic operator. The set of constants in an expression may include constants that are used in statements or the initial values of component state variables in $[G_i]$.

Constraints. We define a *constraint* φ as follows:

$$\varphi ::= \mathit{true} \mid \mathit{false} \mid e \oplus e \mid \varphi \wedge \varphi \mid \varphi \vee \varphi,$$

where $\oplus \in \{<, >, =, \leq, \geq\}$ is a comparison operator.

Guards. Given a method signature $m = \langle Id_m, P_m \rangle$, a guard γ_m is defined as a constraint that only includes parameters from P_m.

Interfaces. Previous work uses LTSs to describe temporal component interfaces. However, as described in Sec. 2, a more precise interface ideally also uses guards to capture constraints on method input parameters.

We define an *interface LTS*, or *iLTS*, to take into account guards, as follows. An iLTS is a tuple $A = \langle M, \mathscr{S}, \Gamma, \Delta \rangle$, where $M = \langle Q, \alpha M, \delta, q_0 \rangle$ is a deterministic and

complete LTS, $\mathscr{S}$ a set of method signatures, Γ a set of guards for method signatures in $\mathscr{S}$, and $\Delta : \alpha M \mapsto \mathscr{S} \times \Gamma$ a function that maps each $\mathtt{a} \in \alpha M$ into a method signature $m \in \mathscr{S}$ and a guard $\gamma_m \in \Gamma$. In addition, the mapping Δ is such that the set of all guards for a given method signature form a partition of the input space of the corresponding method. Let $\Gamma_m = \{\gamma \mid \exists \mathtt{a} \in \alpha M.\Delta(\mathtt{a}) = (m,\gamma)\}$ be the set of guards belonging to a method m. More formally, the guards for a method are (1) non-overlapping:

$$\forall \mathtt{a},\mathtt{b} \in \alpha M,\ \gamma_\mathtt{a},\gamma_\mathtt{b} \in \Gamma,\ m \in \mathscr{S}\ .\ \mathtt{a} \neq \mathtt{b} \wedge \Delta(\mathtt{a}) = (m,\gamma_\mathtt{a}) \wedge \Delta(\mathtt{b}) = (m,\gamma_\mathtt{b}) \Rightarrow \neg\gamma_\mathtt{a} \vee \neg\gamma_\mathtt{b},$$

(2) cover all of the input space: $\forall m \in \mathscr{S}\ .\ \bigvee_{\gamma \in \Gamma_m} \gamma = \mathtt{true}$, and (3) are non-empty.

Given an iLTS $A = \langle M,\mathscr{S},\Gamma,\Delta\rangle$, an execution of A is a sequence of pairs $\sigma = (m_0,\gamma_{m_0}),(m_1,\gamma_{m_1}),\ldots,(m_n,\gamma_{m_n})$, where for $0 \leq i \leq n$, pair (m_i,γ_{m_i}) consists of a method signature $m_i \in \mathscr{S}$ and its corresponding guard γ_{m_i}. Every execution σ has a corresponding trace $\mathtt{a}_0,\mathtt{a}_1,\ldots,\mathtt{a}_n$ in M such that for $0 \leq i \leq n$, $\Delta(\mathtt{a}_i) = (m_i,\gamma_{m_i})$. Then σ is a legal (resp. illegal, unknown) execution in A, if its corresponding trace in M is legal (resp. illegal, unknown). Based on this distinction, we define $\mathscr{L}_{legal}(A)$, $\mathscr{L}_{illegal}(A)$, and $\mathscr{L}_{unknown}(A)$ as the sets of legal, illegal, and unknown executions of A, respectively.

An iLTS $A = \langle M,\mathscr{S},\Gamma,\Delta\rangle$ is an interface for a component $\mathscr{C}$ if $\mathscr{S}$ is a subset of method signatures of the methods $\mathscr{M}$ in $\mathscr{C}$. However, not all such interfaces are acceptable and a notion of interface correctness also needs to be introduced. Traditionally, correctness of an interface for a component $\mathscr{C}$ is associated with two characteristics: *safety* and *permissiveness*, meaning that the interface blocks all erroneous and allows all good executions (i.e., executions that do not lead to an error) of $\mathscr{C}$, respectively. A *full* interface is then an interface that is both safe and permissive [14].

We extend this definition to iLTSs as follows. Let iLTS A be an interface for a component $\mathscr{C}$. An execution $\sigma = (m_0,\gamma_{m_0}),(m_1,\gamma_{m_1}),\ldots,(m_n,\gamma_{m_n})$ of A then represents every concrete sequence $\sigma_c = (m_0,[P_{m_0}]),(m_1,[P_{m_1}]),\ldots,(m_n,[P_{m_n}])$ such that for $0 \leq i \leq n$, $[P_{m_i}]$ satisfies γ_{m_i}. Each such concrete sequence defines an execution of the component $\mathscr{C}$. We say an execution of a component is illegal if it results in an assertion violation; otherwise, the execution is legal. Then, A is a *safe* interface for $\mathscr{C}$ if for every execution $\sigma \in \mathscr{L}_{legal}(A)$, we determine that all the corresponding concrete executions of component $\mathscr{C}$ are legal. It is *permissive* if for every execution $\sigma \in \mathscr{L}_{illegal}(A)$, we determine that all the corresponding concrete executions of component $\mathscr{C}$ are illegal. Finally, A is *tight* if for every execution $\sigma \in \mathscr{L}_{unknown}(A)$, we cannot determine whether the corresponding concrete executions of component $\mathscr{C}$ are legal or illegal; this explicitly captures possible incompleteness of the underlying analysis technique. To conclude, we say A is *full* if it is *safe*, *permissive*, and *tight*. Moreover, we say A is *k-full* for some $k \in \mathbb{N}$ if it is safe, permissive, and tight for all method sequences of length up to k.

5 Symbolic Interface Learning

Let $\mathscr{C}$ be a component and $\mathscr{S}$ the set of signatures of a subset of the methods $\mathscr{M}$ in $\mathscr{C}$. Our goal is to automatically compute an interface for $\mathscr{C}$ as an iLTS $A = \langle M,\mathscr{S},\Gamma,\Delta\rangle$. We achieve this through a novel combination of L* to generate LTS M, and symbolic execution to compute the set of guards Γ and the mapping Δ.

At a high level, our proposed framework operates as follows (see Fig. 4). It uses L* to learn an LTS over an alphabet that initially corresponds to

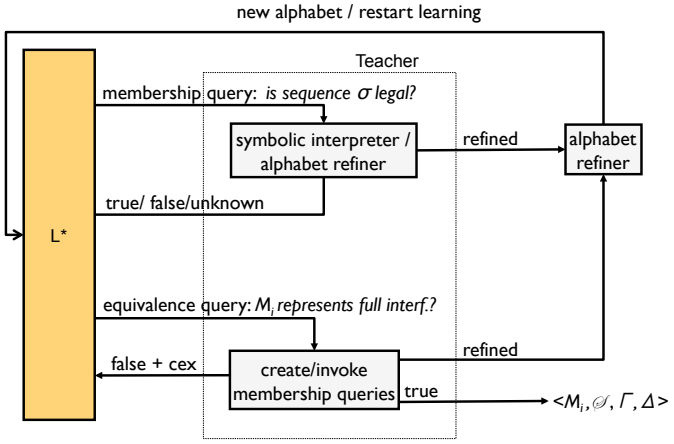

Fig. 4: PSYCO framework during iteration i of learning algorithm.

a set of signatures $\mathscr{S}$ of the methods of $\mathscr{C}$. For our motivating example, we start with the alphabet $\alpha M = \{close, flush, connect, write\}$, set of signatures $\mathscr{S} = \{close(), flush(), connect(snk), write()\}$, and Δ such that $\Delta(close) = (close(), \mathsf{true})$, $\Delta(flush) = (flush(), \mathsf{true})$, $\Delta(connect) = (connect(snk), \mathsf{true})$, and $\Delta(write) = (write(), \mathsf{true})$. As mentioned earlier, L* interacts with a teacher that responds to its membership and equivalence queries. A membership query over the alphabet αM is a sequence $\sigma = \mathtt{a}_0, \mathtt{a}_1, \ldots, \mathtt{a}_n$ such that for $0 \leq i \leq n$, $\mathtt{a}_i \in \alpha M$. Given a query σ, the teacher uses symbolic execution to answer the query. The semantics of executing a query in this context corresponds to exercising all paths through the methods in the query sequence, subject to satisfying the guards returned by the map Δ. Whenever the set of all paths through the sequence can be partitioned into proper subsets that are safe, lead to assertion violations, or to limitations of symbolic execution that prevent further exploration, we refine guards to partition the input space of the methods in the query sequence. We call this process *alphabet refinement*.

For our motivating example, the sequence $\sigma = connect$ will trigger refinement of symbol *connect*. As illustrated in Fig. 2, the input space of method *connect* must be partitioned into the case where: (1) $snk \neq null \wedge snk.connected = false$, which leads to safe executions, and (2) the remaining inputs, which lead to unsafe executions. When a method is partitioned, we replace the symbol in αM corresponding to the refined method with a fresh symbol for each partition, and the learning process is restarted with the new alphabet. For example, we partition the symbol *connect* into *connect_1* and *connect_2*, corresponding to the two cases above, before we restart learning. The guards that define the partitions are stored in Γ, and the mapping from each new symbol to the corresponding method signature and guard is stored in Δ.

Algo. 1 is the top-level algorithm implemented by our interface generation framework. First, we initialize the alphabet αM and the set of guards Γ on line 1. Then, we create a fresh symbol `a` for every method signature m, and use it to populate the alphabet αM and the mapping Δ (lines 2–6). The main loop of the algorithm learns an interface for the current alphabet; the loop either refines the alphabet and reiterates, or

Algo. 1 Learning an iLTS for a component.

Input: A set of method signatures $\mathscr{S}$ of a component $\mathscr{C}$.
Output: An iLTS $A = \langle M, \mathscr{S}, \Gamma, \Delta \rangle$.

```
 1: αM ← ∅, Γ ← {true}
 2: for all m ∈ 𝒮 do
 3:     a ← CreateSymbol()
 4:     αM ← αM ∪ {a}
 5:     Δ(a) ← (m, true)
 6: end for
 7: loop
 8:     AlphabetRefiner.init(αM, Δ)
 9:     SymbolicInterpreter.init(αM, AlphabetRefiner)
10:     Teacher.init(Δ, SymbolicInterpreter)
11:     Learner.init(αM, Teacher)
12:     M ← Learner.learnAutomaton()
13:     if M = null then
14:         (αM, Γ, Δ) ← AlphabetRefiner.getRefinement()
15:     else
16:         return A = ⟨M, 𝒮, Γ, Δ⟩
17:     end if
18: end loop
```

produces an interface and terminates. In the loop, an alphabet refiner is initialized on line 8, and is passed as an argument for the initialization of the *SymbolicInterpreter* on line 9. The *SymbolicInterpreter* is responsible for invoking the symbolic execution engine and interpreting the obtained results. It may, during this process, detect the need for alphabet refinement, which will be performed through invocation of *AlphabetRefiner*. We initialize a teacher with the current alphabet and the *SymbolicInterpreter* on line 10, and finally a learner with this teacher on line 11. The learning process then takes place to generate a classical LTS M (line 12). When learning produces an LTS M that is not *null*, then an iLTS A is returned that consists of M and the current guards and mapping, at which point the framework terminates (line 16). If M is *null*, it means that refinement took place during learning. We obtain the new alphabet, guards, and mapping from the *AlphabetRefiner* (line 14) and start a new learn-refine iteration.

Teacher. As discussed in Sec. 3, the teacher responds to membership and equivalence queries produced by L*. Given a membership query $\sigma = \mathtt{a}_0, \mathtt{a}_1, \ldots, \mathtt{a}_n$, the symbolic teacher first generates a program P_σ. For each symbol $\mathtt{a}_i$ in the sequence, P_σ invokes the corresponding method m_i while assuming its associated guard γ_{m_i} using an assume statement. The association is provided by the current mapping Δ, i.e., $\Delta(\mathtt{a}_i) = (m_i, \gamma_{m_i})$. The semantics of statement **assume** *Expr* is that it behaves as skip if *Expr* evaluates to true; otherwise, it blocks the execution. This ensures that the symbolic execution engine considers only arguments that satisfy the guard, while all other values are ignored.

For the example of Fig. 1, let $\sigma = close, connect\_1, write$ be a query, where $\Delta(close) = (close(), \mathtt{true})$, $\Delta(connect\_1) = (connect(snk), snk \neq null \wedge snk.connected = false)$, and $\Delta(write) = (write(), \mathtt{true})$. Fig. 5 gives the generated program P_σ for this query. Such a program is then passed to the *SymbolicInterpreter* that performs symbolic analysis and returns one of the following: (1) TRUE corresponding

```
void main(PipedInputStream snk) {
  assume true; close();
  assume snk != null && snk.connected == false; connect(snk);
  assume true; write();
}
```

Fig. 5: The generated program P_σ for the query sequence $\sigma = close, connect\_1, write$, where $\Delta(close) = (close(), \mathsf{true})$, $\Delta(connect\_1) = (connect(snk), snk \neq null \wedge snk.connected = false)$, and $\Delta(write) = (write(), \mathsf{true})$.

to a *true* answer for learning, (2) `FALSE` corresponding to a *false* answer, (3) `UNKNOWN` corresponding to an *unknown* answer, and (4) `REFINED`, reflecting the fact that alphabet refinement took place, in which case the learning process must be interrupted, and the learner returns an LTS $M = null$.

An equivalence query checks whether the conjectured iLTS $A = \langle M, \mathscr{S}, \Gamma, \Delta \rangle$, with $M = \langle Q, \alpha M, \delta, q_0 \rangle$, is safe, permissive, and tight. One approach to checking these three properties would be to encode the interface as a program, similar to the program for membership queries. During symbolic execution of this program, we would check whether the conjectured LTS correctly characterizes legal, illegal, and unknown uses of the component. However, conjectured interfaces have unbounded loops; symbolic techniques handle such loops through bounded unrolling. We follow a similar process, but rather than having the symbolic engine unroll loops, we reduce equivalence queries to membership queries of bounded depth. Note that this approach, similar to loop unrolling during symbolic execution, is not complete. Hence, it may fail to detect that an interface is not full, in which case learning terminates early, and produces an LTS that may still be informative.

We generate such bounded membership queries by performing a depth-first traversal of M to some depth k to generate all possible sequences of length k. Each generated sequence belongs to one of three sets $\mathscr{L}_{legal}(M)$, $\mathscr{L}_{illegal}(M)$, or $\mathscr{L}_{unknown}(M)$. Every sequence σ is then queried using the algorithm for membership queries. If the membership query for σ returns `REFINED`, learning is restarted since the alphabet has been refined. Furthermore, if the membership query for a sequence $\sigma \in \mathscr{L}_{legal}(M)$ (resp. $\sigma \in \mathscr{L}_{illegal}(M)$, $\sigma \in \mathscr{L}_{unknown}(M)$) does not return `TRUE` (resp. `FALSE`, `UNKNOWN`), the corresponding interface is not full and σ is returned to L* as a counterexample to the equivalence query. Otherwise, the interface is guaranteed to be k-full, i.e., safe, permissive, and tight up to depth k.

Symbolic Interpreter. Algo. 2 shows the algorithm implemented in *SymbolicInterpreter* and called by the teacher. The algorithm invokes a symbolic execution engine, and interprets its results to determine answers to queries. The input to the algorithm is a program P_σ as defined above, and a set of symbols Σ. The output is either `TRUE`, `FALSE`, or `UNKNOWN`, if no alphabet refinement is needed, or `REFINED`, which reflects that alphabet refinement took place.

Algo. 2 starts by executing P_σ symbolically (line 1), treating main method parameters (e.g., *snk* in Fig. 5) as symbolic inputs. Every path through the program is then characterized by a *path constraint*, denoted by *pc*. A *pc* is a constraint over symbolic parameters, with each conjunct in the constraint stemming from a conditional state-

Algo. 2 Symbolic interpreter.

Input: Program P_σ and set of symbols Σ.
Output: `TRUE`, `FALSE`, `UNKNOWN`, or `REFINED`.

$(PC, \rho) \leftarrow SymbolicallyExecute(P_\sigma)$
$\varphi^{err} \leftarrow \varphi^{ok} \leftarrow \varphi^{unk} \leftarrow \mathsf{false}$
for all $pc \in PC$ **do**
  if $\rho(pc) = \mathsf{error}$ **then**
    $\varphi^{err} \leftarrow \varphi^{err} \vee pc$
  else if $\rho(pc) = \mathsf{ok}$ **then**
    $\varphi^{ok} \leftarrow \varphi^{ok} \vee pc$
  else
    $\varphi^{unk} \leftarrow \varphi^{unk} \vee pc$
  end if
end for
if $\neg(SAT(\varphi^{err}) \vee SAT(\varphi^{unk}))$ **then**
  return `TRUE`
else if $\neg(SAT(\varphi^{ok}) \vee SAT(\varphi^{unk}))$ **then**
  return `FALSE`
else if $\neg(SAT(\varphi^{err}) \vee SAT(\varphi^{ok}))$ **then**
  return `UNKNOWN`
else
  $\Sigma_{new} \leftarrow AlphabetRefiner.refine(\varphi^{err}, \varphi^{unk})$
  if $|\Sigma_{new}| = |\Sigma|$ **then**
    return `UNKNOWN`
  else
    return `REFINED`
  end if
end if

Algo. 3 Symbolic alphabet refinement.

Input: Set of symbols Σ, mapping Δ, and constraints φ^{err}, φ^{unk}.
Output: Refinement Σ_{new}, Γ_{new}, Δ_{new}.

$\Sigma_{new} \leftarrow \Gamma_{new} \leftarrow \emptyset$
for all $\mathtt{a} \in \Sigma$ **do**
  $(m, \gamma) \leftarrow \Delta(\mathtt{a})$
  $\varphi_m^{err} \leftarrow \Pi_m(\varphi^{err})$
  $\varphi_m^{unk} \leftarrow \gamma \wedge \neg\varphi_m^{err} \wedge \Pi_m(\varphi^{unk})$
  if $\neg MP(\varphi_m^{err}) \wedge \neg MP(\varphi_m^{unk})$ **then**
    $\varphi_m^{ok} \leftarrow \gamma \wedge \neg\varphi_m^{err} \wedge \neg\varphi_m^{unk}$
    if $SAT(\varphi_m^{err})$ **then**
      $\mathtt{a}_{err} \leftarrow CreateSymbol()$
      $\Sigma_{new} \leftarrow \Sigma_{new} \cup \{\mathtt{a}_{err}\}$
      $\Gamma_{new} \leftarrow \Gamma_{new} \cup \{\varphi_m^{err}\}$
      $\Delta_{new}(\mathtt{a}_{err}) \leftarrow (m, \varphi_m^{err})$
    end if
    if $SAT(\varphi_m^{unk})$ **then**
      $\mathtt{a}_{unk} \leftarrow CreateSymbol()$
      $\Sigma_{new} \leftarrow \Sigma_{new} \cup \{\mathtt{a}_{unk}\}$
      $\Gamma_{new} \leftarrow \Gamma_{new} \cup \{\varphi_m^{unk}\}$
      $\Delta_{new}(\mathtt{a}_{unk}) \leftarrow (m, \varphi_m^{unk})$
    end if
    if $SAT(\varphi_m^{ok})$ **then**
      $\mathtt{a}_{ok} \leftarrow CreateSymbol()$
      $\Sigma_{new} \leftarrow \Sigma_{new} \cup \{\mathtt{a}_{ok}\}$
      $\Gamma_{new} \leftarrow \Gamma_{new} \cup \{\varphi_m^{ok}\}$
      $\Delta_{new}(\mathtt{a}_{ok}) \leftarrow (m, \varphi_m^{ok})$
    end if
  else
    $\Sigma_{new} \leftarrow \Sigma_{new} \cup \{\mathtt{a}\}$
    $\Gamma_{new} \leftarrow \Gamma_{new} \cup \{\gamma\}$
    $\Delta_{new}(\mathtt{a}) \leftarrow (m, \gamma)$
  end if
end for
return $\Sigma_{new}, \Gamma_{new}, \Delta_{new}$

ment encountered along the path; a path constraint precisely characterizes a path taken through the program. A constraint partitions the set of all valuations over input parameters of the program (i.e., input parameters of the called component methods) into the set of valuations that satisfy the constraint and the set of valuations that do not satisfy the constraint. We denote a set of path constraints as *PC*.

We define a map $\rho : PC \mapsto \{\mathsf{error}, \mathsf{ok}, \mathsf{unknown}\}$ which, given a path constraint $pc \in PC$, returns error (resp. ok) if the corresponding path represents an erroneous (resp. good) execution of the program; otherwise, ρ returns unknown. Mapping pc to unknown represents a case when the path constraint cannot be solved by the underlying

constraint solver used by the symbolic execution engine. Symbolic execution returns a set of path constraints PC and the mapping ρ, which are then interpreted by the algorithm to determine the answer to the query.

After invoking symbolic execution, the algorithm initializes three constraints (φ^{err} for error, φ^{ok} for good, and φ^{unk} for unknown paths) to false on line 2. The loop on lines 3–11 iterates over path constraints $pc \in PC$, and based on whether pc maps into error, ok, or unknown, adds pc as a disjunct to either φ^{err}, φ^{ok}, or φ^{unk}, respectively. Let $SAT : \Phi \mapsto \mathbb{B}$, where Φ is the universal set of constraints, be a predicate such that $SAT(\varphi)$ holds if and only if the constraint φ is satisfiable. In lines 12–17, the algorithm returns TRUE if all paths are good paths (i.e., if there are no error and unknown paths), FALSE if all paths are error paths, or UNKNOWN if all paths are unknown paths.

Otherwise, alphabet refinement needs to be performed; method *refine* of the *AlphabetRefiner* is invoked, which returns the new alphabet Σ_{new} (line 19). If the alphabet size stays the same, no methods have been refined. This can only happen if all potential refinements involve *mixed-parameter* constraints. Informally, a constraint is considered mixed-parameter if it relates symbolic parameters from multiple methods. As explained in Algo. 3, dealing with mixed parameters precisely is beyond the scope of this work. Therefore, Algo. 2 returns UNKNOWN. Otherwise, refinement took place, and Algo. 2 returns REFINED.

Symbolic Alphabet Refinement. The *SymbolicInterpreter* invokes the refinement algorithm using method *refine* of the *AlphabetRefiner*. The current alphabet, mapping, and constraints φ^{err} and φ^{unk} computed by the *SymbolicInterpreter*, are passed as inputs. Method *refine* implements Algo. 3.

In Algo. 3, the new set of alphabet symbols Σ_{new} and guards Γ_{new} are initialized on line 1. The loop on lines 2–31 determines, for every alphabet symbol, whether it needs to be refined, in which case it generates the appropriate refinement. Let $\Delta(\mathtt{a}) = (m, \gamma)$. An operator Π_m is then used to project φ^{err} on the parameters of m (line 4). When applied to a path constraint pc_i, Π_m erases all conjuncts that don't refer to a symbolic parameter of m. If no conjunct remains, then the result is false. For a disjunction of path constraints $\varphi = pc_1 \vee \ldots \vee pc_n$ (such as φ^{err} or φ^{unk}), $\Pi_m(\varphi) = \Pi_m(pc_1) \vee \ldots \vee \Pi_m(pc_n)$. For example, if $m = \langle foo, \{x, y\}\rangle$, then $\Pi_m((s = t) \vee (x < y) \vee (s \leq z \wedge y = z)) \mapsto \mathsf{false} \vee (x < y) \vee (y = z)$, which simplifies to $(x < y) \vee (y = z)$.

We compute φ_m^{unk} on line 5. At that point, we check whether either φ_m^{err} or φ_m^{unk} involve mixed-parameter constraints (line 6). This is performed using a predicate $MP : \Phi \mapsto \mathbb{B}$, where Φ is the universal set of constraints, defined as follows: $MP(\varphi)$ holds if and only if $|Mthds(\varphi)| > 1$. The map $Mthds : \Phi \mapsto 2^{\mathscr{M}}$ maps a constraint $\varphi \in \Phi$ into the set of all methods that have parameters occurring in φ. Dealing with mixed-parameter constraints in a precise fashion would require more expressive automata, and is beyond the scope of this paper. Therefore, refinement proceeds for a symbol only if mixed-parameter constraints are not encountered in φ_m^{err} and φ_m^{unk}. Otherwise, the current symbol is simply added to the new alphabet (lines 27–29).

We compute φ_m^{ok} on line 7 in terms of φ_m^{err} and φ_m^{unk}, so it does not contain mixed-parameter constraints either. Therefore, when the algorithm reaches this point, all of φ_m^{err}, φ_m^{unk}, φ_m^{ok} represent potential guards for the method refinement. Note that φ_m^{err}, φ_m^{unk}, and φ_m^{ok} are computed in such a way that they partition the input space of the

method m, if it gets refined. A fresh symbol is subsequently created for each guard that is satisfiable (lines 8, 14, 20), We update Σ_{new}, Γ_{new}, and Δ_{new} with the fresh symbol and its guard. In the end, the algorithm returns the new alphabet. The computed guards and mapping are stored in local fields that can be accessed through the getter method *getRefinement*() of the *AlphabetRefiner* (see Algo. 1, line 14).

6 Correctness and Guarantees

As discussed, symbolic techniques typically handle loops and recursion by unrolling them a bounded number of times. We consider this source of incompleteness an orthogonal issue to the arguments presented in this section. Hence, we assume loops and recursion are unrolled prior to the application of our algorithm. For simplicity, we also assume method calls are inlined. Our correctness arguments apply to such modified components, whose methods have a finite number of paths due to the unrolling. We provide proofs of our theorems in Appendix A.

We begin by showing correctness of the teacher for L*. In the following lemma, we prove that the program P_σ that we generate to answer a query σ captures all possible concrete sequences for σ. The proof follows from the structure of P_σ.

Lemma 1. (Correctness of P_σ). *Given a component $\mathscr{C}$ and a query σ on $\mathscr{C}$, the set of executions of $\mathscr{C}$ driven by P_σ is equal to the set of concrete sequences for σ.*

The following theorem shows that the teacher correctly responds to membership queries. The proof follows from the finiteness of paths taken through a component and from an analysis of Algo. 2.

Theorem 1. (Correctness of Answers to Membership Queries). *Given a component $\mathscr{C}$ and a query σ, the teacher responds* `TRUE` *(resp.* `FALSE`*,* `UNKNOWN`*) if and only if all executions of $\mathscr{C}$ for σ are legal (resp. illegal, cannot be resolved by the analysis).*

Next, we show that the teacher correctly responds to equivalence queries up to depth k. The proof follows from our reduction of equivalence queries to membership queries since we generate all possible sequences of symbols in the alphabet from a conjectured LTS of length up to k.

Theorem 2. (Correctness to Depth k of Answers to Equivalence Queries). *Let M be an LTS conjectured by the learning process for some component $\mathscr{C}$, Γ the current set of guards, and Δ the current mapping. If an equivalence query returns a counterexample, $A = \langle M, \mathscr{S}, \Gamma, \Delta \rangle$ is not a full interface for $\mathscr{C}$. Otherwise, A is k-full.*

We use the following two lemmas in proving progress and termination of our algorithm. The first lemma is a property of L*, while the second is a property of our alphabet refinement.

Lemma 2. (Termination of Learning). *If the unknown languages are regular, then L* is guaranteed to terminate.*

Lemma 3. (Alphabet Partitioning). *Algo. 3 creates partitions for the alphabet symbols it refines.*

Given that the number of paths through a method is bounded, we can have at most as many guards for the method as the number of these paths, which is bounded. Furthermore, if alphabet refinement is required, Algo. 3 always partitions at least one method. This leads us to the following theorem.

Example	# Methods	Final Alphabet	# Refinements	k	# States	Time (s)
SIGNATURE	5	5	0	3	4	10.5
PIPEDOUTPUTSTREAM	4	5	1	3	3	32.0
INTMATH	8	16	7	1	3	74.1
ALTBIT	2	5	3	4	5	34.8
CEV-FLIGHTRULE	3	5	2	3	3	192.7
CEV	18	23	5	3	9	2846.1

Table 1: Experimental results. "# Methods" is the number of component methods (and also the size of the initial alphabet); "Final Alphabet" the size of the alphabet upon termination; "# Refinements" the number of the performed alphabet refinements; "k" the depth of our equivalence queries; "# States" the number of states in the generated iLTS; "Time" the overall running time in seconds.

Theorem 3. (Progress and Termination of Refinement). *Alphabet refinement strictly increases the alphabet size, and the number of possible refinements is bounded.*

Finally, we characterize the overall guarantees of our framework with the following theorem, whose proof follows from Theorem 2, Theorem 3, and Lemma 2.

Theorem 4. (Guarantees of PSYCO). *If the behavior of a component $\mathscr{C}$ can be characterized by an iLTS, then* PSYCO *terminates with a k-full iLTS for $\mathscr{C}$.*

7 Implementation and Evaluation

We implemented our approach in a tool called PSYCO within the Java Pathfinder (JPF) open-source framework [17]. PSYCO consists of three new, modular JPF extensions: (1) `jpf-learn` implements both the standard and the three-valued version of L*; (2) `jpf-jdart` is our symbolic execution engine that performs concolic execution [11, 21]; (3) `jpf-psyco` implements the symbolic-learning framework, including the teacher for L*. For efficiency, our implementation of L* caches query results in a *MemoizedTable*. To reuse learning results after refinement, PSYCO uses a common *MemoizedTable* for all learning instances. It also uses a naming convention that enables tracing the history of alphabet symbols across refinements. For reuse, it then exploits the fact that if a query σ has a stored result in the *MemoizedTable*, then any future query obtained by substituting any symbol in σ with a refined version of this symbol, will have the same result. Finally, note that programs P_σ are generated dynamically by invoking their corresponding methods using Java reflection.

We evaluated our approach on several realistic examples:

SIGNATURE A class from the *java.security* package used in a paper by Singh et al. [22].

PIPEDOUTPUTSTREAM A class from the *java.io* package and our motivating example (see Fig. 1). Taken from a paper by Singh et al. [22].

INTMATH A class from the Google Guava repository [12]. It implements arithmetic operations on integer types.

ALTBIT Implements a communication protocol that has an alternating bit style of behavior. Howar et al. [16] use it as a case study.

CEV NASA Crew Exploration Vehicle (CEV) 1.5 EOR-LOR example modeling flight phases of a space-craft. The example is based on a Java state-chart model available in the JPF distribution under `examples/jpfESAS`. We translated the example from state-charts into plain Java.

CEV-FLIGHTRULE Simplified version of the example that aims at exposing a flight rule, as in a paper by Giannakopoulou and Pasareanu [10].

Table 1 summarizes the obtained experimental results. For all experiments, `jpf-jdart` used the Yices SMT solver [8]. The experiments were performed on a 2GHz Intel Core i7 laptop with 8GB of memory running Mac OS X. The generated interface automata are shown in Appendix B. For all examples, with the exception of CEV, our technique terminates in a few minutes at most. CEV takes longer since it involves more methods, some of them with a significant degree of branching.

We used a DFS depth $k = 3$ in all examples, except for INTMATH and ALTBIT. For the former, 30 minutes were not sufficient for PSYCO to terminate since the number of conditional branches in method *log10* causes path explosion. However, we noticed by inspection that none of the methods in this class affect global state, and therefore concluded that a depth of $k = 1$ is sufficient. The resulting interface automaton can reach state unknown due to the presence of non-linear constraints, which cannot be solved using Yices. Similarly, by inspection, we identified the need to extend the depth to $k = 4$ for ALTBIT, in order to expose its alternating behavior.

A characteristic of the examples for which PSYCO took longer to terminate was that they involved methods with a large number of conditional branches. This is not particular to our approach, but inherent in any symbolic analysis technique. If n is the number of branches in each method, and a program invokes m methods in sequence, then the number of paths in this program is, in the worst case, exponential in $m * n$. As a result, symbolic analysis of queries is expensive both in branching within each method as well as in the length of the query. Memoization and reuse of learning results after refinement has been the key to ameliorating this problem since a significant number of queries gets answered through lookup in the table rather than through symbolic execution. Furthermore, our `jpf-jdart` project is in its infancy and is being constantly improved, and therefore we expect to see substantial speedups in the near future.

8 Conclusions and Future Work

We have presented the foundations of a novel approach for generating temporal component interfaces enriched with method guards. Significant engineering can be applied from this point on in order to make our approach and tools more efficient in practice. Since symbolic execution is not complete and suffers from path explosion, heuristics will be the key to increasing the applicability of PSYCO. Such heuristics will target both the learning and the symbolic execution part of PSYCO.

PSYCO uses three-valued LTSs to represent interfaces. The unkown state reflects the fact that our analysis has not been able to cover some parts of a component. During compositional verification, such unknowns could be interpreted conservatively as errors, or optimistically as legal states, thus creating bounds within which the behavior of the component lies. Furthermore, future analysis efforts could focus on analyzing these unexplored parts. The interface could also be enriched during testing or usage of the component. Reuse of previous learning results, similar to what is currently performed, could make this process incremental.

In the future, we also intend to investigate ways of addressing mixed parameters more precisely. For example, we plan to combine PSYCO with a learning algorithm for

register automata [15]. This would enable us to relate parameters of different methods through equality and inequality. Finally, we plan to investigate interface generation in the context of compositional verification.

References

1. F. Aarts, B. Jonsson, and J. Uijen. Generating models of infinite-state communication protocols using regular inference with abstraction. In *ICTSS*, pages 188–204, 2010.
2. R. Alur, P. Cerný, P. Madhusudan, and W. Nam. Synthesis of interface specifications for Java classes. In *POPL*, pages 98–109, 2005.
3. D. Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, 1987.
4. S. Chaki and O. Strichman. Three optimizations for assume-guarantee reasoning with L*. *FMSD*, 32(3):267–284, 2008.
5. Y.-F. Chen, E. M. Clarke, A. Farzan, M.-H. Tsai, Y.-K. Tsay, and B.-Y. Wang. Automated assume-guarantee reasoning through implicit learning. In *CAV*, pages 511–526, 2010.
6. Y.-F. Chen, A. Farzan, E. M. Clarke, Y.-K. Tsay, and B.-Y. Wang. Learning minimal separating DFA's for compositional verification. In *TACAS*, pages 31–45, 2009.
7. C. Y. Cho, D. Babić, P. Poosankam, K. Z. Chen, E. X. Wu, and D. Song. MACE: Model-inference-assisted concolic exploration for protocol and vulnerability discovery. In *USENIX Security Symposium*, 2011.
8. B. Dutertre and L. D. Moura. The Yices SMT solver. Technical report, SRI International, 2006.
9. M. Gheorghiu, D. Giannakopoulou, and C. S. Pasareanu. Refining interface alphabets for compositional verification. In *TACAS*, pages 292–307, 2007.
10. D. Giannakopoulou and C. S. Pasareanu. Interface generation and compositional verification in JavaPathfinder. In *FASE*, pages 94–108, 2009.
11. P. Godefroid, N. Klarlund, and K. Sen. DART: Directed automated random testing. *SIGPLAN Not.*, 40(6):213–223, 2005.
12. Guava: Google core libraries. http://code.google.com/p/guava-libraries/.
13. A. Gupta, K. L. McMillan, and Z. Fu. Automated assumption generation for compositional verification. In *CAV*, pages 420–432, 2007.
14. T. A. Henzinger, R. Jhala, and R. Majumdar. Permissive interfaces. In *ESEC/FSE*, pages 31–40, 2005.
15. F. Howar, B. Steffen, B. Jonsson, and S. Cassel. Inferring canonical register automata. In *VMCAI*, 2012.
16. F. Howar, B. Steffen, and M. Merten. Automata learning with automated alphabet abstraction refinement. In *VMCAI*, pages 263–277, 2011.
17. Java PathFinder (JPF). http://babelfish.arc.nasa.gov/trac/jpf.
18. J. C. King. Symbolic execution and program testing. *Commun. ACM*, 19(7):385–394, 1976.
19. C. S. Pasareanu, D. Giannakopoulou, M. G. Bobaru, J. M. Cobleigh, and H. Barringer. Learning to divide and conquer: applying the L* algorithm to automate assume-guarantee reasoning. *FMSD*, 32(3):175–205, 2008.
20. R. L. Rivest and R. E. Schapire. Inference of finite automata using homing sequences. *Inf. Comput.*, 103(2):299–347, 1993.
21. K. Sen, D. Marinov, and G. Agha. CUTE: A concolic unit testing engine for C. In *ESEC/FSE*, pages 263–272, 2005.
22. R. Singh, D. Giannakopoulou, and C. S. Pasareanu. Learning component interfaces with may and must abstractions. In *CAV*, pages 527–542, 2010.

A Proofs

Lemma 1. (Correctness of P_σ). *Given a component $\mathscr{C}$ and a query σ on $\mathscr{C}$, the set of executions of $\mathscr{C}$ driven by P_σ is equal to the set of concrete sequences for σ.*

Proof. Let $\sigma = \mathtt{a}_0, \mathtt{a}_1, \ldots, \mathtt{a}_n$, where for $0 \leq i \leq n$, $\Delta(\mathtt{a}_i) = (m_i, \gamma_{m_i})$. We first show that any $\sigma_c = (m_0, [P_{m_0}]), (m_1, [P_{m_1}]), \ldots, (m_n, [P_{m_n}])$ such that for $0 \leq i \leq n$, $[P_{m_i}]$ satisfies γ_{m_i}, is an execution of $\mathscr{C}$ driven by P_σ. The program P_σ invokes methods of $\mathscr{C}$ in the same order as the query does. Therefore, the only concrete executions that P_σ does not drive are ones that violate the assumptions associated with each method. Since each valuation $[P_{m_i}]$ satisfies guard γ_{m_i}, it follows that P_σ drives σ_c. Reversely, each execution of $\mathscr{C}$ driven by P_σ is a concrete execution σ_c defined by σ. This does not hold either if (1) P_σ calls the methods in a different order, which is not the case, or if (2) the values of arguments passed to methods in an execution of P_σ do not satisfy the corresponding guards, which is not possible given the program's assume statements. □

Theorem 1. (Correctness of Answers to Membership Queries). *Given a component $\mathscr{C}$ and a query σ, the teacher responds* `TRUE` *(resp.* `FALSE`*,* `UNKNOWN`*) if and only if all executions of $\mathscr{C}$ for σ are legal (resp. illegal, cannot be resolved by the analysis).*

Proof. For a membership query σ, our teacher generates and analyzes the program P_σ. Given that the length of the query σ is bounded, the number of compoment method invocations in P_σ is bounded as well. The number of possible paths through a method and the number of statements on each path are bounded (see assumptions in Sec. 4). Consequently, the number of paths through P_σ is bounded, as well as the number of statement on each path. Therefore, the number of path constraints is bounded and the constraints will be exhaustively enumerated by symbolic execution. In Algo. 2, the constraint φ^{ok} is a disjuction of the path constraints for paths mapped to `ok` (see line 7). The constraints φ^{err} and φ^{unk} are computed in the same fashion. The teacher responds with `TRUE` if constraints φ^{err} and φ^{unk} are unsatisfiable, meaning that all paths in P_σ are legal. Similar claims hold for `FALSE` and `UNKNOWN` responses. Therefore, the anwer to a membership query is correct if and only if the program P_σ contains exactly the concrete executions σ_c defined by σ, which follows from Lemma 1. □

Theorem 2. (Correctness to Depth k of Answers to Equivalence Queries). *Let M be an LTS conjectured by the learning process for some component $\mathscr{C}$, Γ the current set of guards, and Δ the current mapping. If an equivalence query returns a counterexample, $A = \langle M, \mathscr{S}, \Gamma, \Delta \rangle$ is not a full interface for $\mathscr{C}$. Otherwise, A is k-full.*

Proof. For an equivalence query, we generate *all* traces of the conjectured LTS of length k, and convert them to membership queries. If the result of a membership query is different from the result expected from the LTS, then this is a witness to the fact that A is not full, by definition. When no counterexample is produced, A is full for all sequences of method invocations of length up to k (the legal language of a component is prefix-closed, so if a sequence of length k is legal, then all its prefixes are also legal). □

Lemma 2. (Termination of Learning). *If the unknown languages are regular, then L* is guaranteed to terminate.*

Proof. Each L* instance is either interrupted for refinement (and thus terminates), or eventually terminates if the unknown languages are regular, and with a minimal (in number of states) LTS, as guaranteed by the characteristics of the algorithm. □

Lemma 3. (Alphabet Partitioning). *Algo. 3 creates partitions for the alphabet symbols it refines.*

Proof. By analyzing lines 4–7 of Algo. 3, it is straightforward to show that every two guards generated for each method are disjoint, and that the union of all three guards $\varphi_m^{err}, \varphi_m^{ok}, \varphi_m^{unk}$ equals γ. Moreover, refinement only takes place when none of $\varphi_m^{err}, \varphi_m^{ok}, \varphi_m^{unk}$ contain mixed parameter constraints. The algorithm then only generates new alphabet symbols for satisfiable guards (lines 8, 14, 20), and therefore only generates non-empty partitions. Note that it is possible for a single partition to be generated, which corresponds to no refinement of the method. □

Theorem 3. (Progress and Termination of Refinement). *Alphabet refinement strictly increases the alphabet size, and the number of possible refinements is bounded.*

Proof. **Progress.** Refinement is triggered by the symbolic execution of a program P_σ generated for a membership query σ. All computed path constraints $pc \in PC$ have the same root, and no two path constraints are equivalent. Any two path constraints pc_i and pc_j diverge at a point that corresponds to a branching condition within some method m, where different outcomes of the condition are stored in each path.

In Algo. 3, at least two of the constraints φ^{ok}, φ^{err}, φ^{unk} are satisfiable whenever refinement is invoked. For simplicity, assume that the two satisfiable constraints are φ^{ok} and φ^{err}. These constraints must contain at least one path condition each, say pc_1 and pc_2. The path conditions pc_1 and pc_2 are not equivalent since they were generated by symbolic execution of the same program, and since each path condition is added to only one constraint by Algo. 2. Let pc_1 and pc_2 diverge at a condition from a method m. Based on our initial observation, the projections $\Pi_m(pc_1)$ and $\Pi_m(pc_2)$ will at least contain constraints that reflect the two outcomes of the condition at which they diverge. So the method m is refined, unless these constraints involve mixed parameters.

As a consequence, either at least one method gets refined by Algo. 3 in more than one partition, or all methods involve mixed parameters. In the former case, the current symbol for method m is substituted by at least two fresh symbols in the new alphabet. Therefore, the new alphabet has at least one symbol more than the previous one. In the later case, the size of the new alphabet returned by Algo. 3 is equal to the size of the current alphabet Algo. 2, in which case Algo. 2 returns `UNKNOWN` to L*, and L* continues the learning process.

Termination. Since mixed-parameter constraints do not trigger refinement, our framework refines a method only with respect to its paths. Since each method has a bounded number of paths, we can have at most as many partitions of a method as the number of its paths. □

Theorem 4. (Guarantees of PSYCO). *If the behavior of a component $\mathscr{C}$ can be characterized by an iLTS, then* PSYCO *terminates with a k-full iLTS for $\mathscr{C}$.*

Proof. Overall termination follows from Theorem 3 and Lemma 2. The returned iLTS is guaranteed to be k-full by Theorem 2. □

B Generated Interfaces

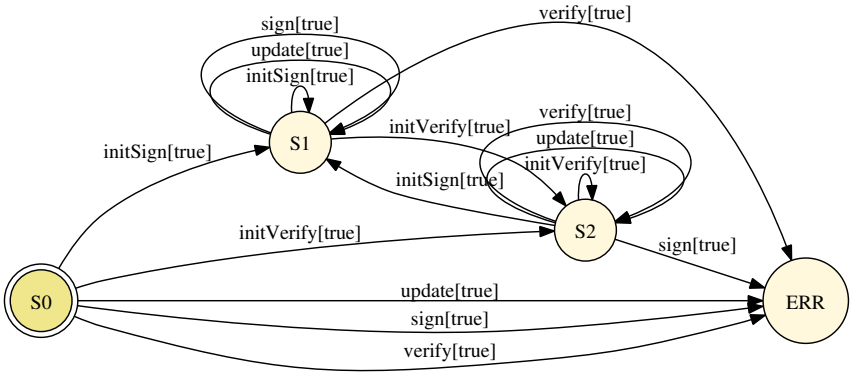

Fig. 6: iLTS for Signature.

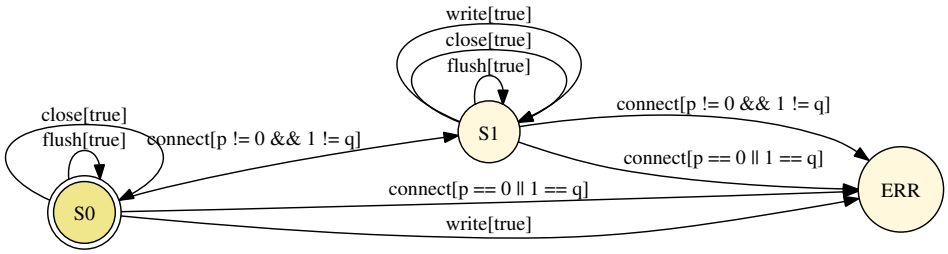

Fig. 7: iLTS for PipedOutputStream.

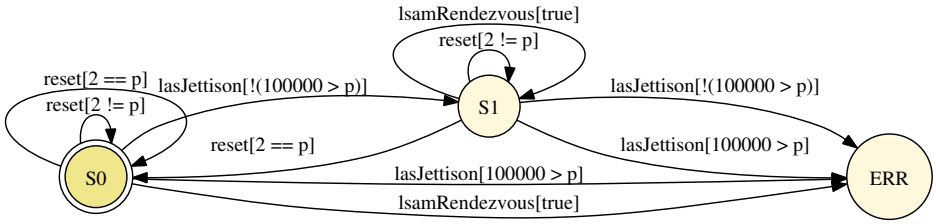

Fig. 8: iLTS for CEV Flight Rule.

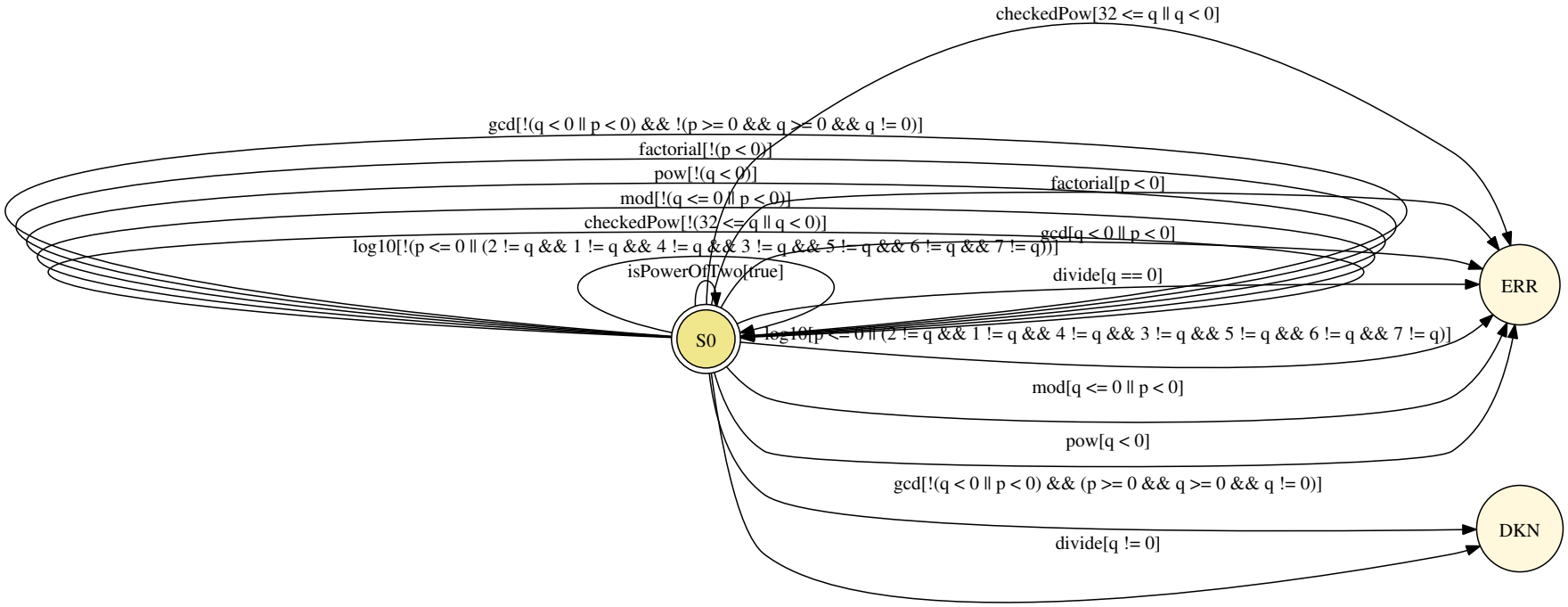

Fig. 9: iLTS for IntMath.

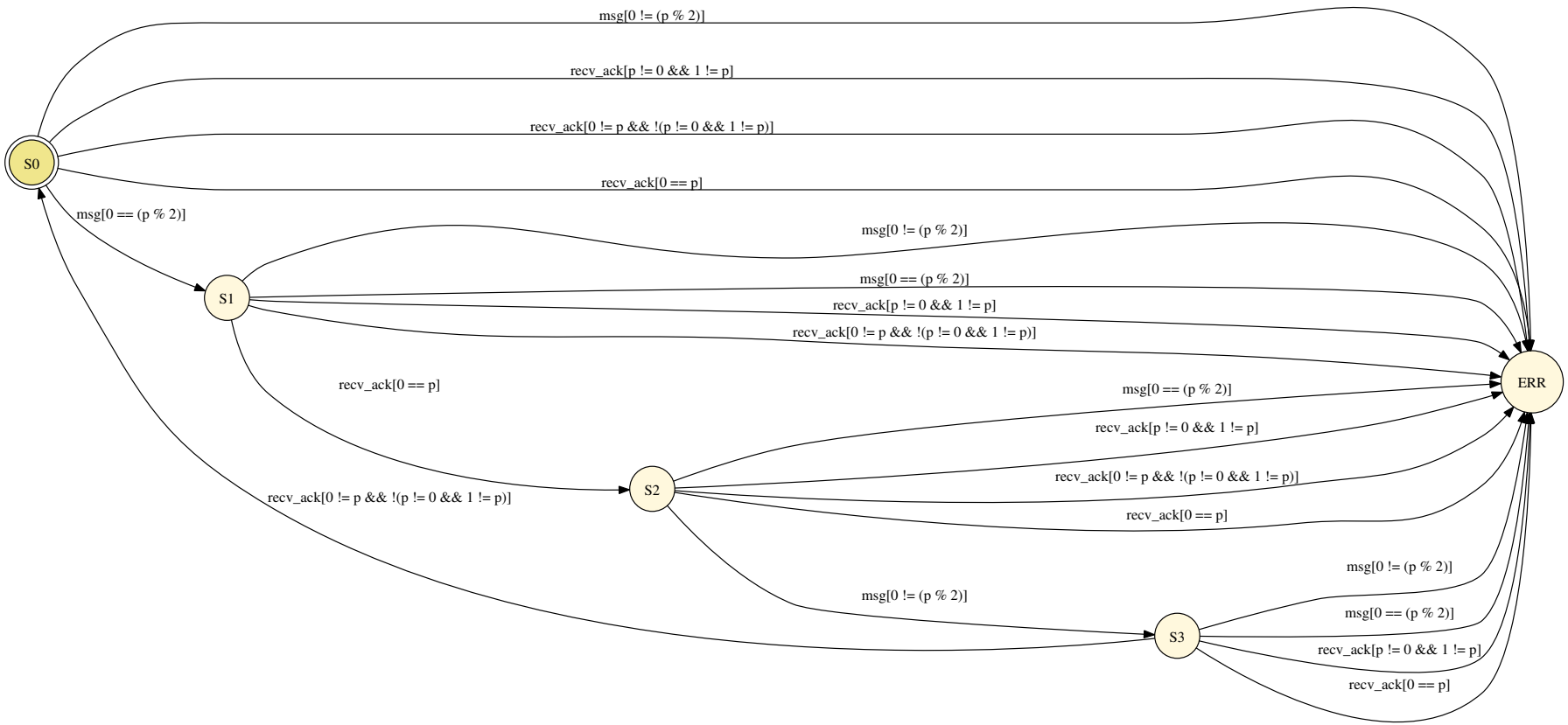

Fig. 10: iLTS for AltBit.

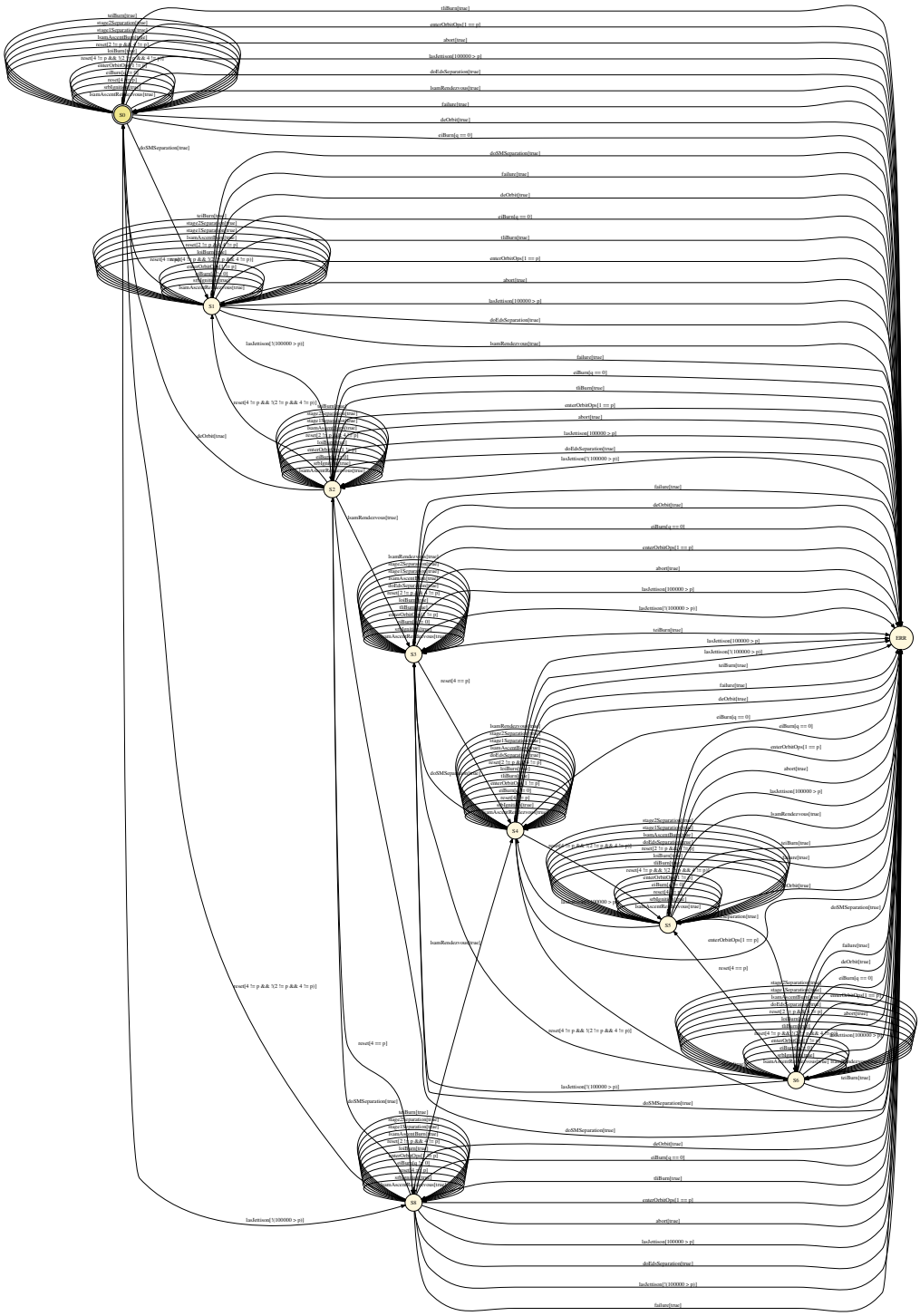

Fig. 11: iLTS for CEV.