Property Differencing for Incremental Checking

Guowei Yang1, Sarfraz Khurshid2, Suzette Person3, Neha Rungta4
1Department of Computer Science, Texas State University, San Marcos, TX 78666, USA
2Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712, USA
3NASA Langley Research Center, Hampton, VA 23681, USA
4NASA Ames Research Center, Moffett Field, CA 94035, USA
1gyang@txstate.edu, 2khurshid@utexas.edu, 3suzette.person@nasa.gov, 4neha.s.rungta@nasa.gov

ABSTRACT
This paper introduces iProperty, a novel approach that facilitates incremental checking of programs based on a property differencing technique. Specifically, iProperty aims to reduce the cost of checking programs as they are initially developed and as they co-evolve with the program. The key novelty of iProperty is to compute the differences between the new and old versions of expected properties to reduce the size and number of the properties that need to be checked during the initial development of the properties. Furthermore, property differencing is used in synergy with program behavior differencing techniques to optimize common regression scenarios, such as detecting regression errors or checking feature additions for conformance to new expected properties. Experimental results in the context of symbolic execution of Java programs annotated with properties written as assertions show the effectiveness of iProperty in utilizing change information to enable more efficient checking.

Categories and Subject Descriptors
D.2.4 [Software Engineering]: Software/Program Verification—assertion checkers; D.2.5 [Software Engineering]: Testing and Debugging—symbolic execution, testing tools; F.3.1 [Logics and Meanings of Programs]: Specifying and Verifying and Reasoning about Programs—assertions, invariants

General Terms
Verification, Algorithms

Keywords
Incremental symbolic execution, assertions, change-impact analysis, Symbolic PathFinder, Daikon

1. INTRODUCTION

Annotating functional correctness properties of code, e.g., using assertions [17] or executable contracts, such as those supported by the Java Modeling Language [37] or Eiffel [39], offers a number of benefits in automated conformance checking of program behaviors to expected properties to support bug finding [18, 22, 29, 30, 38, 57]. Effectively utilizing such properties in practice is complicated by two basic factors. One, it requires the properties to be written and maintained meticulously, so they correctly reflect the expected behaviors of the code, even as it evolves. Two, it requires efficient and cost-effective techniques to check the actual behaviors of the code with respect to the given properties.

We believe an approach that integrates automated program analyses with the development and maintenance of code annotated with properties allows us to address both factors by making it possible for developers to efficiently check and correct the properties and the code they write. Our focus in this paper is on reducing the computational cost of such analyses, specifically in the context of change to properties or to code—due to bugs fixed or feature additions (i.e., addition of new functionality).

In this work, we introduce iProperty, a novel approach to compute differences between properties of related programs in a manner that facilitates more efficient incremental checking of conformance of programs to properties. The key novelty of iProperty is to compute logical differences between old and new versions of properties that undergo change and facilitate the two development scenarios that form our focus: (1) writing properties correctly for an existing program, where the developer writes partial correctness properties, checks them, and then iteratively refines and re-checks them; and (2) checking changes to code and properties as they co-evolve to, for example, check feature additions for conformance to new expected properties or to check bug fixes for regression errors.

In a program annotated with properties that was checked previously but has since undergone change, iProperty utilizes the results of the previous analysis and information about the change to the properties to reduce the number of properties that need to be checked. Our insight is that if a property $\Phi$ about program $p$ holds at a control point $l$ in $p$ and $\Phi'$ implies another property $\Phi''$, $\Phi''$ also holds at $l$ in $p$. This simple observation forms the basis of our algorithm for computing property differences, which checks for implications between corresponding old and new properties to minimize the set of properties that must be checked after
a property change is made and hence reduce the overall cost of checking.

To improve incremental property checking, we leverage the distinction between changes to properties, i.e., assertions, and changes to the code that implements the program functionality. iProperty, uses the property differences in conjunction with code-based change impact analysis techniques to direct checking onto relevant parts of code and properties, thereby removing much of the redundancy otherwise present in re-applying the checking technique. iProperty uses the property differencing algorithm together with a previously developed behavioral differencing technique [45, 50].

iProperty supports incremental checking of properties written as assertions or that can be mechanically transformed into assertions. The assert statements in the source code are de-sugared into if-then statements in the object-code, which allows the use of traditional code change impact analyses, e.g., to compute behavioral differences [14, 45, 50, 51]. Change impact analysis techniques do not distinguish between code that implements functionality and code that represents expected behaviors and thus, are unable to characterize whether an assertion violation represents a regression error, incorrect update to the property, or non-conformance between addition of new features and properties.

To evaluate the efficacy of iProperty, we apply it in the context of symbolic execution of Java programs annotated with assertions to check for assertion violations. We use subject programs from recent work on incremental symbolic execution, which focused on changes to the code [45, 60]. To annotate these subjects with assertions and to simulate assertion changes, we (a) write some assertions manually, and (b) synthesize some assertions automatically using the Daikon tool [19]. While our application of Daikon for assertion synthesis was inspired by Daikon’s spirit of discovering likely invariants, our specific context of symbolic execution enabled us to systematically validate the output generated by Daikon and make it output more precise, thereby laying the foundation of a future approach for more accurate discovery of likely invariants [61]. This paper makes the following contributions:

- **Property differencing.** We introduce the idea of incremental checking driven by property differencing.
- **iProperty.** We present algorithms to compute logical differences between old and new properties and utilize checking results from the previous analysis
- **Assertion-driven incremental checking.** Assertion differences are combined with a change impact analysis technique [45] to detect regression errors and check conformance of new functionality to new expected properties.
- **Experimental evaluation.** The experimental results show that iProperty can (1) provide a reduction in the number of SMT solver calls and total analysis time required for symbolic execution of Java programs compared to the conventional approach of using the Symbolic PathFinder tool [42, 43], and (2) enable verification of properties that could not be verified in a non-incremental manner.

2. MOTIVATING EXAMPLE

This section presents a motivating example to illustrate iProperty and how it is applied in the context of symbolic execution.

## Figure 1: Method to compute the middle value of three input numbers [31] and expected properties (Lines [16] [19].)

```java
1 int median(int x, int y, int z) {
2     int m = z;
3     if (y < z) {
4         if (x < y)
5             m = y;
6         else if (x < z)
7             m = x;
8     }
9     else {
10         if (x > y)
11             m = y;
12         else if (x > z)
13             m = x;
14     }
15
16     assert x == y | x == z ? m == x : true; //#1
17     assert y == z ? m == y : true; //#2
18     assert x <= z && z <= x ? m == z : true; //#3
19     assert y <= z && z <= y ? m == z : true; //#4;
20     return m;
21 }
```

Symbolic execution uses symbolic values in lieu of concrete values for program inputs and builds a path condition along feasible execution paths. A path condition consists of a set of constraints over the symbolic input values and constants. Satisfiability of the constraints is checked using off-the-shelf SMT solvers. In this work, we use Symbolic PathFinder (SPF) [42, 43], an open-source tool for symbolic execution of Java programs built on top of the Java PathFinder (JPF) model checker [57].

Figure 1 shows the Java implementation of median, which computes the middle value of its three integer inputs; the code for median appears in previous work [31], but without property annotations. We manually add partial correctness properties, specifically post-conditions, to this example in the form of assertions just before the method exit.

**Version 1.** Assume the user writes four assertions to check expected properties. The two assertions on Lines [16] [17] check median when at least two inputs are equal, i.e., \( x = y \lor x = z \Rightarrow m = x \), and \( y = z \Rightarrow m = y \). The two assertions on Lines [18] [19] check median when the definition of \( m \) at Line 2 reaches the return statement, i.e., the program behaves correctly when \( z \) is the middle value; specifically, the user asserts \( x \leq z \land z \leq y \Rightarrow m = z \), and \( y \leq z \land z \leq x \Rightarrow m = z \).

Symbolically executing this method using SPF checks the given properties (Lines [16] [19]) and reports no assertion failures after 2 seconds. SPF explores 173 states and makes 172 calls to the underlying SMT solver.

**Version 2: Property modification and addition.** Having gained some confidence in the implementation, assume now the user decides to check the program after editing some assertions and adding new assertions that check more program behaviors. Specifically, the user (1) modifies the third and fourth assertions to use ‘\<‘ instead of ‘\<=‘, e.g., to reduce the overlap between the different sets of inputs each assertion checks, and (2) adds four new assertions to check cases previously unchecked. The following sequence of assertions replaces the assertions in Version 1 (Lines [16] [19]).

```java
23    assert x == y | x == z ? m == x : true; //#1
24    assert y == z ? m == y : true; //#2
25    assert x < z && z < y ? m == z : true; //#3'
26    assert y < z && z < x ? m == z : true; //#4'
```
Applying the conventional symbolic execution approach to check the median program we invoke SPF once which reports no assertion failures after 3 seconds; for this check, SPF explores 301 states and makes 300 SMT solver calls. In contrast, applying iProperty to check the median program with this modified sequence of assertions we invoke SPF three times. Two of these invocations focus on the new assertions that were added (Lines \[25\], \[26\]) and the third invocation focuses on the old assertions that were not modified (Lines \[23\], \[24\]).

The first invocation directly checks whether the Version 1 assertion \#3 (Line \[18\]) implies the Version 2 assertion \#3' (Line \[23\]) by checking if the following method causes an assertion failure:

```java
void checkImplies(int x, int y, int z, int m) {
    if (x <= z && z <= y ? m == z : true )
        assert x < z && z < y ? m == z : true;
}
```

This SPF invocation takes less than 1 second and reports no assertion failure, i.e., the implication holds. Thus, the validity of Version 1 assertion \#3 guarantees the validity of Version 2 assertion \#3' so it is not necessary to check it against the implementation of median. In this case, SPF explores 19 states and invokes the CVC3 solver 18 times.

The second invocation of SPF similarly checks that Version 1 assertion \#4 (Line \[19\]) implies Version 2 assertion \#4' (Line \[26\]) and hence Version 2 assertion \#4' is also valid and does not need to be checked explicitly against the implementation of median. SPF takes less than 1 second, explores 19 states, and makes 18 CVC3 solver calls to perform the check.

The third invocation of SPF checks only the Version 2 assertions \#5, \#6 against the implementation of median. This invocation takes less than 1 second and reports no assertion failure. In this invocation, SPF explores 121 states and makes 120 calls to the CVC3 solver.

Overall, using iProperty to check the Version 2 properties makes three invocations of SPF taking a total of 2 seconds (versus 3 seconds in the conventional approach), requires SPF to explore 159 states (versus 301 states in conventional approach), and makes 156 solver calls to CVC3 (versus 300 calls in conventional approach). Thus, iProperty provides a 48% reduction in SMT solver calls compared to the conventional approach using SPF for this small example.

### 3. APPROACH

In this section, we present the details of the iProperty approach that computes property differences for more efficient incremental checking of evolving programs. iProperty operates on two program versions, \(p\) and \(p'\), leverages the results from checking \(p\) in order to compute the logical differences in evolving properties. iProperty then combines the property differences and a code change impact analysis to enable efficient incremental property checking of \(p'\).

Given two related program versions \(p\) and \(p'\), where \(p\) is annotated with one or more properties, and the properties in \(p'\) are changed with respect to \(p\); for example, properties are added, deleted, or modified in \(p'\). Moreover, \(p'\) may contain changes to the code that result in functional differences, e.g., added, changed, or removed program behaviors. Assume \(p\) was previously checked for property failures and the results from checking \(p\) are available. Then the basic problem addressed by our technique is how to efficiently check the properties in \(p'\). Our technique addresses this problem in two ways: 1) by defining a mechanism to compute property differences between \(p\) and \(p'\) that leverages the results from checking \(p\) to automatically re-write the properties in \(p'\), retaining only the necessary parts of the properties, and 2) by checking \(p'\) incrementally with respect to the impact of the code changes along with the property differences.

Section 3.1 describes the initial property checking. Section 3.2 describes the algorithmic details of computing differences in properties specified as \texttt{assert} statements in Java source code. Section 3.3 describes three checks that can be performed using the property differences and the results of a change impact analysis to efficiently check for regression errors and conformance of new features in the code to the additional properties.

#### 3.1 Initial Property Checking

In this work, we analyze properties specified as Java \texttt{assert} statements. We assume the specified properties are side-effect free.

**Definition 3.1.** A is a set of assertions in a Java program \(p\) such that each \(a \in A\) is an Java assertion “\texttt{assert } \Phi\)” where \(\Phi\) specifies one or more properties as a conjunction of clauses, \(\Phi_a := \phi_0 \land \ldots \land \phi_m (m \geq 0)\).

Each clause, \(\phi\), in a conjunction \(\Phi\) can be an arbitrarily complex boolean expression and may include one or more invocations of pure functions (side-effect free).

iProperty uses the results from checking the previous version of the properties in the property differencing algorithm presented in Section 3.2. The results from analyzing a program, \(p\), with a given set of initial properties, \(A\), are stored for re-use by iProperty. During the initial analysis of the program we record the \texttt{outcome} from checking a property — whether the property passed or failed. In the case a property fails, program analysis techniques such as symbolic execution generate a \texttt{counterexample} to show how the property failed. We first define two data structures to store the analysis results: \texttt{Outcome} and \texttt{Counterexample}.

**Definition 3.2.** Outcome is a map \(\text{Out}_p : A \mapsto \{\text{Fail}, \text{Not Fail}, \text{Unknown}\}\) that returns one of three possible outcomes of checking an assertion \(a\) with respect to a program \(p\) (1) \texttt{Fail} indicates that a violation of \(a\) was detected, (2) \texttt{Not Fail} indicates no failure of \(a\) was found, and (3) \texttt{Unknown} indicates that the result of checking \(a\) are not known.

**Definition 3.3.** Counterexample is a map \(\text{Cer}_p : A \mapsto T\) that returns a set of traces where each trace represents a feasible execution path from the start of the program \(p\) to the violation of an assertion, \(a\) and \(\text{Out}_p(a) := \text{Fail}\).

The \texttt{Outcome} map for \(p\) is populated with the outcomes generated when checking the corresponding assertions in \(A\) against program \(p\); while the \texttt{Counterexample} map is populated with assertions which have the \texttt{Fail} outcome.

**Definition 3.4.** Transform is a map \(\text{trans}_p : A \mapsto T\) that returns the set of clauses contained in the conjunction, \(\Phi_a := \phi_0 \land \ldots \land \phi_m\), for the assertion \(a \in A\).
procedure PropertyDifferencing(A’, Cex_p, Out_p)
1: ∀a’ ∈ A’. Cex_p(a’) := ∅
2: for each a’ ∈ A’ do
3: if matchedAssertions(a’) ≠ ⊥ then
4:   (outcome, a, δ) := AssertDiff(a’, a’)
5:   else (outcome, a, δ) := (Unknown, transform(a’))
6:   Out_p(a’) := outcome
7:   ΔA’\A(a’) := δ a
8: if outcome == Fail then
9:   Cex_p(a’) := Cex_p(a
10: procedure AssertDiff(a’, a)
11: if Outp(a) == Unknown then
12: return (Unknown, transform(a’))
13: if LogicalDifference(a’, a) == ⊥ then
14: return (Outp(a), ⊥)
15: if Outp(a) == Not Fail then
16: return (Unknown, {LogicalDifference(a’, a)})
17: else if Outp(a) == Fail then
18: return (Unknown, transform(a’))

Figure 2: The property differencing algorithm.

3.2 Property Differencing (No Code Changes)

The key idea our approach is to automatically re-write assertions in p’, based on the logical differences between assertions in p and p’. The logical differences are combined with the information gathered from the previous analysis of p, in order to reduce the number of properties that need to be checked in p’. In this section we assume that there are no code changes between programs p and p’, i.e., they are syntactically the same and semantically equivalent when assertions are turned off. In Section 3.3 we take into account code changes.

At the heart of computing property differences is the notion of logical difference between two properties. In this work we define logical difference between two Java assertions as follows:

Definition 3.5. LogicalDifference is a function that takes as input two assertions a’ and a, and returns the set of clauses in a’ where \( \{ φ’ ∈ transform(a’) | Φ_a \nrightarrow φ’ \} \).

The logical difference function returns the set of clauses in a’ that are not subsumed, i.e., logically implied, by the clauses in a. Moreover, if the logical difference function returns the empty set, it follows that \( Φ_a \nrightarrow Φ_a’ \), i.e., all of the clauses in the conjunction of a’ are subsumed by those in a; indicating there is no functional change going from a to a’.

Lemma 3.6. Assume checking p with a at location l shows that assertion a holds. Let δ = LogicalDifference(a’, a). Then, checking p with conjunction \( \bigwedge_{i=0}^{\delta} δ_i \) at control-point l is equivalent to checking p with assertion a’ at l, i.e., the set of executions of p that violate a’ at l equals the set of executions of p that violate \( \bigwedge_{i=0}^{\delta} δ_i \) at l.

Lemma 3.7. If checking p with a at location l results in a violation of a and a’ ⇒ a, then a’ is also violated at l in p.

The PropertyDifferencing algorithm is shown in Figure 2. The inputs to the algorithm are the set of all assertion statements in p’ (A’), and the Outcome and Counterexample maps from checking p, (Out_p) and (Cex_p) respectively. As the PropertyDifferencing algorithm iterates over each assertion in p’, it computes the logical difference between matching assertions in p and p’ and uses the results from analyzing p to update i) the outcome map of p’, Out_p, ii) the counter-example map, Cex_p, and iii) the map of assertions differences, ΔA’\A, that contains for each assertion in p’, the set of clauses in a’ that need to be checked against p’. The entries in Out_p and Cex_p are initialized to Unknown and ⊥ respectively for each assertion at line 1 during the initialization of the PropertyDifferencing algorithm.

The algorithm checks if a’ is a new assertion added to p’ by calling the function matchedAssertions at line 3 in Figure 2. This function takes as input an assertion statement in p’ and returns the corresponding assertion statement in p; when no corresponding assert statement is found, it returns ⊥. Assert statements can be matched using various techniques such as an AST matching algorithm 20 or an IDE monitor that tracks the changes to the code. When no matching assert statement in present in p, then the analysis concludes that a’ is a new assertion added to p’.

In this case, the outcome for a’ is set to Unknown (lines 5 and 6) and the set of clauses in a’ is added to ΔA’\A in order to be checked against p’ (line 7). When the matchedAssertions function returns a matching assert statement, a, at line 3, the PropertyDifferencing algorithm computes the differences between a and a’ by invoking the AssertDiff function.

The AssertDiff procedure shown at lines 11–18 in Figure 2 computes and returns the outcome and the logical difference of a’ with respect to the matched assertion in p. When the outcome of the matched assertion, a, is Unknown with respect to p then a’ is treated as a new assertion by setting its outcome to Unknown (lines 11–12) and adding it to the set of assertions to be checked in p’ (lines 6–7).

When there is no logical difference between assertions a and a’, and the outcome of a is either Fail or Not Fail, then the outcome of a’ is returned as the outcome of a’ at lines 13–14 in Figure 2. There are two cases to consider when, however, there is a logical difference between assertions a and a’. First consider the Not Fail case at lines 15 and 16. In the absence of code changes, the logical difference between the assertions is added to the difference map ΔA’\A to be checked for failures, and its outcome is set to Unknown. Next, consider the Fail case at lines 17 and 18. Since there is a difference between the two assertions we cannot reason about whether the new assertion a’ will fail or not. Hence, the outcome for a’ is set to Unknown and the set of clauses in a’ is added to ΔA’\A.

We also use previously generated counterexamples to potentially reduce the number of properties that need to be checked using an an expensive analysis technique such as symbolic execution. Often a program execution that is successful in eliciting a violation of one property may be able to detect a violation of another property as well. In order to leverage this, we concretely execute inputs of the program that lead to the counter-examples in Cex_p and check for all assertions, a’\f, such that Out_p(a’\f) := Unknown if a violation of the assertion is detected. For these additional failing assertions their corresponding entries in the (a) outcome map are marked as Fail, (b) difference map are marked as empty, i.e., ΔA’\A(a’\f) := ⊥, and (c) counterexample map are updated with the error trace.

The algorithm in Figure 2 computes the map of assertion differences, ΔA’\A. For every assertion a’, the map contains a set of clauses (conjunction of clauses) that we need to check with respect to program p’. Each entry in the map ΔA’\A is mapped into assert statement assert \( φ \) at the location of a’ in p’. The clauses set is re-written as a Java expression.
analyze program $p'$ with re-written assertions using symbolic execution for more efficient checking of the properties.

### 3.3 Property Checking (with Code Changes)

In this section, we present our incremental property checking approach in the presence of code changes. Recall that earlier, while computing the property differences we assumed there are no code changes between $p$ and $p'$—only property changes. Our goal is to partition the incremental property checking in a manner that allows more efficient detection of regression errors introduced during bug fixes and checking conformance of feature additions to new properties.

The incremental property checking algorithm combines in a novel way results computed by the property differing algorithm in Section 3.2 and our previous work on Directed Incremental Symbolic Execution (DiSE) that generates the set of program behaviors impacted by the code changes. The combination of the property differences with DiSE allows us to partition the incremental analysis into three different checks: (i) **Check I** analyzes the changed or impacted program behaviors with respect to the assertion differences; (ii) **Check II** analyzes the changed or impacted behaviors with respect to the original (preserved) assertions; and (iii) **Check III** analyzes the unchanged program behaviors with respect to the assertion differences. The three checks can be performed independent of each other. The partitioning also provides insight into whether the property violation was caused by the property differences (incorrect update of properties), code changes (regression errors), or a combination of property differences and code changes (new feature additions do not conform to the new properties).

#### 3.3.1 Directed Incremental Symbolic Execution

We first present a brief overview of the DiSE methodology. Given two related program versions $p$ and $p'$, (1) DiSE constructs the change set $C$, which contains the set of program statements that are syntactically different between $p$ and $p'$, (2) DiSE then uses the change set $C$ as the slicing criterion to statically compute the impact set, $I$, the set of program statements that may be impacted by the changes in $p'$, and (3) Symbolic execution of $p'$ is then directed along the program locations in the impact set, $I$, in order to generate the program behaviors impacted by the changes in $C$.

```java
1: int a, b;
2: public int test(int x, int y) {
3:   if (x > 0) a = x+1;
4:   else a = x+2;
5:   if (y == 0) b = y+1;
6:   else b = y+2; // change
7: }
```

The simple program shown above contains two global variables $a$ and $b$, and two input parameters $x$ and $y$. Updates to the global variables are made based on the values of the input variables. Suppose the assignment $b := y + 2$ is changed to $b := y - 1$ (line 7). The program statement impacted by this change is the conditional statement ($y == 0$). The execution of the changed statement is control dependent on the false branch of the conditional statement. DiSE uses this impact information to direct symbolic execution. There are four possible paths through the program defined by these path constraints: (a) $x > 0 \land y == 0$, (b) $x > 0 \land y \neq 0$, (c) $x \leq 0 \land y == 0$, and (d) $x \leq 0 \land y \neq 0$. DiSE, however, generates only two paths out of the four. In its default search strategy DiSE will explore paths (a) and (b), and then prune the other paths since all the impacted program behaviors (operations on $y$ and $b$) have been explored in paths (a) and (b). DiSE has been shown to be an efficient technique for characterizing evolving program behaviors.

#### 3.3.2 Three Types of Property Checks

The three types of property checks in our incremental analysis are shown in Figure 3 (Check I) analyze the impacted program behaviors with respect to the property differences, (Check II) analyze the impacted program behaviors with respect to the preserved properties, and (Check III) analyze unchanged program behaviors with respect to the property differences. Programs $p_f$ and $p'_f$ are assertion-free program versions of $p$ and $p'$ respectively, $A$ and $A'_f$ are the sets of assertions in $p$ and $p'$ respectively, and $\Delta A'_f A$ is the set of assertion differences computed by the algorithm in Figure 2.

We leverage the conservative approximation of impacted behaviors computed by DiSE in the checks in Figure 3. At lines 4 and 5 in Figure 3 the change and impact sets ($C_{p_f}$ and $I_{p_f}$) are computed from the assertion-free program versions $p_f$ and $p'_f$. This generates the set of program statements in the code that may be impacted by changes to only the code and not to the assertions. The Impact Analysis computes the set of program statements within the program slice using the change set as a slicing criterion. A forward as well as a backward slice is computed in order to account for all program statements that may impact the program behavior. Control and data dependence information is used by the impact analysis to construct the program slice. This is a conservative analysis—any program statement that may be dependent on a program statement in the change set is included in the impact set. The forward slice computes the flow of impact from the change set to the other statements while the backward slice accounts for control and data flow from program statements to those in the change set.

A violation of an assertion during Check I indicates that a new feature addition or modification does not conform to the new properties. In Check I, we check the assertion differences with respect to the changed behaviors in $p'_f$, shown at line 8 in Figure 3. A safe approximation of the impacted behaviors is generated by the symbolic execution performed by DiSE (the `DirectedSymbolicExecution` procedure). Symbolic execution is directed along the program statements in

```java
1: Input : Assertion free $p \rightarrow p_f$; Assertions $A$
2: $\text{Assumption free } p \rightarrow p'_f$; Assertions $A'$
3: 
4: $C_{p_f} := \text{SyntacticDiff}(p_f, p'_f)$
5: $I_{p_f} := \text{ImpactAnalysis}(C_{p_f})$
6: 
7: $/$ *Check I: Impacted behaviors & assertion diffs*/
8: $\text{DirectedSymbolicExecution}(p'_f \oplus \Delta A'_f A, I_{p_f})$
9: 
10: $/$ *Check II: Impacted behaviors & original assertions*/
11: $\text{OldAssertions} := \emptyset$
12: for each $a \in A \land \text{matchedAsserts}(a) \neq \emptyset$ do
13:   $\text{OldAssertions} := \text{OldAssertions} \cup \{a\}$
14: $\text{DirectedSymbolicExecution}(p'_f \oplus \text{OldAssertions}, I_{p_f})$
15: 
16: $/$ *Check III: Unchanged behaviors & assertion diffs*/
17: $U_{p_f} := \text{Statements}(p'_f) \setminus I_{p_f}$
18: $\text{DirectedSymbolicExecution}(p'_f \oplus \Delta A'_f A, U_{p_f})$
```

Figure 3: Three cases for incremental property checking.
the impact set \( I_{P'} \) to generate the set of impacted behaviors in \( P' \). The assertion differences, \( \Delta A',A \), which have been rewritten as assertions in \( P' \) are checked on-the-fly as the search is directed along the impacted program executions by DirectedSymbolicExecution.

A violation of an assertion during Check II indicates a regression error. In Check II, we analyze the original (preserved) assertions with respect to the changed behaviors at lines 10 through 13. The set \( \text{OldAssertions} \) is initialized as empty at line 10. We add to this set all assertions in \( A \) that have a matched assertion in \( P' \). This check ensures that assertions removed in \( A' \) are not added to the \( \text{OldAssertions} \) set. We check the assertions in the \( \text{OldAssertions} \) set with respect to the changed behaviors in \( P' \). Check II serves as a regression metric where the analysis ensures that changes to the code do not violate the preserved properties.

A violation of an assertion during Check III indicates a potential incorrect update to the property. In Check III, we analyze the assertion differences with respect to the behaviors that are unchanged between \( P \) and \( P' \). In order to explore the preserved behaviors we generate the set of program statements that are not impacted by the code changes. At line 17 in Figure 3 the set, \( U_{P'} \), contains any program statement that is not in the impact set \( (I_{P'}) \). Now we direct symbolic execution along program statements in \( U_{P'} \) to generate the set of unchanged behaviors. We check the assertion differences with respect to these unchanged behaviors.

4. EVALUATION

We empirically evaluate the effectiveness of our technique for incremental property differencing and checking. Our evaluation addresses the following research questions:

- RQ1: How does the efficiency of checking assertion differences compare with checking the full assertions using symbolic execution?
- RQ2: How does the cost of checking an implication between an old assertion and a clause in a new assertion compare with checking the clause against the program?
- RQ3: How does the performance of the incremental property checking, in the presence of code changes, compare with full symbolic execution in detecting (a) regression errors and (b) conformance of new features to additional properties?

For the first two research questions, we restrict the program differences to assertion changes only; for the third research question we analyze programs with changes to both the code and the assertions.

4.1 Tool Support

We implement our approach using Symbolic PathFinder (SPF) \([12,13]\), the symbolic execution extension to the Java PathFinder (JPF) model checker \([57]\) for analyzing Java programs. SPF uses lazy initialization \([53]\) to handle symbolic inputs of complex (user-defined) types. We use CVC3 \([8]\) as the backend solver for all subjects except Apollo which contains non-linear constraints. We use the CORAL \([53]\) solver for Apollo. We use the DiSE extension to JPF to perform incremental property checking \([44,50]\).

Checking assertions with SPF: When checking assertions with SPF, we use an interactive approach, stopping at the first violation of an assertion. We then remove the failing assertion, and recheck the program until the next assertion fails or SPF completes, e.g., reaches a depth bound. SPF supports finding multiple assertion violations in a single invocation, however, this approach would also find all possible ways of violating each assertion and result in unnecessary overhead, particularly when our goal is simply to determine for each assertion, if there exists any feasible execution that violates it.

4.1.1 Property Differencing: Checking Implications

To check if the property \( \Phi_a \) in assertion \( a \) implies a clause, \( \phi' \), (RQ2) we (1) create a symbolic driver method with body “if (\( \Phi_a \)) assert \( \phi' \)” where \( \Phi_a \) is the conjunction of clauses in \( a \), and the formal parameters are based on the free variables of \( \Phi_a \) and \( \phi' \); and (2) symbolically execute the driver. If symbolic execution does not lead to an assertion violation, the implication holds; otherwise, it may not hold. Conceptually, symbolic execution of the driver checks the formula \( \forall \psi \in \Phi \exists \forall \phi \phi' \Rightarrow \phi \) where \( \psi = \psi_0 \ldots \psi_k \) are the free variables in \( \Phi_a \) and \( \phi' \). If this formula holds, then the clauses in \( a \) imply \( c' \) since the formula checks the implication for all possible values for \( \psi_0 \ldots \psi_k \) and only a subset of these values can reach the control point of \( a \).

If \( \Phi_a \) or \( \phi' \) invoke other method(s), symbolic execution dynamically traces through the bodies of the invoked methods. However, the cost of implication checking then depends on the number of (bounded) paths in the methods invoked by assertions. To optimize implication checking, we over-approximate the outcome of the calls to boolean methods by introducing non-deterministic boolean variables to conservatively model the results of method invocation \([6]\). Specifically, (1) for each unique boolean method invocation \( m(...) \) (based on method name and arguments) in “if (\( \Phi_a \)) assert \( \phi' \)” declare a fresh boolean variable and initialize it using a non-deterministic choice (i.e., \( \text{Verify.getBoolean()} \) in SPF); (2) insert these variable declarations before the implication check; and (3) replace all occurrences of boolean method invocations in the implication check by the corresponding boolean variables. Thus, we conservatively capture the behavior of the boolean methods invoked: if no assertion violation is found in the generated program, then the original implication holds. Recall we assume the assertions are without side-effects.

4.2 Artifacts

In our evaluation, we use four artifacts with manually created assertions and three artifacts with dynamically inferred invariants that we transform into Java assert statements.

4.2.1 Triangle Classification Artifact

The artifacts with manually created assertions include a Java version of the classic triangle classification program by Ammann and Offutt \([2]\) annotated with assertions. The triangle classification program \texttt{trityp} takes as inputs three numbers and outputs whether they represent sides of a valid \texttt{scalene}, \texttt{isosceles}, or \texttt{equilateral} triangle, or an \texttt{invalid} triangle. The classification logic of the \texttt{trityp} program is implemented by the \texttt{Triang} method using 53 lines of Java code\(^1\) which seem deceptively simple, but are non-trivial to reason about. The assertions for this artifact were developed following an evolutionary process.

\(^1\)In the textbook by Ammann and Offutt (Figures 3.2–3.3, Lines 29–81, \([2]\))
## Assertions for trityp

We consider three versions of assertions for trityp (based on the actual experiences of one of the authors). Let $A$, $B$, and $C$ be the three input numbers and $T$ represent the output, where $T = 1$ means “scalene”, $T = 2$ means “isosceles”, and $T = 4$ means “equalateral”, and $T = 4$ means “invalid”.

**Version V1.** The initial set of the trityp assertions includes the following assertions:

- `assert A != B && A != C && B != C && A + B + C > 0 ; T = 4 ; true ;`
- `assert B == C && A < B + C && B > 0 ; T = 2 ; true ;`
- `assert A == B && C < A + B ; T = 2 ; true ;`
- `assert A + B <= C ? T = 4 : true ;`
- `assert A + C <= B ? T = 4 : true ;`
- `assert A + B <= C ? T = 4 : true ;`
- `assert A <= 0 || B <= 0 || C <= 0 ? T = 4 : true ;`
- `assert B + C <= A ? T = 4 : true ;`
- `assert A + C <= B ? T = 4 : true ;`
- `assert A + B <= C ? T = 4 : true ;`

**Version V2.** Analysis of the trityp example revealed that assertions A2–A8 are invalid in the initial version and are therefore edited to:

- `assert A + B <= C ? T = 4 : true ;`
- `assert A + C <= B ? T = 4 : true ;`
- `assert B + C <= A ? T = 4 : true ;`
- `assert A == B && C < A + B ; T = 2 ; true ;`
- `assert A + B <= C && A + B + C > A ? T = 1 : true ;`

**Version V3.** After checking version V2, assertions A6'–A8' in V2 are still invalid and edited to:

- `assert B + C <= A ? T = 2 ; true ;`
- `assert B <= C || A + B + C > 0 ; T = 2 ; true ;`
- `assert A + B + C > A && A + B + C < A + B && C > 0 ; T = 2 ; true ;`
- `assert A + C <= B && A + C < B + C && A > 0 ; T = 2 ; true ;`
- `assert B + C <= A && B + C < A + B && C > 0 ; T = 2 ; true ;`
- `assert A + C <= B && A + C < B + C && A > 0 ; T = 2 ; true ;`

V2 of the assertions thus includes the following 10 assertions in order: A0, A1, A2'–A8', A9.

**Version V3.** After checking version V2, assertions A6'–A8' in V2 are still invalid and edited to:

- `assert A + B <= C ? T = 4 : true ;`
- `assert A + C <= B ? T = 4 : true ;`
- `assert B + C <= A ? T = 4 : true ;`
- `assert A == B && C < A + B ; T = 2 ; true ;`
- `assert A + B <= C && A + B + C > A ? T = 1 : true ;`

V3 of the assertions thus includes the following ten assertions in order: A0, A1, A2'–A8', A6''–A8'', A9.

### 4.2.2 Data Structure Artifacts

Three artifacts with manually created assertions are data structure subjects with assertions based on structural representation invariants. Some of the data structure invariants were developed in previous work [13]; we adapt and augment them in this work to create multiple assertion versions.

**Assertions for Data Structure Artifacts.** Table 1 summarizes the data structure artifacts. For each artifact, we create a symbolic execution driver, which is common across all versions: “$T_0 = new T() ; o.add(x_1) ; ... ; o.add(x_k)$” for symbolic variables $x_1, ..., x_k$ and data structure type $T$.

Next, we create $k$ versions, where $k$ is the number of invariants for the subject. Let the invariants be $I_1, ..., I_k$ and versions be $V_1, ..., V_k$. Each version consists of the symbolic execution driver followed by a sequence of `assert` statements, where $V_k$ contains “`assert I_k`, and version $V_i$ ($1 < i <= k$) contains “`assert I_{i-1}; ...; assert I_1`”. Thus, each new version adds the corresponding new assertion.

### 4.3 Results and Analysis

#### 4.3.1 RQ1: Checking Assertion Differences

We check the trityp program with respect to each of the three assertion versions comparing checking assertion differences with checking the full assertions using symbolic execution. The results are shown in Table 2. For each artifact and its versions, we show the results for checking full assertions (CheckFull) and the results for checking assertions differences that are computed by our property differing algorithm (CheckDiff). #SMT is the number of SMT calls made. #States is the number of states explored during symbolic execution, $T[s]$ is the wall clock time in seconds for symbolic execution, and Red. is the reduction in SMT calls achieved by checking only the assertion differences. The #Conc. column indicates the number of concrete executions used to check the counterexamples from the previous version.

The results in Table 2 show the cumulative results for all invocations of SPF. For the trityp artifact, our property differing algorithm reduces the number of assertions that need to be checked by symbolic execution; consequently it reduces the SMT calls by 95.8% for V2 and 73.0% for V3.

---

### Table 1: Data structure subjects.

<table>
<thead>
<tr>
<th>Subject</th>
<th>#Versions</th>
<th>Invariants</th>
</tr>
</thead>
<tbody>
<tr>
<td>SortedList</td>
<td>5</td>
<td>Acyclicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First element is smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Last element is largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List is sorted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List has unique elements (if unique inserts)</td>
</tr>
<tr>
<td>HeapArray</td>
<td>3</td>
<td>Root is largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Max-heap property</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse-sorted array is max-heap</td>
</tr>
<tr>
<td>BST</td>
<td>4</td>
<td>Acyclicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Left-most leaf is smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Right-most leaf is largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Binary search constraints</td>
</tr>
</tbody>
</table>

### Table 2: Results for checking the tritype assertions.

<table>
<thead>
<tr>
<th>Version</th>
<th>CheckFull</th>
<th>CheckDiff</th>
<th>#SMT</th>
<th>#States</th>
<th>T[s]</th>
<th>Red.</th>
<th>#States</th>
<th>T[s]</th>
<th>Red.</th>
<th>#Conc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>812</td>
<td>821</td>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>V2</td>
<td>763</td>
<td>767</td>
<td>16</td>
<td>32</td>
<td>95.8%</td>
<td>33</td>
<td>1</td>
<td>6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>V3</td>
<td>786</td>
<td>787</td>
<td>13</td>
<td>212</td>
<td>73.0%</td>
<td>213</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 3: Subjects with mechanically synthesized assertions using Daikon.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Method</th>
<th>Daikon application</th>
<th># Invariants</th>
<th># Versions</th>
<th>Version generation</th>
<th># Invariants in each version</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>update</td>
<td>1000</td>
<td>4</td>
<td>V1: 1, V2: 2, V3: 3, V4: 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS</td>
<td>alt_sep_test</td>
<td>1438</td>
<td>35</td>
<td>V1: 10, V2: 20, V3: 30, V4: 40, V5: 53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non_Crossing_Biased_Climb</td>
<td>1439</td>
<td>55</td>
<td>V1: 15, V2: 25, V3: 35, V4: 45, V5: 56</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non_Crossing_Biased_Bascend</td>
<td>1438</td>
<td>56</td>
<td>V1: 1, V2: 2, V3: 3, V4: 4, V5: 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apollo</td>
<td>Reaction_Jet_Control0.Main1</td>
<td>1000</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Results for checking data structure assertions.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Version</th>
<th>CheckFull</th>
<th>CheckDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>SortedList</td>
<td>V1</td>
<td>252</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>1438</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>4630</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>8958</td>
<td>209</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>10316</td>
<td>201</td>
</tr>
<tr>
<td>HeapArray</td>
<td>V1</td>
<td>1422</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>3510</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>4292</td>
<td>99</td>
</tr>
<tr>
<td>BST</td>
<td>V1</td>
<td>2162</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>9366</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>16566</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>36256</td>
<td>836</td>
</tr>
</tbody>
</table>

### Table 5: Results for checking WBS, TCAS, and Apollo.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Version</th>
<th>CheckFull</th>
<th>CheckDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>V1</td>
<td>4976</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>96</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>188</td>
<td>1</td>
</tr>
<tr>
<td>TCAS</td>
<td>V1</td>
<td>3182</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>4270</td>
<td>210</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>4594</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>6582</td>
<td>323</td>
</tr>
<tr>
<td>Apollo</td>
<td>V1</td>
<td>1014</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>1537</td>
<td>946</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>1728</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>7330</td>
<td>4100</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>TO</td>
<td>TO&gt;28800</td>
</tr>
</tbody>
</table>

In Table 4, we present the results of checking full assertions and checking assertion differences on the data structure artifacts. Versions V1 and V2 of SortedList and BinarySearchTree (BST) artifacts contain properties that assert acyclicity. These assertions are checked using lazy initialization that concretizes complex data structures and does not require any constraint solving during symbolic execution. Currently our reductions are targeted for programs and properties that do use constraint solving; hence the number of states generated and the number of SMT calls in the CheckFull and CheckDiff columns for version V2 of SortedList and BinarySearchTree (BST) artifacts are the same. Checking assertion differences in other versions requires fewer SMT calls than checking all assertions and costs less in time and states visited. Checking assertion differences in the SortedList example provides up to 93.6% reduction in the number of SMT calls compared to checking full assertions. In the HeapArray artifact, we observe up to 74.2% reduction in the number of SMT calls compared to checking full assertions. In the BinarySearchTree example, checking assertion differences provides up to 43.4% reduction in the number of SMT calls compared to checking full assertions. Overall, experiments with trityp, SortedList, HeapArray, and BinarySearchTree show that checking assertion differences based on the property differing algorithm outperforms checking full assertions in terms of the number of SMT solver calls, number of states explored, and the time taken for symbolic execution.

Table 5 summarizes the artifacts with mechanically synthesized assertions and lists the methods we consider, the number of tests used to run Daikon, the relevant valid invariants reported by Daikon, the number of assertion versions we create using those invariants, and the number of assertions in each version. We create four and five versions of the WBS and Apollo artifacts respectively, using the relevant and valid invariants generated by Daikon. The first version consists of one randomly selected invariant, and each subsequent version contains all the invariants from the previous version and one additional invariant also selected at random. Daikon’s invariant generation is particularly effective for TCAS [19], so we are able to create versions with larger numbers of relevant valid invariants. We create five versions for each chosen method in TCAS, where the first version consists of randomly selected invariants, and each subsequent version contains a number of additional invariants also selected at random.

In all the versions, each assertion contains a single clause that corresponds to one relevant valid invariant. We use information about how the different versions are constructed in order to manually mark the changed and added assertions between consecutive versions. However, a Unix diff could be used to compute the changes, and it only takes a few milliseconds to run Unix diff for any two consecutive versions.

Table 5 presents the results of checking full assertions and checking assertion differences on WBS, TCAS, and Apollo. TCAS1 refers to first TCAS method listed in Table 3, i.e., TCAS.alt_sep_test, TCAS2 refers the second, etc. Checking full assertions in V5 of Apollo times out (TO) after 8 hours, whereas, checking assertion differences based on property differing successfully compiles for all artifacts. Moreover, checking assertion differences outperforms checking full assertions in terms of states visited and time taken for symbolic execution, and provides between 5.3% and 74.2% reduction in SMT calls across the different versions. While these results show the benefit of using iProperty for more efficient checking of these subjects, the underlying cost of symbolic execution still is a considerable factor in the scalability of the overall incremental approach.
Table 6: Implication checking results using symbolic execution. “Exp1” is the first experiment. “Exp2” is the second experiment.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Check implication ( \phi_a \Rightarrow \phi_{a'} ) in code</th>
<th>Check “assert ( \phi_a )” in code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#SMT</td>
<td>#States</td>
</tr>
<tr>
<td>Exp.: WBS</td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>Exp.: WBS</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Exp.: TCAS\textsuperscript{1}</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Exp.: TCAS\textsuperscript{2}</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Exp.: TCAS\textsuperscript{3}</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Exp.: Apollo</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Exp.: Apollo</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7: Results for checking regression errors.

<table>
<thead>
<tr>
<th>Version</th>
<th>SPF</th>
<th>DiSE</th>
<th>#SMT</th>
<th>#States</th>
<th>( T[s] )</th>
<th>SPF</th>
<th>DiSE</th>
<th>#SMT</th>
<th>#States</th>
<th>( T[s] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS\textsuperscript{1}</td>
<td>15</td>
<td>16</td>
<td>1</td>
<td>15</td>
<td>0%</td>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS\textsuperscript{1}</td>
<td>3238</td>
<td>3239</td>
<td>138</td>
<td>3238</td>
<td>0%</td>
<td>3239</td>
<td>163</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS\textsuperscript{2}</td>
<td>2198</td>
<td>2199</td>
<td>155</td>
<td>107</td>
<td>95.1%</td>
<td>108</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS\textsuperscript{3}</td>
<td>7958</td>
<td>7959</td>
<td>506</td>
<td>158</td>
<td>98.0%</td>
<td>159</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apollo\textsuperscript{1}</td>
<td>158</td>
<td>159</td>
<td>456</td>
<td>4</td>
<td>97.5%</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Results for checking conformance between code change and property change.

<table>
<thead>
<tr>
<th>Version</th>
<th>SPF</th>
<th>DiSE</th>
<th>#SMT</th>
<th>#States</th>
<th>( T[s] )</th>
<th>SPF</th>
<th>DiSE</th>
<th>#SMT</th>
<th>#States</th>
<th>( T[s] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS\textsuperscript{1}</td>
<td>15</td>
<td>16</td>
<td>1</td>
<td>15</td>
<td>0%</td>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS\textsuperscript{1}</td>
<td>1878</td>
<td>1879</td>
<td>94</td>
<td>1878</td>
<td>0%</td>
<td>1879</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS\textsuperscript{2}</td>
<td>2198</td>
<td>2199</td>
<td>83</td>
<td>107</td>
<td>95.1%</td>
<td>108</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAS\textsuperscript{3}</td>
<td>7062</td>
<td>7063</td>
<td>418</td>
<td>142</td>
<td>98.0%</td>
<td>143</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apollo\textsuperscript{1}</td>
<td>158</td>
<td>159</td>
<td>254</td>
<td>4</td>
<td>97.5%</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3.2 RQ2: Implication Checking

To address RQ2 we analyzed assertion pairs, \((a, a')\) where assertion \(a\) is in program \(p\) and its matching assertion \(a'\) is in program \(p'\). In our experiments \(\Phi_a\) and \(\Phi_{a'}\) each contain a single clause \(\phi_a\) and \(\phi_{a'}\) respectively. Hence, for each assertion pair we symbolically execute “if \(\phi_a\) assert \(\phi_{a'}\)”. To check if the implication holds as defined in Section 3.2.

We conduct two different experiments, each with five assertion pairs. In the first experiment, we create assertion pairs by selecting an invariant \(\phi_a\) at random from the set of relevant valid invariants. In the second experiment, we create pairs by applying mechanical transformations: (1) for invariant \(\phi_a\) selected at random: \(\phi_a\) contains the ‘>’ comparison operator, we create \(\phi_{a'}\) by mutating ‘>’ to ‘\(\leq\)’ to form assertion pair \(\langle a, a'\rangle\); (2) if \(\phi_a\) contains the ‘<’ comparison operator, we create \(\phi_{a'}\) by mutating ‘<’ to ‘\(\geq\)’ to form assertion pair \(\langle a, a'\rangle\); (3) if \(\phi_a\) contains a numeric expression of the form “\(t \text{ op } r\)” we generate \(\phi_{a'}\) by adding a constant \(c\in\{-100,\ldots,100\}\), selected at random, to both sides of the expression \(\phi_a\) to form assertion pair \(\langle a, a'\rangle\); and (4) if \(\phi_a\) invokes a Daikon library method, we select an invariant \(\phi_{a'}\neq\phi_a\) at random to create assertion pair \(\langle \phi_a \wedge \phi_{a'} \rangle \forall (\phi_a \wedge \neg \phi_{a'})\).

Table 9 shows experimental results for implication checking for the two experiments. All implications for the first experiment were false (as expected since implication is unlikely to hold between two invariants selected at random); all implications for the second experiment were true (as expected given the design of the transformations). Thus, the two experiments together cover both possible outcomes for implication checking. Some invariants, including all of the Apollo invariants, invokes Daikon library methods, e.g., daikon.Quant.monthOf(this.NumeratorTerms_55575_.arg1), and therefore, the resulting implication checking is handled by our implication check optimization which uses non-deterministic choice to model boolean method invocations (Section 3.3). Thus, some cells have 0 SMT calls – SPF state exploration does not require SMT if the implication only consists of boolean methods and boolean operators.

The results show that the number of SMT calls for the implication check \(\phi_a \Rightarrow \phi_{a'}\) is at most 8.3%, 1%, and 0% for WBS, TCAS, and Apollo respectively, of the cost to check “assert \(\phi_{a'}\)” in the code. Overall, the cost of checking an implication between an assertion in \(p\) and a new clause in the assertion \(a'\) is substantially less than the cost of checking the clause against \(p'\).

4.3.3 RQ3: Incremental Assertion Checking

To address RQ3 we analyze programs that contain changes to both code and properties. We compare the performance of incremental property checking with full symbolic execution in (a) detecting regression errors, and (b) checking conformance of new features to additional properties. We compare the version pairs by manually making changes to the code that were annotated with relevant and valid Daikon invariants; we then also make changes to the properties by adding additional Daikon invariants. We choose to make manual changes to the code in order to simulate potential regression errors and non-conformance between code and properties.

We create one version for each subject listed in Table 5.

To compare the performance in finding regression errors we perform DiSE and full symbolic execution as implemented in SPF on the program \(p'\) annotated with the OldAssertions as defined in Check II Figure 4 of Section 3.3. An assertion violation in \(p'\) is indicative of a regression error. Table 6 shows the experimental results for this experiment. DiSE and SPF make the same number of SMT calls for WBS’ and TCAS\textsuperscript{1}. For TCAS\textsuperscript{1}, all the paths in the code are marked impacted by the code changes, thus DiSE explores all the paths; for WBS’, both SPF and DiSE detect an assertion violation. While in other versions, DiSE achieves substantial reduction, for example, there is 97.5% reduction in the number of SMT calls in Apollo\textsuperscript{1}. The same pattern of results is seen in the the number of states explored by symbolic execution in SPF and DiSE and also total wall clock time taken in seconds. Due to the overhead of static analysis performed by DiSE, for TCAS\textsuperscript{1}, DiSE takes a small amount of
additional analysis time compared to symbolic execution in SPF; however, even when reductions are not possible, the overhead is very small.

In order to check conformance between code changes and properties changes, we perform DiSE and full symbolic execution as implemented in SPF on \( P' \) with the annotated property differences \( \Delta_{\mathcal{A}', \mathcal{A}} \) as described in Check I of Figure 3. We compare the cost of performing DiSE with the cost of performing symbolic execution in SPF in each experiment. Table 8 shows the results. In a manner similar to the results observed in Table 7 for some versions, e.g., WBS', and TCAS\(_2\)', DiSE and symbolic execution in SPF explore the same number of states and make the same number of SMT calls. For some versions, however, such TCAS\(_2\)', DiSE makes fewer SMT calls, explores fewer states, and takes less time compared to symbolic execution in SPF.

4.4 Threats to Validity

The primary threats to external validity in our study are the use of SPF and DiSE, the selection of artifacts and properties, and the changes made to create versions of the code and properties. Although our observations may not generalize to other artifacts, properties, and changes, we attempted to mitigate these threats by analyzing multiple artifacts, all of which have been used in previous studies of symbolic execution based techniques, with respect to properties (assertions) that were both manually developed and mechanically synthesized. Further evaluation of our technique on a broader range of program types, property specifications, and change types would address this threat.

The primary threats to internal validity are the potential faults in the implementation of our algorithms and in SPF and DiSE. We controlled for this threat by testing the tools and implementations of the algorithms on examples that could be manually verified. With respect to threats to construct validity, the metrics we selected to evaluate the cost of our algorithms are commonly used to measure the cost of symbolic execution based techniques.

5. RELATED WORK

Incremental program analysis techniques which leverage change impact analysis results \[1, 5, 28, 14, 45, 48, 80\] have been widely studied to help reduce the cost of program analysis by enabling incremental analysis of previously checked programs. Recent work has also focused on techniques which reuse verification results to help reduce the cost of program analysis \[10, 56, 60\] and reuse of reachability information through similarity checking of test goals represented as automata to reduce the cost of test input generation \[9\]. iProperty partitions the analysis of code and property specifications, leveraging the results of a change impact analysis and reusing the verification results from checking the previous version of the program against the property specifications to reduce the cost of property checking.

A number of recent techniques optimize symbolic execution for one program version using compositional analysis \[23\], abstraction \[3\], and partitioning \[54\]. The idea of optimizing symbolic execution in the context of code that undergoes change was introduced by DiSE \[45\], which first uses a static analysis as we describe and utilize in this paper. An alternative optimization of storing results of symbolic execution and re-using them is taken by Green \[56\] and Memoise \[60\]. Since property differencing likely reduces the number of paths to explore using symbolic execution, we expect it to offer a reduction in the exploration space for Memoise and Green, similar to DiSE.

Recent work explores program differencing in the context of assertions \[32, 35, 39, 47\]. Lahiri et al. present differential contract checking \[36\] which uses static analysis to determine possible inputs that will cause a contract to evaluate differently between two versions. Joshi et al. employ assertion checking to limit false alarms during static concurrency analysis of open systems \[32\]. In more recent work, Lahiri et al. present a differential assertion checking technique (DAC) for comparing relative correctness of the program \[35\]: do the two programs versions perform consistent memory accesses?, at a cost that is lower than absolute correctness.

We believe that core algorithms in iProperty can be extended to handle other lightweight specifications as well, e.g., JML \[37\], Eiffel \[39\], and Spec# \[7\]. These specification languages provide developers the ability to specify properties at a component, module, or method-level in a manner very similar to assertions.

There is a large body of the work on property-based slicing and property-aware testing and verification of programs \[12, 15, 18, 21, 25, 26\] as well as on change-impact analysis, e.g., for regression test selection \[27, 41, 49\]. The key difference between iProperty and previous work is our property differencing technique and its synergistic application with change-impact analysis that characterizes impacted behaviors to enable more efficient checking of code conformance to behavioral properties. Property differing as a stand-alone problem has also been addressed by a number of previous projects, e.g., in the context of UML models \[40, 58\] and Object-Z \[55\]; however, previous work on property differing has not considered using the differences to optimize checking of code.

6. CONCLUSION

This paper introduced iProperty, a novel approach to compute differences between properties of related programs in a manner that facilitates more efficient incremental checking of conformance of programs to properties. The key novelty of iProperty is to compute logical differences between old and new versions of properties that undergo change and facilitate two development scenarios that form our focus: (1) writing properties correctly for an existing program, and (2) checking changes to code and properties as they co-evolve. We evaluate iProperty on Java programs with manually written and mechanically synthesized assertions. The experimental results show that iProperty reduces the number of SMT solver calls – a key metric for the cost of symbolic execution – in comparison with the conventional approach. We believe iProperty provides a promising approach to not only reduce the cost of checking properties but also to make it easier to formulate correct properties – manually or mechanically. While our focus in this paper was on symbolic execution, we believe property differing can enable scalable incremental checking using other software analyses, such as model checking and static analysis.

7. ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation (NSF Grant Nos. CCF-0845628, CCF-1310688, and CNS-1239498).
8. REFERENCES


