Nopol: Automatic Repair of Conditional Statement Bugs in Java Programs
Jifeng Xuan, Matias Martinez, Favio Demarco, Maxime Clément, Sebastian Lamelas, Thomas Durieux, Daniel Le Berre, Martin Monperrus

To cite this version:
Jifeng Xuan, Matias Martinez, Favio Demarco, Maxime Clément, Sebastian Lamelas, et al.. Nopol: Automatic Repair of Conditional Statement Bugs in Java Programs. IEEE Transactions on Software Engineering, Institute of Electrical and Electronics Engineers, 2017, 43 (1), pp.34-55. <10.1109/TSE.2016.2560811>. <hal-01285008v2>

HAL Id: hal-01285008
https://hal.archives-ouvertes.fr/hal-01285008v2
Submitted on 2 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Nopol: Automatic Repair of Conditional Statement Bugs in Java Programs

Jifeng Xuan, Member, IEEE, Matias Martinez, Favio DeMarco†, Maxime Clément†, Sebastian Lamelas Marcote†, Thomas Durieux†, Daniel Le Berre, and Martin Monperrus, Member, IEEE,

Abstract—We propose Nopol, an approach to automatic repair of buggy conditional statements (i.e., if-then-else statements). This approach takes a buggy program as well as a test suite as input and generates a patch with a conditional expression as output. The test suite is required to contain passing test cases to model the expected behavior of the program and at least one failing test case that reveals the bug to be repaired. The process of Nopol consists of three major phases. First, Nopol employs angelic fix localization to identify expected values of a condition during the test execution. Second, runtime trace collection is used to collect variables and their actual values, including primitive data types and object-oriented features (e.g., nullness checks), to serve as building blocks for patch generation. Third, Nopol encodes these collected data into an instance of a Satisfiability Modulo Theory (SMT) problem; then a feasible solution to the SMT instance is translated back into a code patch. We evaluate Nopol on 22 real-world bugs (16 bugs with buggy if conditions and 6 bugs with missing preconditions) on two large open-source projects, namely Apache Commons Math and Apache Commons Lang. Empirical analysis on these bugs shows that our approach can effectively fix bugs with buggy if conditions and missing preconditions. We illustrate the capabilities and limitations of Nopol using case studies of real bug fixes.

Index Terms—Automatic repair, patch generation, SMT, fault localization

1 INTRODUCTION

Automatic software repair aims to automatically fix bugs in programs. Different kinds of techniques are proposed for automatic repair, including patch generation [27], [41] and dynamic program state recovery [42], [7].

A family of techniques has been developed around the idea of “test-suite based repair” [27]. The goal of test-suite based repair is to generate a patch that makes passing test cases pass and keeps the other test cases satisfied. Recent test-suite based repair approaches include the work by Le Goues et al. [27], Nguyen et al. [38], Kim et al. [24].

In recent work [32], we have shown that if conditions are among the most error-prone program elements in Java programs. In our dataset, we observed that 12.5% of one-change commits simply update an if condition. This motivates us to study the automatic repair of conditional statements in real-world bugs.

In this paper, we present a novel automatic repair system called Nopol.† This system fixes conditional bugs in object-oriented programs and is evaluated on real bugs from large-scale open-source programs. For instance, Nopol can synthesize a patch that updates a buggy if condition as shown in Fig. 1 or adds a guard precondition as in Fig. 2. Both figures are excerpts of real-world bugs taken from the bug tracking system of Apache Commons Math.

Nopol takes a buggy program as well as a test suite as input and generates a conditional patch as output. This test suite must contain at least one failing test case that embodies the bug to be repaired. Then, Nopol analyzes program statements that are executed by failing test cases to identify the source code locations where a patch may be needed.

For each statement, the process of generating the patch consists of three major phases. First, we detect whether there exists a fix location for a potential patch in this statement with a new and scalable technique called “angelic fix localization” (Section 3.2). For one fix location, this technique reveals angelic values, 1. Nopol is an abbreviation for “no polillas” in Spanish, which literally means “no moth anymore”. 

† F. DeMarco, M. Clément, S. Lamelas Marcote, and T. Durieux have contributed to this work during their internship at INRIA Lille – Nord Europe.
- if \((u * v == 0)\) {
  \+ if \((u == 0 || v == 0)\) {
    return \(\text{Math.abs}(u) + \text{Math.abs}(v)\);
  }
}

Fig. 1. Patch example of Bug CM5: a bug related to a buggy \(\text{IF}\) condition. The original condition with a comparison with == is replaced by a disjunction between two comparisons.

\+ if \((\text{specific} != \text{null})\) {
  sb.append("; "); //sb is a string builder in Java
  \}

Fig. 2. Patch example of Bug PM2: a precondition is added to avoid a null dereference.

which make all failing test cases pass.

Second, NOPOL collects runtime traces from test suite execution through code instrumentation (Section 3.3). These traces consist of snapshots of the program state at all candidate fix locations. The collected trace consists of both primitive data types (e.g., integers and booleans) and object-oriented data (e.g., nullness or object states obtained from method calls).

Third, given the runtime traces, the problem of synthesizing a new conditional expression that matches the angelic values is translated into a Satisfiability Modulo Theory (SMT) problem (Section 3.4). Our encoding extends the technique by Jha et al. [19] by handling rich object-oriented data. We use our own implementation of the encoding together with an off-the-shelf SMT solver (Z3 [12]) to check whether there exists a solution.

If such a solution exists, NOPOL translates it back to source code, i.e., generates a patch. We re-run the whole test suite to validate whether this patch is able to make all test cases pass and indeed repairs the bug under consideration.

To evaluate and analyze our repair approach NOPOL, we collect a dataset of 22 bugs (16 bugs with buggy \(\text{IF}\) conditions and 6 bugs with missing preconditions) from real-world projects. Our result shows that 17 out of 22 bugs can be fixed by NOPOL, including four bugs with manually added test cases. Four case studies are conducted to present the benefits of generating patches via NOPOL and five bugs are employed to explain the limitations.

The main contributions of this paper are as follows.

- The design of a repair approach for fixing conditional statement bugs of the form of buggy \(\text{IF}\) conditions and missing preconditions.
- Two algorithms of angelic fix localization for identifying potential fix locations and expected values.
- An extension of the SMT encoding in [19] for handling nullness and certain method calls of object-oriented programs.
- An evaluation on a dataset of 22 bugs in real-world programs with an average of 25K executable lines of code for each bug.
- A publicly-available system for supporting further replication and research.
- An analysis of the repair results with respect to fault localization.

This paper is an extension of our previous work [14]. This extension adds an evaluation on a real-world bug dataset, four detailed case studies, a discussion of the limitations of our approach, and a detailed analysis on patches.

The remainder of this paper is organized as follows. Section 2 provides the background of test-suite based repair. Section 3 presents our approach for repairing bugs with buggy \(\text{IF}\) conditions and missing preconditions. Section 4 details the evaluation on 22 real-world bugs. Section 5 further analyzes the repair results. Section 6 presents potential issues and Section 7 lists the related work. Section 8 concludes.

2 Background

We present the background on test-suite based repair and the two kinds of bugs targeted in this paper.

2.1 Test-Suite Based Repair

Test-suite based repair consists in repairing programs according to a test suite, which contains both passing test cases as a specification of the expected behavior of the program and at least one failing test case as a specification of the bug to be repaired. Failing test cases can either identify a regression bug or reveal a new bug that has just been discovered. Then, a repair algorithm searches for patches that make all the test cases pass.

The core assumption of test-suite based repair is that the test suite is good enough to thoroughly model the program to repair [36]. This is a case when the development process ensures a very strong programming discipline. For example, most commits of Apache projects (e.g., Apache Commons Lang) contain a test case specifying the change. If a commit is a bug fix, the commit contains a test case that highlights the bug and fails before the fix.

Test-suite based repair, which has been popularized by the work of GenProg by Le Goues et al. [27], has become an actively explored research area [38], [24], [44], [45], [29]. The approach presented in this paper, NOPOL, is also an approach to test-suite based repair. Other kinds of repair methods include repair based on formal models [20] and dynamic repair of the program state [42].
2.2 Buggy IF Condition Bugs

Conditional statements (e.g., if (condition)(...) else (...)), are widely-used in programming languages. Pan et al. [39] show that among seven studied Java projects, up to 18.6% of bug fixes have changed a buggy condition in IF statements. A buggy IF condition is defined as a bug in the condition of an if-then-else statement.

The bug in Fig. 1 is a real example of a buggy IF condition in Apache Commons Math. This bug is a code snippet of a method that calculates the greatest common divisor between two integers. The condition in that method is to check whether either of two parameters u and v is equal to 0. In the buggy version, the developer compares the product of the two integers to zero. However, this may lead to an arithmetic overflow. A safe way to proceed is to compare each parameter to zero. This bug was fixed by NOPOL (see Bug CM5 in Table 2).

2.3 Missing Precondition Bugs

Another class of common bugs related to conditions is the class of missing preconditions. A precondition aims to check the state of certain variables before the execution of a statement. Examples of common preconditions include detecting a null pointer or an invalid index in an array. In software repositories, we can find commits that add preconditions (i.e., which were previously missing).

The bug in Fig. 2 is a missing precondition with the absence of null pointer detection. The buggy version without the precondition throws an exception signaling a null pointer at runtime. NOPOL fixed this bug by adding the precondition (see Bug PM2 in Table 2).

3 OUR APPROACH

This section presents our approach to automatically repairing buggy IF conditions and missing preconditions. Our approach is implemented in a tool called NOPOL that repairs Java code.

3.1 Overview

NOPOL is a repair approach, which is dedicated to buggy IF conditions and missing preconditions. As input, NOPOL requires a test suite which represents the expected program functionality with at least one failing test case that exposes the bug to be fixed. Given a buggy program and its test suite, NOPOL returns the final patch as output. Fig. 1 and Fig. 2 are two examples of output patches for buggy IF conditions and missing preconditions by NOPOL, respectively.

How to use NOPOL. From a user perspective, given a buggy program with a test suite, including failing test cases, the user would run NOPOL and obtain a patch, if any. Before applying NOPOL to the buggy program, the user does not need to know whether the bug relates to conditions. Instead, the user runs NOPOL for any buggy program. If NOPOL finds a patch, then the user would manually inspect and validate it before the integration in the code base. As further discussion in Section 4.4, the user can also add a pre-defined timeout, e.g., 90 seconds as suggested in experiments or a longer timeout like five hours instead of exhaustively exploring the search space.

Fig. 3 shows the overview of NOPOL. NOPOL employs a fault localization technique to rank statements according to their suspiciousness of containing bugs. For each statement in the ranking, NOPOL considers it as a buggy IF condition candidate if the statement is an IF statement; or NOPOL considers it as a missing precondition candidate if the statement is any other non-branch or non-loop statement (e.g., an assignment or a method call). NOPOL processes candidate statements one by one with three major phases.

First, in the phase of angelic fix localization, NOPOL arbitrarily tunes a conditional value (true or false) of an IF statement to pass a failing test case. If such a conditional value exists, the statement is identified as a fix location and the arbitrary value is viewed as the expected behavior of the patch. In NOPOL, there are two kinds of fix locations, IF statements for repairing buggy conditions and arbitrary statements for repairing missing preconditions.

Second, in the phase of runtime trace collection, NOPOL runs the whole test suite in order to collect the execution context of each fix location. The context includes both variables of primitive types (e.g., booleans or integers) and a subset of object-oriented features (nullness and certain method calls); then such runtime collection will be used in synthesizing the patch in the next phase.

Third, in the phase of patch synthesis, the collected
trace is converted into a Satisfiability Modulo Theory (SMT) formula. The satisfiability of SMT implies that there exists a program expression that preserves the program behavior and fixes the considered bug. That is, the expression makes all test cases pass. If the SMT formula is satisfiable, the solution to the SMT is translated as a source code patch; if unsatisfiable, NOPOL goes to the next statement in the statement ranking, until all statements are processed.

After the above three phases, the whole test suite is re-executed to validate that the patch is correct. This validation could be skipped if the SMT encoding is proven to be correct. Indeed, theoretically, if the SMT solver says “satisfiable”, it means that a patch exists. However, there could be an implementation bug in the off-the-shelf SMT solver, or in the patch synthesis. Consequently, we do make the final validation by re-executing the test suite.

3.2 Angelic Fix Localization

In NOPOL, we propose to use value replacement [18] to detect potential fix locations. Value replacement [18] comes from fault localization research. It consists in replacing at runtime one value by another one. More generally, the idea is to artificially change the program state for localizing faults. There are a couple of papers that explore this idea. For instance, Zhang et al. [55] use the term “predicate switching” and Chandra et al. [9] use the term “angelic debugging”.

NOPOL replaces conditional values in IF statements. We refer to conditional values that make test cases pass as angelic values.

Definition (Angelic Value) An angelic value is an arbitrarily-set value at a given location during test execution, which enables a failing test case to pass.

To facilitate the description of our approach, we follow existing work [17] to introduce the concept of locations. A location is an integer value, which identifies the absolute position of a statement in source code.

Definition (Angelic Tuple) An angelic tuple is a triplet \((loc, val, test)\), where the statement at a location \(loc\) is evaluated to a value \(val\) to make a failing test case \(test\) pass.

In this paper, we refer to the technique of modifying the program state to find the values for angelic tuples \((loc, val, test)\) as angelic fix localization. If an angelic tuple \((loc, val, test)\) is found, there may exist a patch in the location \(loc\) in source code. In the phase of angelic fix localization, only failing test cases are needed, not the whole test suite.

A single test case \(test\) may evaluate the statement at the location \(loc\) several times. Consequently, according to our definition, the value \(val\) is fixed across all evaluations of a given statement for one test case. This is the key point for having a tractable search space (will be discussed in Section 3.2.3). On the other hand, one angelic value is specific to a test case: for a given location \(loc\), different failing test cases may have different angelic values.

3.2.1 For Buggy IF Conditions

For buggy IF conditions, angelic fix localization works as follows. For each IF condition that is evaluated during test suite execution, an angel forces the IF condition to be true or false in a failing test case. An angelic tuple \((loc, val, test)\), i.e., (IF condition location, boolean value, failing test case), indicates that a fix modifying this IF condition may exist (if the subsequent phase of patch synthesis succeeds, see Section 3.4).

Algorithm 1 is the pseudo-code of angelic fix localization for buggy IF conditions.
Input: 
stmt, a candidate non-IF statement; 
T_f, a set of failing test cases.

Output: 
R, a set of angelic tuples.

1 \( R \leftarrow \emptyset; \)
2 Initialize a test case set \( T_{\text{pre}} \leftarrow \emptyset; \)
3 Let \( \text{loc} \) be the location of a potential precondition of \( \text{stmt}; \)
4 Force \( \text{stmt} \) to be skipped and execute \( T_f; \)
5 foreach failing test case \( t_i \in T_f \) do
6     if \( t_i \) passes then
7         \( T_{\text{pre}} \leftarrow T_{\text{pre}} \cup \{ t_i \}; \)
8     end
9 end 
// All test cases in \( T_f \) are passed
10 if \( T_{\text{pre}} = T_f \) then
11     foreach \( t_i \in T_{\text{pre}} \) do
12         \( R \leftarrow R \cup \{(\text{loc}, \text{false}, t_i)\}; \)
13     end
14 end

Algorithm 2: Angelic Fix Localization Algorithm for Missing Preconditions

Lines 17 to 22, for further patch synthesis; otherwise, there exists no angelic value for the test case and the location under consideration. The same idea of forcing the execution can be used to identify angelic values for loop conditions [30].

3.2.2 For Missing Preconditions

Angelic fix localization for missing preconditions is slightly different from that for IF conditions. For each non-branch and non-loop statement that is evaluated during test suite execution, an angel forces to skip it. If a failing test case now passes, it means that a potential fix location has been found. The oracle for repair is then "false"; that is, the added precondition must be false, i.e., the statement should be skipped. Then, an angelic tuple \((\text{loc}, \text{val}, \text{test})\) is (precondition location, false, failing test case).

Algorithm 2 is the pseudo-code of this algorithm. Given a non-IF statement \( \text{stmt} \), we skip this statement to check whether all failing test cases are passed, i.e., Lines 4 to 9. If yes, the location \( \text{loc} \) of the precondition as well as its angelic value false is collected, i.e., Lines 11 to 13. If skipping the statement does not pass all the failing test cases, no angelic values will be returned. This technique also works for missing preconditions for entire blocks since blocks are just specific statements in Java. In our implementation, we only consider adding missing preconditions for single statements rather than blocks. Manual examination on the dataset in Section 4.4 will show that our dataset does not contain missing preconditions for blocks.

3.2.3 Characterization of the Search Space

We now characterize the search space of angelic values. If an IF condition is executed more than once in a failing test case, there may exist a sequence of multiple different angelic values resulting in a passing test case. For example, a buggy IF condition that is executed three times by one failing test case may require a sequence of three different angelic values to pass the test case.

Search space for buggy IF conditions. In general, if one failing test case executes a buggy IF condition for \( t_c \) times, the search space of all sequences of angelic values is \( 2^{t_c} \). To avoid the problem of combinatorial explosion, NOPOL assumes that, for a given failing test case, the angelic value is the same during the multiple executions on one statement. The search space size becomes 2 for one failing test case instead of \( 2^{t_c} \). Under this assumption, the search space is shown as follows.

For buggy IF condition, the search space is \( 2 \times n_c \) where \( n_c \) is the number of executed IF statements by a given failing test case.

Search space for missing preconditions. Similarly to angelic fix localization for buggy IF conditions, if a statement is executed several times by the same failing test case, angelic fix localization directly adds a precondition (with a false value) and completely skips the statement for a given test case.

For missing precondition bugs, the search space size is \( n_p \) where \( n_p \) is the number of executed statements by test cases. It is not \( 2 \times n_p \) because we only add a precondition and check whether the false value passes the failing test case.

NOPOL does not decide a priority between updating existing conditions or adding new preconditions. A user can try either strategy, or both. There is no analytical reason to prefer one or the other; our evaluation does not give a definitive answer to this question. In our experiment, we perform both strategies for statements one by one (see Section 3.1).

If no angelic tuple is found for a given location, there are two potential reasons. First, it is impossible to fix the bug by changing the particular condition (resp. adding a precondition before the statement). Second, only a sequence of different angelic values, rather than a single angelic value, would enable the failing test case to pass. Hence, NOPOL is incomplete: there might be a way to fix an IF condition by alternating the way of finding angelic values, but we have not considered it in this paper.

3.3 Runtime Trace Collection for Repair

Once an angelic tuple is found, NOPOL collects the values that are accessible at this location during program execution. Those values are used to synthesize a correct patch (in Section 3.4). In our work, different kinds of data are collected to generate a patch.
3.3.1 Expected Outcome Data Collection

As mentioned in Section 3.2, an angelic value indicates that this value enables a failing test case to pass. To generate a patch, NOPOL collects the expected outcomes of conditional values to pass the whole test suite: angelic values for failing test cases as well as actual execution values for the passing ones.

Let $O$ be a set of expected outcomes in order to pass all test cases. An expected outcome $O_{loc,m,n} \in O$ refers to the value at location $loc$ during the $m$-th execution in order to pass the $n$-th test case. NOPOL collects $O_{loc,m,n}$ for all executions of location $loc$.

For buggy if conditions, $O_{loc,m,n}$ is the expected outcome of the condition expression at $loc$. For a failing test case, the expected outcome is the angelic value; for a passing test case, the expected outcome is the runtime value $eval(loc)$, i.e., the result of the evaluation during the actual execution of the if condition expression.

$$O_{loc,m,n} = \begin{cases} eval(loc), & \text{for passing test cases} \\ \text{angelic value} & \text{for failing test cases} \end{cases}$$

For missing preconditions, $O_{loc,m,n}$ is the expected value of the precondition at $loc$, i.e., $true$ for passing test cases and $false$ for failing test cases. The latter comes from angelic fix localization: if the precondition returns $false$ for a failing test case, the buggy statement is skipped and the test case passes.

$$O_{loc,m,n} = \begin{cases} true & \text{for passing test cases} \\ false & \text{for failing test cases} \end{cases}$$

Note that not all the bugs with missing preconditions can be fixed with the above definition. Section 4.6.3 will present the limitation of this definition with a real example.

3.3.2 Primitive Type Data Collection

At the location of an angelic tuple, NOPOL collects the values of all local variables, method parameters, and class fields that are typed with a basic primitive type (booleans, integers, floats, and doubles).

Let $C_{loc,m,n}$ be the set of collected values at location $loc$ during the $m$-th execution of the $n$-th test case. In order to synthesize conditions that use literals (e.g., $if (x > 0)$), $C_{loc,m,n}$ is enriched with constants for further patch synthesis. First, NOPOL collects static values that are present in the program. Second, we add three standard values $\{0, -1, 1\}$, which are present in many bug fixes in the wild (for instance for well-known off-by-one errors). Based on these standard values, other values can be formed via wiring building blocks (in Section 3.4.2). For example, a value 2 in a patch can be formed as $1 + 1$, if 2 is not collected during runtime trace collection.

3.3.3 Object-Oriented Data Collection

NOPOL aims to support automatic repair for object-oriented programs. In particular, we would like to support nullness checks and some particular method calls. For instance, NOPOL is able to synthesize the following patch containing a method call.

```java
+ if (obj.size() > 0) {
    compute(obj);
+ }
```

To this end, in addition to collecting all values of primitive types, NOPOL collects two kinds of object-oriented features. First, NOPOL collects the nullness of all variables of the current scope. Second, NOPOL collects the output of “state query methods”, defined as the methods that inspect the state of objects and are side-effect free. A state query method is an argumentless method with a primitive return type. For instance, methods $size()$ and $isEmpty()$ of Collection are state query methods. The concept of state query methods is derived from “argument-less boolean queries” of “object states” by Pei et al. [41].

NOPOL is manually fed with a list of such methods. The list is set with domain-specific knowledge. For instance, in Java, it is easy for developers to identify such side-effect free state query methods on core library classes such as $String, File$ and $Collection$. For each object-oriented class $T$, those predefined state query methods are denoted as $sqm(T)$.

NOPOL collects the nullness of all visible variables and the evaluation of state query methods for all objects in the scope (local variables, method parameters, and fields) of a location where an angelic tuple exists. Note that this incorporates inheritance; the data are collected based on the polymorphism in Java. For instance, when the value of $obj.size()$ is collected, it may be for one implementation of $size()$ based on array lists and for another implementation of $size()$ based on linked lists. This means that a patch synthesized by NOPOL can contain polymorphic calls.

3.3.4 On the Size of Collected Data

Let us assume there are $w$ primitives values and a set $O$ of $w$ objects in the scope of an angelic tuple. In total, NOPOL collects the following values:

- $w$ primitive variables in the scope;
- $w$ boolean values corresponding to the nullness of each object;
- $\sum_{o \in O} \left| sqm(class(o)) \right|$ values corresponding to the evaluation of the state query methods of all objects available in the scope, where $class(o)$ denotes the class of the object $o$;
- constants, i.e., 0, -1, and 1 in our work.

2. Besides collecting static fields in a class, we have also tried to collect other class fields of the class under repair, but the resulting patches are worse in readability than those without collecting class fields. Hence, no class fields other than static fields are involved in data collection.
3.4 Patch Synthesis: Encoding Repair in SMT

The patch synthesis of buggy IF conditions and missing preconditions consists of synthesizing an expression \( \text{exp} \) such that

\[
\forall \text{loc,m,n} \quad \text{exp}(\text{C}_{\text{loc,m,n}}) = \text{O}_{\text{loc,m,n}} \tag{1}
\]

As a synthesis technique, NOPOL, as SemFix [38], uses a variation of oracle-guided component-based program synthesis [19], which is based on SMT.

The solution of the SMT problem is then translated back to a boolean source code expression \( \text{exp} \) representing the corrected IF condition or the added precondition.

3.4.1 Building Blocks

We define a building block (called a component in [19]) as a type of expression that can appear in the boolean expression to be synthesized. For instance, the logical expression to be synthesized. For instance, the logical

\[
\text{building block} \quad \text{IF} \quad \text{expression} \quad \text{condition} \quad \text{exp} \quad \text{in} [19]
\]

We define a building block that includes values of different types: BOOL, INT, and arithmetic operators (+, −, and \( \times \)) and boolean operators (\( \land \), \( \lor \), and \( \neg \)). The same type of building blocks can appear multiple times in one expression.

Let \( b_i \) be the \( i \)-th building block (\( i = 1, 2, \ldots, k \)). Then \( b_i \) is identified as a tuple of a set of input variables \( I_i \), an output variable \( r_i \), and an expression \( \phi_i(I_i, r_i) \) encoding the operation of the building block. That is \( b_i = (\phi_i(I_i, r_i), I_i, r_i) \) (\( b_i = (\phi_i, I_i, r_i) \) for short).

For example, given a boolean output \( r \) and an input \( I = \{I_1, I_2\} \) consisting of two boolean values, a building block could be \( b_i = (\phi_i, \{I_1, I_2\}, r_i) \), where \( \phi_i \) is implemented with the operator \( \land \) i.e., \( I_1 \land I_2 \).

Let \( r \) be the final value of the synthesized patch. Hence there exists one building block \( b_i \) whose output is bound to the return value \( r_i = r \).

Suppose we are given a set \( B \) of building blocks and a list \( \text{CO} \) of pairs \( (\text{C}_{\text{loc,m,n}}, \text{O}_{\text{loc,m,n}}) \), i.e., pairs of collected values and expected values at the location \( \text{loc} \) during the \( m \)-th execution of the \( n \)-th test case.

\( \text{C}_{\text{loc,m,n}} \) includes values of different types: BOOL, INT, or REAL. A patch is a sequence of building blocks \( < b_1, b_2, \ldots, b_k > \) with \( b_i \in B \), whose input values are taken from either \( \text{C}_{\text{loc,m,n}} \) or other building blocks.

3.4.2 Wiring Building Blocks

The problem of patch synthesis is thus to wire the input of building blocks \( < b_1, b_2, \ldots, b_k > \) to the input values \( I_0 \) or the output values of other building blocks. To synthesize a condition, we need to make sure that the types of the variables are valid operands (e.g., an arithmetic operator only manipulates integers).

Example. Let us assume that \( C_{\text{loc,m,n}} \) has three values, an integer variable \( I_0 \), a boolean constant \( c_1 \leftarrow \text{False} \), and an integer constant \( c_2 \leftarrow 3 \). Assume we have two building blocks, \( \text{BOOL} \leftarrow f_1(\text{BOOL}) \) and \( \text{BOOL} \leftarrow f_2(\text{BOOL}) \). Then the goal of patch synthesis is to find a well-formed expression, such as \( \text{False}, f_1(\text{False}), f_2(\text{INT}, \text{INT}), f_1(f_2(3,3)) \); meanwhile, one of these expressions is expected to match the final output \( r \).

3.4.3 Mapping Inputs and Outputs with Location Variables

Let \( I = \bigcup I_i \) and \( O = \bigcup \{r_i\} \) be the sets of input and output values of all building blocks \( b_i \in B \). Let \( I_0 \) be the input set \( \{C_{\text{loc,m,n}}\} \) and let \( r \) be the output in the final patch. We define \( IO \) as \( IO = I \cup O \cup I_0 \cup \{r\} \). We partition the elements of \( IO \) according to their types in BOOL, INT, and REAL.

The SMT encoding relies on the creation of location variables. A location variable \( l_{x} \in L \) represents an index of an element \( x \in IO \). Note that the concept of location variables in SMT encoding is different from the concept of locations in Section 3.2. A location variable \( l_{x} \) indicates a relative position, i.e., an index, of \( x \) in a patch while a location \( l_{x} \) indicates an absolute position of a statement in a source file. A value variable \( v_{x} \in V \) represents a value taken by an element \( x \in IO \). Values of location variables are actually integers (\( L \subseteq \text{INT} \)); value variables are of any supported type, i.e., BOOL, INT, or REAL.

Informally, a location variable \( l_{x} \) serves as an index of \( x \in IO \) in a code fraction while a value variable \( v_{x} \) indicates its value during test execution. Section 3.4.4 will further illustrate how to index the code via location variables. Location variables are invariants for the execution of all test cases: they represent the patch structure. Value variables are used internally by the SMT solver to ensure that the semantics of the program is preserved.

3.4.4 Domain Constraints

Let us first define the domain constraints over the location variables. Given the input set \( I_0 \) and the building block set \( B \), let \( p \) be the number of possible inputs and \( p = |I_0| + |B| \). The location variables of the elements of \( I_0 \) and \( r \) are fixed:

\[
\phi_{\text{FIXED}}(I_0, r) = (\bigwedge_{i=1}^{\text{|I_0|}} l_{0,i} = i) \land l_r = p
\]

Given building blocks \( b_i \in B \), the location variable \( l_{r_i} \) for the output \( r_i \) \( (r_i \in O) \) of \( b_i \) belongs to a range of \( [|I_0| + 1, p] \):

\[
\phi_{\text{OUTPUT}}(O) = \bigwedge_{i=1}^{\text{|O|}} (|I_0| + 1 \leq l_{r_i} \leq p)
\]

Handling types. Only the location variables corresponding to the values of the same type are allowed. Suppose that \( \text{type}(x) \) returns the set of elements with
the same type of $x$ among BOOL, INT, and REAL. Then we can restrict the values taken by the location variables of the input values of building blocks using the following formula:

$$\phi_{INPUT}(I) = \bigwedge_{x \in I} \bigvee_{y \in \text{type}(x), x \neq y} (l_x = l_y)$$

Recall the example in Section 3.4.2, we have the input $I_0 = \{i_0, c_1, c_2\}$, the output $r$, and two building blocks $f_1$ and $f_2$. We assume that each building block is involved in the patch for at most one for simplifying the example; in our implementation, a building block can be used once or more in a synthesized patch. Then we have the location variables as follows. The location variables of $i_0$, $c_1$, and $c_2$ are 1, 2, and 3; the locations variables of building blocks are 4 and 5, respectively. Based on the types, candidate values of location variables of $I_{f_1, 1}$, $I_{f_2, 1}$, and $I_{f_2, 2}$ are calculated:

\begin{align*}
  l_{i_0} & = 1 & \text{input variable, integer} \\
  l_{c_1} & = 2 & \text{boolean constant, False} \\
  l_{c_2} & = 3 & \text{integer constant, 3} \\
  l_{I_{f_1}} & \in [4, 5] & \text{output of } f_1, \text{ boolean} \\
  l_{I_{f_2}} & \in [4, 5] & \text{output of } f_2, \text{ boolean} \\
  l_r & = 5 & \text{expected output value, boolean} \\
  l_{I_{f_1}, 1} & \in \{l_{c_1}, l_{I_{f_1}}, l_{I_{f_2}}\} & \text{the parameter of } f_1, \text{ boolean} \\
  l_{I_{f_2}, 1} & \in \{l_{i_0}, l_{I_{f_2}}\} & \text{the first parameter of } f_2, \text{ integer} \\
  l_{I_{f_2}, 2} & \in \{l_{i_0}, l_{I_{f_2}}\} & \text{the second parameter of } f_2, \text{ integer}
\end{align*}

The following additional constraints are used to control the values of location variables. First, we ensure that each output of a building block is mapped to one distinct input (wires are one-to-one).

$$\phi_{CONS}(L, O) = \bigwedge_{x, y \in O, x \neq y} l_x \neq l_y$$

Second, we need to order the building blocks in such a way that its arguments have already been defined.

$$\phi_{ACYC}(B, L, I, O) = \bigwedge_{(\phi, I_r, r) \in B} \bigwedge_{x \in I} l_x < l_{r_i}$$

Then, we combine all constraints together.

$$\phi_WFF(B, L, I, O, I_0, r) = \phi_{FIXED}(I_0, r) \land \phi_{OUTPUT}(O) \land \phi_{INPUT}(I) \land \phi_{CONS}(L, O) \land \phi_{ACYC}(B, L, I, O)$$

An assignment of $L$ variables respecting the predicate $\phi_{WFF}(B, L, I, O, I_0, r)$ corresponds to a syntactically correct patch.

Value variables corresponding to the input and the output of a building block are related according to the functional definition of a predicate $pb_i$. Given a building block $b_i = \{\phi, i_r, r_i\}$, let $value(I_i)$ be a function that returns the value for the input $I_i$. For a value variable $v_{r_i}$, let $pb_i(value(I_i), v_{r_i}) = \text{true}$ iff \(\phi_i(v_{r_i} = r_i)\). Given $V_{IO} = \{v_x | x \in I \cup O\}$, we define the following constraint.

$$\phi_{LIB}(B, V_{IO}) = \bigwedge_{(\phi, I_r, r_i) \in B, v_r \in V_{IO}} pb_i(value(I_i), v_{r_i})$$

The location variables and the value variables are connected together using the following rule which states that elements at the same position should have the same value. Note that we need to limit the application of that rule to values of the same type because in our case, input or output values can be of different types. Such a limitation to the elements of the same type is valid since the domain of the location variables are managed using constraints $\phi_{INPUT}(I)$.

$$\phi_{CONN}(L, V_{IO}) = \bigwedge_{s \in \{\text{BOOL, INT, REAL}\}} \bigwedge_{x, y \in s} l_x = l_y \Rightarrow v_x = v_y$$

Let the notation $a[v \leftarrow x]$ mean that the variable $v$ in the constraint $a$ has been set to the value $x$. For a given location $loc$, the patch for a given input $I_0$ and a given output $r$ is preserved using the following existentially quantified constraint.

$$\phi_{FUNC}(B, L, C_{loc,m,n}, O_{loc,m,n}) = \exists V_{IO}(\phi_{LIB}(B, V_{IO}) \land \phi_{CONN}(L, V_{IO})[value(I_0) \leftarrow C_{loc,m,n}, v_r \leftarrow O_{loc,m,n}])$$

Finally, finding a patch which satisfies all expected input-output pairs $(C_{loc,m,n}, O_{loc,m,n})$ requires to satisfy the following constraint.

$$\phi_{PATCH}(B, I, O, CO, I_0, r) = \exists L\left(\bigwedge_{(C_{loc,m,n}, O_{loc,m,n}) \in CO} \phi_{FUNC}(B, L, C_{loc,m,n}, O_{loc,m,n}) \land \phi_{WFF}(B, L, I, O, I_0, r)\right)$$

3.4.5 Complexity Levels of Synthesized Expressions

Ideally, we could feed SMT with many instances of all kinds of building blocks (see Section 3.4.1). Only the required building blocks would be wired to the final result. This is an inefficient strategy in practice: some building blocks require expensive computations, e.g., a building block for multiplication (which is a hard problem in SMT).

To overcome this issue, we use the same technique of complexity levels as SemFix [38]. We first try to synthesize an expression with only one instance of easy building blocks ($\leq, \leq, \neq,$ and $=$)\(^5\). Then, we add new building blocks (e.g., building blocks of logical operators and arithmetic operators, successively) and eventually we increase the number of instances of building blocks. We refer to those successive SMT satisfaction trials as the “SMT level”.

\(^5\) $>$ and $\geq$ are obtained by symmetry, e.g., $a \geq b$ as $b \leq a$. 

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 2016 8
Input: 
L, an assignment of location variables, i.e., an SMT solution;
r, a final and expected output variable of patch;
Output:patch, a source code patch.
1. Find a location variable \( l_x = l_r \);
2. \( patch = traverse(l_x) \);

Function \( traverse(l_x) \)
1. if \( x \in O \) then // Output of a building block
   2. Find the expression \( \phi_x(I_x, x) \) and \( I_x = (I_{x,1}, I_{x,2}, ...) \);
   3. return \( code(\phi_x(traverse(I_{x,1}), traverse(I_{x,2}, ...))) \);
4. else if \( x \in I \) then // Input of a building block
   5. Find \( y \) for \( l_y = l_x \); // \( l_y \in O \cup I_0 \)
   6. return \( traverse(l_y) \);
7. else // \( x \in I_0 \), from collected runtime trace
   8. return \( code(x) \);
end

Algorithm 3: Translation Algorithm from an SMT Solution to a Source Code Patch.

3.4.6 Patch Pretty-Printing

NOPOL translates a solution to a patch in source code if there is a feasible solution to the SMT problem. Since NOPOL repairs bugs with buggy IF conditions and missing preconditions, the patch after translation is a conditional expression, which returns a boolean value.

The translation is obtained with a backward traversal starting at the final output location \( l_r \), as explained in Algorithm 3. A function \( traverse \) returns the traversal result according to the location variables while a function \( code \) converts a variable into source code. For example, for a variable \( a \), \( code(a) \) is translated to \( a \); if \( \phi \) denotes the conjunction of boolean values, \( code(\phi(traverse(a), traverse(b))) \) is translated to \( a \land b \).

As shown in Algorithm 3, patch translation from an SMT solution to a patch is a deterministic algorithm, which generates an identical patch for a given SMT solution. Once a patch is translated from the SMT solution, NOPOL returns this patch to developers as the final patch.

Here is a possible solution to the SMT instance for our running example (in Section 3.4.2): \( l_{i_0} = 1, l_{e_1} = 2, l_{r_2} = 3, l_r = 5, l_{r_{f_1}} = 4, l_{r_{f_2}} = 5, l_{I_{f_1,1}} = 2, l_{I_{f_2,1}} = 1, l_{I_{f_2,2}} = 1 \).

In our example, the output is bound to \( l_r = 5 \) that is the output of \( f_2 \). Then \( f_2 \) takes the integer input value \( i_0 \) in \( l_{i_0} \) as a parameter. The final patch is thus the expression \( f_2(i_0, i_0) \) which returns a boolean. This patch could be the repair of a bug, i.e., a fixed IF condition or an added precondition. In this example, \( f_1 \) is never used.

3.5 Fault Localization

NOPOL uses an existing fault localization technique to speed up finding an angelic value, if one exists. In fault localization, statements are ranked according to their suspiciousness. The suspiciousness of a statement measures its likelihood of containing a fault.

In NOPOL, a spectrum-based ranking metric, Ochiai [4], is used as the fault localization technique. Existing empirical studies [47], [52] show that Ochiai is more effective on localizing the root cause of faults in object-oriented programs than other fault localization techniques. In Section 5.2, we will compare the effectiveness among different fault localization techniques.

Given a program and a test suite, the suspiciousness \( susp(s) \) of a statement \( s \) is defined as follows.

\[
susp(s) = \frac{\text{failed}(s)}{\sqrt{\text{total\_failed} + (\text{failed}(s) + \text{passed}(s))}}
\]

where \( \text{total\_failed} \) denotes the number of all the failing test cases and \( \text{failed}(s) \) and \( \text{passed}(s) \) respectively denote the number of failing test cases and the number of passing test cases, which cover the statement \( s \). Note that \( 0 \leq susp(s) \leq 1 \) where \( susp(s) = 1 \) indicates the highest probability of localizing the bug and \( susp(s) = 0 \) indicates there is no likelihood between this statement and the bug.

We rank all the statements based on their suspiciousness in descending order. For all the statements with the suspiciousness over zero, we detect whether this statement is an IF statement or not. As previously mentioned in Section 3.1, for an IF condition, NOPOL tries to synthesize a new condition while for a non-IF statement, NOPOL tries to add a precondition.

4 Automatic Repair of Real-World IF Bugs

We now evaluate our repair approach, NOPOL, on a dataset of 22 real-world bugs. First, we describe our evaluation methodology in Section 4.1; second, we introduce the setup of our dataset in Section 4.2 and the implementation details in Section 4.3; third, we present the general description of the synthesized patches in Section 4.4; fourth, four bugs are employed as case studies in Section 4.5 and five bugs are used to illustrate the limitations in Section 4.6.

4.1 Evaluation Methodology

Our evaluation methodology is based on the following principles.

P1. We evaluate our tool, NOPOL, on real-world buggy programs (Section 4.4).

P2. For bugs that NOPOL can fix, we examine the automatically generated patches, and compare them with human-produced patches (Section 4.5).

P3. For bugs that NOPOL cannot correctly fix, we check the details of these bugs and highlight the
reasons behind the unrepairability (Section 4.6). When
the root cause is an incorrect test case (i.e., an incom-
plete specification), we modify the test case and re-run
Nopol.

P4. We deliberately do not compute a percentage of
repaired bugs because this is a potentially unsound
measure. According to our previous investigation [36],
this measure is sound if and only if 1) the dataset is
only composed of bugs of the same kind and 2) the distribution of complexity within the dataset
reflects the distribution of all in-the-field bugs within
this defect class. In our opinion, the second point is
impossible to achieve.

We have not quantitatively compared our approach
to existing repair approaches on the same dataset
because 1) either existing approaches are inapplicable
on this dataset (e.g., GenProg [27] and SemFix [38] are
designed for C programs); 2) or these approaches are
not publicly available (e.g., PAR [24] and mutation-
based repair [13]).

4.2 Dataset of Real-World Bugs

Nopol focuses on repairing conditional bugs, i.e.,
bugs in IF conditions and preconditions. Hence, we
build a dataset of 22 real-world bugs of buggy IF
conditions and missing preconditions. Since our pro-
totype implementation of Nopol repairs Java code,
these 22 bugs are selected from two open-source Java
projects, Apache Commons Math [6] and Apache Com-
mons Lang [7] (Math and Lang for short, respectively).

Both Math and Lang manage source code using
Apache Subversion [8] (SVN for short) and manage bug
reports using Jira. [9] Jira stores the links between bugs
and related source code commits. In addition, these
projects use the FishEye browser to inspect source
code and commits.

In our work, we employ the following four steps to
collect bugs for the evaluation. First, we automatically
extract small commits that modify or add an IF con-
dition using Abstract Syntax Tree (AST) analysis [15].
We define a small commit as a commit that modifies
at most 5 files, each of which introduces at most 10
AST changes (as computed by the analytical method,
GumTree [15]). In Math, this step results in 161 com-
mits that update IF conditions and 104 commits that
add preconditions; in Lang, the commits are 165 and
91, respectively. The lists of commits are available
at the Nopol project [1]. Second, for each extracted
commit, we collect its related code revision, i.e., the
source program corresponding to this commit. We
manually check changes between the code revision
and its previous one; we only accept changes that
contain an IF condition or a missing precondition
and do not affect other statements. Those commits
could also contain other changes that relate to neither
a bug nor a patch, such as a variable renaming or
the addition of a logging statement. In this case,
changes of the patch are separated from irrelevant
changes. Third, we extract the test suite at the time
of the patch commit, including failing test cases. [11]
Fourth, we manually configure programs and test
suites to examine whether bugs can be reproduced.
Note that the reproducibility rate is very low due to
the complexity of the projects Math and Lang.

Table 1 summarizes the 22 bugs in two categories,
i.e., bug types of buggy IF conditions and missing
preconditions. We index these bugs according to their
types and projects. A bug index (Column 3) is named
based on the following rule. Letters C and P indicate
bugs with buggy IF conditions and missing precondi-
tions, respectively; M and L are bugs from Math and
Lang, respectively. For instance, CM1 refers to a bug
with a buggy IF condition in the project Math. We also
record the number of executable Lines of Code (LoC,
i.e., the number of lines that exclude empty lines and
comment lines) for each source program (Column 6).
Moreover, we show the number of classes, the number
of methods in the buggy program, and the number of
unit test cases (Columns 7-9). For each method that
contains the buggy code, we describe the functionality
of this method and record its Cyclomatic Complexity
(Columns 10 and 11). The Cyclomatic Complexity [34]
is the number of linearly independent paths through
the source code of a method. This complexity indi-
cates the testability of a method and the difficulty of
understanding code by developers.

As shown in Table 1, the dataset contains 16 bugs
with buggy IF conditions and 6 bugs with missing
preconditions. Among these bugs, 12 bugs are from
Math and 10 bugs are from Lang. In average, a buggy
program consists of 25.48K executable lines of code.
The average complexity is 8.6; that is, a buggy method
consists of 8.6 independent paths in average. Note
that the method complexity of Bug PM1 is 1 since
its buggy method contains only one throw statement
(which misses a precondition); the method complexity
of Bug CM9 is 38 and its buggy method contains 30
IF statements.

4.3 Implementation Details

Our approach, Nopol, is implemented with Java 1.7
on top of Spoon 3.1.0. [12] Spoon [40] is a library for
transforming and analyzing Java source code. It is
used for angelic fix localization, instrumentation, and
final patch synthesis in our work. Fault localization

11. In considered commits, bug fixes are always committed to-
gether with originally failing test cases (which are passed after
fixing the bugs). This is a rule in Apache development process [3].
TABLE 1
The Evaluation Dataset of NOPOL. It contains 22 bugs related to buggy IF conditions and missing preconditions.

<table>
<thead>
<tr>
<th>Bug type</th>
<th>Project</th>
<th>Bug description</th>
<th>Executable</th>
<th>#Classes</th>
<th>#Methods</th>
<th>#Unit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>CM1</td>
<td>141003</td>
<td>4611</td>
<td>153</td>
<td>947</td>
<td>363</td>
</tr>
<tr>
<td></td>
<td>CM2</td>
<td>142127</td>
<td>5539</td>
<td>212</td>
<td>1390</td>
<td>543</td>
</tr>
<tr>
<td></td>
<td>CM3</td>
<td>141473</td>
<td>6633</td>
<td>191</td>
<td>1504</td>
<td>654</td>
</tr>
<tr>
<td></td>
<td>CM4</td>
<td>159727</td>
<td>7442</td>
<td>206</td>
<td>1640</td>
<td>704</td>
</tr>
<tr>
<td></td>
<td>CM5</td>
<td>735178</td>
<td>250534</td>
<td>468</td>
<td>3684</td>
<td>1502</td>
</tr>
<tr>
<td></td>
<td>CM6</td>
<td>791766</td>
<td>37918</td>
<td>632</td>
<td>5123</td>
<td>1992</td>
</tr>
<tr>
<td></td>
<td>CM7</td>
<td>835105</td>
<td>38841</td>
<td>650</td>
<td>5275</td>
<td>2105</td>
</tr>
<tr>
<td></td>
<td>CM8</td>
<td>13686250</td>
<td>64709</td>
<td>975</td>
<td>7374</td>
<td>3902</td>
</tr>
<tr>
<td></td>
<td>CM9</td>
<td>1413196</td>
<td>70317</td>
<td>1037</td>
<td>7928</td>
<td>4263</td>
</tr>
<tr>
<td></td>
<td>CM10</td>
<td>1453527</td>
<td>79701</td>
<td>1158</td>
<td>9074</td>
<td>4827</td>
</tr>
<tr>
<td></td>
<td>CL1</td>
<td>137231</td>
<td>10362</td>
<td>156</td>
<td>2132</td>
<td>781</td>
</tr>
<tr>
<td></td>
<td>CL2</td>
<td>137371</td>
<td>11035</td>
<td>169</td>
<td>2240</td>
<td>793</td>
</tr>
<tr>
<td></td>
<td>CL3</td>
<td>137552</td>
<td>12852</td>
<td>173</td>
<td>2579</td>
<td>994</td>
</tr>
<tr>
<td></td>
<td>CL4</td>
<td>230921</td>
<td>15816</td>
<td>215</td>
<td>3516</td>
<td>1437</td>
</tr>
<tr>
<td></td>
<td>CL5</td>
<td>825420</td>
<td>17376</td>
<td>86</td>
<td>3744</td>
<td>1678</td>
</tr>
<tr>
<td></td>
<td>CL6</td>
<td>1075673</td>
<td>18884</td>
<td>211</td>
<td>3918</td>
<td>1934</td>
</tr>
</tbody>
</table>

| Lang     | PM1     | 620221          | 165759     | 396      | 2803     | 1168        |
|          | PM2     | 1035009         | 44347      | 745      | 5536     | 2236        |
|          | PL1     | 304351          | 17286      | 233      | 4056     | 1710        |
|          | PL2     | 643953          | 17790      | 240      | 4285     | 1829        |
|          | PL3     | 655246          | 18533      | 257      | 4443     | 1899        |
|          | PL4     | 1142389         | 18674      | 218      | 3887     | 1877        |

|             | Average | 25480.6        | 399.1      | 3960.4   | 1784.6   | 8.6         |
|             | Median   | 17585.0        | 225.5      | 3818.5   | 1694.0   | 5.5         |

† A commit ID is an identifier that indicates the commit of the patch, in both SVN and the FishEye system. According to this commit, we can manually check relevant patched code and test cases. For instance, the commit of Bug CM1 can be found at [https://fishcottonissues.com/browse/commits/141003].‡ For some bugs, bug IDs are not obviously identified in the bug tracking system. These bugs can be found in the version control system. For example, Apache projects previously used Bugzilla as a bug tracking system before moving to Jira. The Bugzilla system is not available anymore.

is implemented with GZoltar 0.0.10.13 GZoltar [6] is a fault localization library for ranking faulty statements. The SMT solver inside NOPOL is Z3 4.3.2.14 We generate SMT-LIB15 files using jSMTLIB.16 jSMTLIB [10] is a library for checking, manipulating, and translating SMT-LIB formatted problems. The test driver is JUnit 4.11. For future replication of the evaluation, the code of NOPOL is available on GitHub [2].

All experiments are run on a PC with an Intel Core i7 3.60 GHz CPU and a Debian 7.5 operating system. The maximum heap size of Java virtual machine was set to 2.50 GB.

4.4 Main Research Questions

We present the general evaluation of NOPOL on the dataset via answering six Research Questions (RQs).

RQ1: Can NOPOL fix real bugs in large-scale Java software?

In test-suite based repair, a bug is fixed if the patched program passes the whole test suite [27]. Table 2 presents the evaluation of patches on 22 bugs. Column 3 shows the buggy code (the condition for each bug with a buggy IF condition and the statement for each bug with a missing precondition). Column 4 shows the patches that were manually-written by developers as found in the version control system:


the updated condition for each bug with a buggy IF condition and the added precondition for each bug with a missing precondition. Column 5 presents the generated patches by NOPOL. Column 6 is the result of our manual analysis of the correctness of the patches (will be explained in RQ2). Finally, Column 7 shows whether we had to modify existing test cases: “A” stands for additional test cases, “T” for transformed test cases, and “D” for deleted test cases. The purpose of test case modification is to yield a correct repair (will be explained in RQ3).

As shown in Table 2, among 22 bugs, NOPOL can fix 17 bugs: 13 out of 16 bugs with buggy IF conditions and 4 out of 6 bugs with missing preconditions. Meanwhile, four out of five unfixed bugs relate to timeout. In our work, the execution time of NOPOL is limited to up to five hours. We will empirically analyze the fixed bugs in Section 4.5 and explore the limitations of our approach as given by the five unfixed bugs in Section 4.6.

Table 2 also shows that patches generated by NOPOL consist of both primitive values and object-oriented features. For the object-oriented features, two major types can be found in the generated patches: nullness checking (patches of Bugs CL4 and PM2) and the length() method of strings (patches of Bugs CL1, CL2, CL3, and CL5).

Note that six bugs (Bugs CM2, CM3, CM4, CM5, CM6, and CM10) with buggy IF conditions are fixed by adding preconditions rather than updating conditions. One major reason is that a non-IF statement is ranked above the buggy IF statement during the fault localization; then NOPOL adds a patch, i.e.,
RQ2: Are the synthesized patches as correct as the manual patches written by the developer?

In practice, a patch should be more than making the test suite pass since test cases may not be enough for specifying program behaviors [36], [45]. In this paper, a generated patch is correct if and only if the patch is functionally equivalent to the manually-written patch by developers.

For each synthesized patch, we have followed Qi et al. [45] to perform a manual analysis of its correctness. The manual analysis consists of understanding the domain (e.g., the mathematical function under test for a bug in the project Math), understanding the role of the patch in the computation, and understanding the meaning of the test case as well as its assertions.

As shown in Table 2, 13 out of 17 synthesized patches are as correct as the manual patches. Among these 13 correct patches, five patches (for Bugs CM1, CM2, CM3, CM5, and CM7) are generated based on not only the original test suite but also additional test cases. The reason is that the original test suite is too weak to drive the synthesis of a correct patch; then we had to manually write additional test cases (all additional test cases are publicly-available on the companion website [1]17). This will be discussed in next research question.

For four bugs (Bugs CM4, CM6, CL1, and CL5), we are not able to synthesize a correct patch. This will be further discussed in Section 5.4.

RQ3: What is the root cause of test case modification?

As shown in Table 2, some bugs are correctly repaired only after the test case modification (including test case addition, transformation, and deletion). The most important modification is test case addition. Six bugs (Bugs CM1, CM2, CM3, CM4, CM5, and CM7) with additional test cases correspond to too weak specifications. We manually added test cases for these bugs to improve the coverage of buggy statements. Without the additional test case, the synthesized patch is degenerated. A case study of Bug CM1 (Section 4.5.4) will further illustrate how additional test cases help to synthesize patches. Note that all additional test cases appear in the bugs, which are reported in the early stage of the project Math. One possible reason is that the early version of Math is not in test-driven development and the test suites are not well-

-designed.

In test case transformation (Bugs CM2, CM3, CM5, CM6, CM7, CL1, CL2, CL3, CL4, CL5, CL6, and PL3), we simply break existing test cases into smaller ones, in order to have one assertion per test case. This is important for facilitating angelic fix localization since our implementation of NOPOL has a limitation, which collects only one runtime trace for a test case (Section 3.2). We note that such test case transformation can even be automated [51].

The reason behind the four bugs with deleted test cases (Bugs CM6, CM7, CL1, and CL4) is accidental and not directly related to automatic repair: these deleted test cases are no longer compatible with the Java version and external libraries, which are used in NOPOL.

RQ4: How are the bugs in the dataset specified by test cases?

To further understand the repair process of bugs by NOPOL, Tables 3 and 4 show the detailed analysis for patched statements in 17 fixed bugs and buggy statements in five non-fixed bugs, respectively. Table 3 gives the following information: whether a synthesized patch is at the same location as the one from the patch which is manually-written by developers. This failure is caused by an incorrectly identified output of a precondition. Section 4.6.3 will explain the reason behind this failure.

RQ5: Where are the synthesized patches localized? How long is the repair process?

A patch could be localized in a different location from the patch which is manually-written by developers. We present the details for patch locations for all the 17 patched bugs in Table 3. For 11 out of 17 fixed bugs, the locations of patched statements (i.e., locations of fixes) are exactly the same as those of the buggy ones. For the other six bugs, i.e., Bugs CM2, CM3, CM4, CM5, CM6, and CL1, NOPOL generates patches by adding new preconditions rather than updating existing conditions, as mentioned in Table 2.

For 17 fixed bugs in Table 3, the average execution time of repairing one bug is 24.8 seconds while for five non-fixed bugs in Table 4, four bugs are run out of time and the other one spends 37 seconds. The execution time of all the 22 bugs ranges from 6 to 84 seconds. We consider that such execution time, i.e., fixing one bug within 90 seconds, is acceptable.

In practice, if applying NOPOL to a buggy program, we can directly set a timeout, e.g., 90 seconds (over 84 seconds as shown in Table 3) or a longer timeout like five hours in our experiment. Then for any kind of buggy program (without knowing whether the bug is with a buggy condition or a missing precondition), NOPOL will synthesize a patch, if it finds any. Then
a human developer can check whether this patch is correct from the user perspective.

RQ6: How effective is fault localization for the bugs in the dataset?

Fault localization is an important step in our repair approach. As shown in Table 3, for patched statements in 17 fixed bugs, the average fault localization rank is 32.8. In four out of 17 bugs (Bugs CM1, CM2, CM7, and CM10), patched statements are ranked over 50. This fact indicates that there is room for improving the fault localization techniques. Among the five unfixed bugs, the buggy statements of Bugs CM9 and PM1 are ranked over 50. Section 5.2 will further compare the effectiveness of six fault localization techniques on 22 bugs.

Note that in Tables 3 and 4, Bugs CM6 and PM1 have no passing test cases. Bug PM1 cannot be fixed by our approach while Bug CM6 can be still fixed because the two failing test cases give a non-trivial input-output specification. The reason behind the unfixed bugs statement is not $e_p = 0$, but the multiple executions of the buggy code by one test case. This reason will be discussed in Section 4.6.1.

4.5 Case Studies for Fixed Bugs

We conduct four case studies to show how NOPOL fixes bugs with buggy IF conditions and missing preconditions. These bugs are selected because they highlight different facets of the repair process. The patch of Bug PL4 (Section 4.5.1) is syntactically the same as the manually-written one; the patch of Bug CL4 (Section 4.5.2) is as correct as the manually-written patch (beyond passing the test suite); the patch of Bug CM2 (Section 4.5.3) is correct by adding a precondition rather than updating the buggy condition, as written by developers; and the patch of Bug CM1 (Section 4.5.4) is correct, but its patch requires an additional test case.

4.5.1 Case Study 1, Bug PL4

NOPOL can generate the same patches for three out of 22 bugs as the manually-written ones. We take Bug PL4 as an example to show how the same patch is generated. This bug is fixed by adding a precondition. Fig. 4 shows a method translate() at Line 1 and the buggy method translateInner() at Line 9 of Bug PL4. The method translate() is expected to translate a term in the regular expression of "\&\#\{x\}\d+" into codepoints, e.g., translating the term "\&\#x30" into "\u0030".

To convert from input to codepoints, the characters in input are traversed one by one. Note that for a string ending with "\&\#x", no codepoint is returned. Lines 15 to 20 in Fig. 4 implement this functionality. However, the implementation at Lines 17 to 19 ignores a term in a feasible form of "\&\#\{x\}\d+", e.g., a string like "\&\#x30". A precondition should be added to detect this feasible form, i.e., the comment at Line 18 of start == seqEnd.

The buggy code at Line 19 is executed by one passing test case and one failing test case. Table 5 shows these two test cases. For the passing test case, the behavior of the method is expected not to change the variable input while for the failing test case, the input is expected to be converted. In the passing test case, the value of the precondition of the statement at Line 19 is expected to be true, i.e., both start and seqEnd equal to 8, while in the failing test case, the condition is expected to be false, i.e., start and seqEnd are 8 and 19, respectively. The false value is the angelic value for a missing precondition.

According to those expected precondition values for test cases, NOPOL generates a patch via adding a pre-
4.5.2 Case Study 2, Bug CL4

For several bugs, NOPOL generates patches that are literally different from the manually-written patches, but these generated patches are as correct as manually-written patches. In this section, we present a case study where NOPOL synthesizes a correct patch for a bug with a buggy IF condition. Bug CL4 in Lang fails to find the index of a matched string in a string builder. Fig. 5 presents the buggy method of Bug CL4: to return the first index of substr in a parent string builder from a given basic index startIndex. The condition at Line 4 contains a mistake of startIndex >= size, which omits checking whether substr == null. A variable size is defined as the length of the parent string builder. The manually-written fix is shown at Line 3.

The buggy code at Line 4 in Fig. 5 is executed by 23 passing test cases and two failing test cases. One of the passing test cases and two failing test cases are shown in Table 6. For the passing test case, a value -1 is expected because no matched string is found. For the two failing test cases, each input substr is a null value, which is also expected to return a non-found index -1. This requires the checking of null to avoid NullPointerException, i.e., the condition at Line 3.

For the passing test case in Table 6, the condition at Line 4 is false. For the two failing test cases, NOPOL extracts the angelic value true to make both failing test cases pass. According to these condition values, a patch of ! (substr != null) || startIndex >= size can be synthesized. This synthesized patch is equivalent to substr == null || startIndex >= size, which is correct. The resulted expression based on the solution to the SMT does not weaken the repairability of synthesized patches. A recent method for finding simplified patches, proposed by Mechtaev et al. [35], could be used to avoid such a redundant expression.

4.5.3 Case Study 3, Bug CM2

In this section, we present Bug CM2, a correctly patched bug via adding a precondition, rather than updating an existing condition, as written by developers. The buggy method in Bug CM2 is to calculate the value of Binomial Coefficient by choosing k-element subsets from an n-element set. Fig. 6 presents the buggy method. The input number of elements n should be no less than zero. But the condition at Line 4 reads n <= 0 instead of n < 0. The manually-written patch by developers is in the FIX comment at Line 4.

The buggy code at Line 4 in Fig. 6 is executed by two passing test cases and one failing test case.
4.5.4 Case Study 4, Bug CM1

Insufficient test cases lead to trivial patch generation. We present Bug CM1 in Math, with a buggy `if` condition. This bug cannot be correctly patched with the original test suite due to the lack of test cases. In our work, we add two test cases to support the patch generation. Fig. 7 presents the buggy source code in the method `evaluate()` of Bug CM1. This method returns an estimate of the percentile \( p \) of the values stored in the array `values`.

According to the API document, the algorithm of `evaluate()` is implemented as follows. Let \( n \) be the length of the (sorted) array. The algorithm computes the estimated percentile position \( \text{pos} = p \times (n + 1) / 100 \) and the difference \( \text{dif} \) between \( \text{pos} \) and \( \text{floor}(\text{pos}) \). If \( \text{pos} >= n \), then the algorithm returns the largest element in the array; otherwise the algorithm returns the final calculation of percentile. Thus, the condition at Line 15 in Fig. 7 contains a bug, which passes all test cases. This patch is incorrect. To obtain a correct patch (the one shown in Table 2), we add one test case of \( \{1,2,3\} 100 \), as shown in Table 8. Then the expected value of `evaluate()` is 3.0. After running NOPOL, a patch of `length <= fpos`, which is different from the manually-written one, i.e., \( \text{pos} >= n \). However, from the source code at Line 7, \( \text{fpos} = \text{floor}(\text{pos}) \), i.e., \( \text{fpos} \) is the largest integer that no more than \( \text{pos} \). That is, \( \text{fpos} <= \text{pos} \). Meanwhile, \( n == \text{length} \) holds according to Line 4. As a result, the generated patch `length <= fpos` implies the manually-written one, i.e., \( \text{pos} >= n \). We can conclude that NOPOL can generate a correct patch for this bug by adding one test case.
### 4.6 Limitations

As shown in Section 4.4, we have collected five bugs that reveal five different limitations of NOPOL. Table 9 lists these five bugs in details. We analyze the related limitations in this section.

#### 4.6.1 No Angelic Value Found

In our work, for a buggy IF condition, we use angelic fix localization to flip the boolean value of conditions for failing test cases. For Bug PM1, no angelic value is found as shown in Table 2. The reason is that both then and else branches of the IF condition are executed by one failing test case. Hence, no single angelic value (true or false) can enable the test case to pass. As discussed in Section 3.2.3, the search space of a sequence of angelic values is exponential and hence discarded in our implementation of NOPOL.

To mitigate this limitation, a straightforward solution is to discard the failing test case, which leads to no angelic values (keeping the remaining failing ones). However, this may decrease the quality of the generated patch due to the missing test data and oracles. Another potential solution is to refactor test cases into small snippets, each of which covers only then or else branches [51]. A recent proposed technique SPR [29] could help NOPOL to enhance its processing of sequential angelic values.

#### 4.6.2 Performance Bugs Caused by Angelic Values

NOPOL identifies angelic values as the input of patch synthesis. In the process of angelic fix localization, all failing test cases are executed to detect conditional values that make failing test cases pass (see Section 3.2). However, sometimes the trial of angelic fix localization (forcing to true or false) may result in a performance bug. In our work, Bug CM9 cannot be fixed due to this reason, i.e., an infinite loop caused by angelic fix localization.

A potential solution to this issue is to set a maximum execution time to avoid the influence of performance bugs. But a maximum execution time of test cases may be hard to be determined according to different test cases. For instance, the average execution time of test cases in Math 3.0 is much longer than that in Math 2.0. We leave the setting of maximum execution time as one piece of future work.

#### 4.6.3 Incorrectly Identified Output of a Precondition

As mentioned in Section 3.3.1, the expected output of a missing precondition is set to be true for a passing test case and is set to be false for a failing one. The underlying assumption for a passing test case is that the true value keeps the existing program behavior. However, it is possible that given a statement, both true and false values can make a test case pass. In these cases, synthesis may not work for bugs with missing preconditions.

This is what happens to Bug PL3. Fig. 8 shows a code snippet of Bug PL3. The manually-written patch is a precondition at Line 3; Table 10 shows one passing test case and one failing test case. Based on the angelic fix localization in Algorithm 2, the expected precondition values of all passing test cases are set to be true. However, in the manually-written patch, the precondition value by the passing test case in Table 10 is false, i.e., lower > str.length() where lower is 0 and str.length() is 10. Thus, it is impossible to generate a patch like the manually-written one, due to a conflict in the input-output specification. Consequently, in the phase of patch synthesis, the SMT solver executes with timeout.

The example in Bug PL3 implies that for some bugs, the assumption (i.e., a missing precondition is expected to be true for passing test cases) can be violated. For Bug PL3, we have temporarily removed this assumption and only used the failing test cases to synthesize a patch. The resulting patch is if(lower >= str.length()) lower = str.length(), which has the same program behavior as the manually-written patch, i.e., lower > str.length(). In NOPOL, we are conservative and assume that the expected value of a precondition by passing test cases is true (in Section 3.2.2).

#### 4.6.4 Complex Patches using Method Calls with Parameters

In our work, we support the synthesis of conditions that call unary methods (without parameters). However, our approach cannot generate a patch if a method with parameters has to appear in a condition. For example, for Bug CM8, the patch that is written by developers contains a method abs(x) for computing the absolute value. Our approach cannot provide such kinds of patches because methods with parameters

---

**TABLE 9**

<table>
<thead>
<tr>
<th>Bug index</th>
<th>Root cause</th>
<th>Result of repair</th>
<th>Reason for the unfixed bug</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM1</td>
<td>Angelic fix localization</td>
<td>No angelic value found. Termination before runtime trace collection and patch synthesis</td>
<td>One failing test case executes the missing precondition for more than once.</td>
</tr>
<tr>
<td>CM9</td>
<td>Angelic fix localization</td>
<td>Timeout during test suite execution</td>
<td>An infinite loop is introduced during the trial of angelic values.</td>
</tr>
<tr>
<td>PL3</td>
<td>Runtime trace collection</td>
<td>Timeout during SMT solving</td>
<td>The expected value of a precondition is incorrectly identified.</td>
</tr>
<tr>
<td>CM8</td>
<td>Patch synthesis</td>
<td>Timeout during SMT solving</td>
<td>A method call with parameters is not handled by SMT.</td>
</tr>
<tr>
<td>CL6</td>
<td>Patch synthesis</td>
<td>Timeout during SMT solving</td>
<td>A method of a null object yields an undefined value for SMT.</td>
</tr>
</tbody>
</table>
String abbreviate(String str, int lower, int upper) {
    ...  
    // FIX: if (lower > str.length())
    lower = str.length();
    ...  
    if (upper == -1 || upper > str.length())
        upper = str.length();
    if (upper < lower)
        upper = lower;
    StringBuffer result = new StringBuffer();
    int index = String.valueOf(str, "_.", lower);
    if (index == -1)
        result.append(str.substring(0, upper));
    else ...
    return result.toString();
}

Fig. 8. Code snippet of Bug PL3. The manually-written patch is shown in the FIX comment at Line 3.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output abbreviate(str, lower, upper)</th>
<th>Test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0123456789</td>
<td>0123456789</td>
<td>pass</td>
</tr>
<tr>
<td>0123456789 *</td>
<td>0123456789</td>
<td>fail</td>
</tr>
</tbody>
</table>

Table 10 Sample of test cases for Bug PL3

cannot be directly encoded in SMT. Then the lack of information of method calls leads to the timeout of an SMT solver.

A workaround would generate a runtime variable to collect existing side-effect free method calls with all possible parameters. For example, one could introduce a new variable double tempVar = abs(x) and generate a patch with the introduced variable tempVar. However, this workaround suffers from the problem of combinatorial explosion.

4.6.5 Unavailable Method Values for a Null Object

Our repair approach can generate a patch with object-oriented features. For example, a patch can contain state query methods on Java core library classes, such as String.length(), File.exists() and Collection.size(). We map these methods to their return values during the SMT encoding. However, such methods require that the object is not null; otherwise, a null pointer exception in Java is thrown.

Let us consider Bug CL6, whose manually-written patch is cs == null || cs.length() == 0. For this bug, one passing test case detects whether the object cs is null. For this test case, the value of cs.length() is undefined and not given to SMT. Thus, it is impossible to generate a patch, which contains cs.length() if cs is null by at least one test case. Consequently, the SMT solver times-out because it tries to find a complex patch that satisfies the constraints.

A possible solution is to encode the undefined values in the SMT. Constraints should be added to ensure that the unavailable values are not involved in the patch. This needs important changes in the design of the encoding, which is left to future work.

5 DISCUSSIONS

We now discuss NOPOL features with respect to four important aspects.

5.1 Differences with SemFix

As mentioned in Section 3.4, NOPOL uses the same technique in the phase of patch synthesis as SemFix [38], i.e., component-based program synthesis [19]. However, there exist a number of important differences between NOPOL and SemFix.

First, SemFix does not address missing pre-conditions. As shown in Table 2, adding preconditions enables us to repair more bugs than only updating conditions. We think that it is possible to extend the SemFix implementation to support repairing missing preconditions via adding the encoding strategy as in NOPOL.

Second, NOPOL does not use symbolic execution to find an angelic value. It is known that symbolic execution may have difficulties with the size and complexity of analyzed programs [28]. According to our work, we have the following observations. The angelic value in angelic fix localization is possible only when the domain is finite. For booleans, the domain of variables is not only finite but also very small. This results in a search space that can be explored dynamically and exhaustively as in NOPOL. Symbolic execution as done in SemFix is capable of also reasoning on integer variables, because the underlying constraint solver is capable of exploring the integer search space. To sum up, our analytical answer is that for boolean domains, angelic fix localization is possible and probably much faster (this is claimed, but not empirically verified). For integer domains, only symbolic execution is appropriate. Meanwhile, NOPOL can also handle the right-hand side of assignments as in SemFix. If we encode the synthesis as in SemFix, the right-hand side of assignments can be directly processed by NOPOL.

Third, NOPOL supports object-oriented code. We have adapted the code synthesis technique so that the generated patch can contain null checks and method calls.

Finally, the evaluation of NOPOL is significantly larger than that of SemFix. We have run NOPOL on larger programs and real bugs. In SemFix, 4/5 of subject programs have less than 600 LoC and the bugs are artificially seeded. In the evaluation in our paper, the average number of lines of code per subject program is 25K LoC and the bugs are extracted from real programs that happened in practice.
5.2 Effectiveness of Fault Localization Techniques

Fault localization plays an important role during the repair process. In our approach, a fault localization technique ranks all the suspicious statements and NOPOL attempts to generate patches by starting with analyzing the most suspicious statement first. We use Ochiai [4] as the fault localization technique. For our dataset, we wonder whether there is a difference between Ochiai and other techniques. In this section, we study the accuracy of different fault localization techniques on the bugs of our dataset.

We employ the absolute wasted effort to measure fault localization techniques. The wasted effort is defined as the ranking of the actual buggy statement. Given a set \( S \) of statements, the wasted effort is expressed as follows,

\[
effort = |\{\text{susp}(x) > \text{susp}(x^*)\}| + 1
\]

where \( x \in S \) is any statement, \( x^* \) is the actual buggy statement, and \(|\cdot|\) calculates the size of a set. A low value indicates that the fault localization technique is effective.

In our experiment, we compare six well-studied fault localization techniques: Ochiai [4], Tarantula [21], Jaccard [4], Naish2 [37], Ochiai2 [37], and Kulczynski2 [50]. Table 11 presents the comparison on the two types of bugs considered in this paper.

As shown in Table 11, for bugs with buggy IF conditions, Jaccard obtains the best average wasted effort while Jaccard and Naish2 get the same median value. For bugs with missing preconditions, Tarantula obtains the best average wasted effort while Ochiai, Jaccard, and Naish2 get the same median value. Those results assess that, according to our dataset of real bugs, the sensitivity of NOPOL with respect to fault localization is not a crucial point and Ochiai is an acceptable choice.

<table>
<thead>
<tr>
<th>Fault localization technique</th>
<th>Buggy IF condition</th>
<th>Missing precondition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ochiai</td>
<td>Average</td>
<td>Median</td>
</tr>
<tr>
<td>Tarantula</td>
<td>127.63</td>
<td>45.50</td>
</tr>
<tr>
<td>Jaccard</td>
<td>121.44</td>
<td>25.50</td>
</tr>
<tr>
<td>Naish2</td>
<td>135.06</td>
<td>25.50</td>
</tr>
<tr>
<td>Ochiai2</td>
<td>133.06</td>
<td>44.50</td>
</tr>
<tr>
<td>Kulczynski2</td>
<td>127.44</td>
<td>45.50</td>
</tr>
</tbody>
</table>

Second, our synthetic technique is based on different SMT levels (see Section 3.4.5). Synthesis at each SMT level corresponds to one independent SMT instance. Hence, the synthesis can also be run in parallel. However, we should mention that parallelizing the synthesis may lead to multiple resulted patches. Synthesis at a low SMT level can generate a simple patch; for the same bug, synthesis at a higher SMT level may not generate a better patch and may waste the running cost.

5.4 Insight on Test-Suite Based Repair

NOPOL is a test-suite based repair approach, as other existing work ([27], [38], [24], [44], etc.) in the field of automatic software repair. However, the foundations of test-suite based repair are little understood. Our experience with NOPOL enables us to contribute to better understanding the strengths and the weaknesses of test-suite based repair.

There are two grand research questions behind test-suite based repair. The first one is about the quality of test suites [36]: do developers write good-enough test suites for automatic repair? Qi et al. [45] have shown that the test suites considered in the GenProg benchmark are not good enough, in the sense that they accept trivial repairs such as directly removing the faulty code. The experiments we have presented in this paper shed a different light. For nine bugs considered in this experiment, the test suite leads to a correct patch. For four additional bugs, a slight addition in the test suite allows for generating a correct patch. We consider this as encouraging for the future of the field. However, there is a need for future work on recommendation systems that tell when to add additional test cases for the sake of repair, and what those test cases should specify.

The second grand research question goes beyond standard test suites such as JUnit ones and asks whether repair operators do not overfit the inputs of input-output specifications [46]. For instance, for Bug CM6, one of the test inputs always equals to \(-1\) when the buggy code is executed. As a result, the patch simply uses this value (corresponding to the number of rows) to drive the control flow, which is wrong. On the other hand, there are other cases when the repair
operator yields a generic and correct solution upfront. This fact indicates that the same repair operator may overfit or not according to different bugs. It is necessary to conduct future research on the qualification of repair operators according to overfitting.

6 Threats to Validity

We discuss the threats to the validity of our results along four dimensions.

6.1 External Validity

In this work, we evaluate our approach on 22 real-world bugs with buggy IF conditions and missing preconditions. One threat to our work is that the number of bugs is not large enough to represent the actual effectiveness of our technique. While the number of bugs in our work is fewer than that in previous work [27], [38], [24], the main strength of our evaluation is twofold. On one hand, our work focuses on two specific types of bugs, i.e., buggy IF conditions and missing preconditions (as opposed to general types of bugs in [24]); on the other hand, our work is evaluated on real-world bugs in large Java programs (as opposed to bugs in small-scale programs in [38] and bugs without object-oriented features in [27]). We note that it is possible to collect more real-world bugs, with the price of more human labor. As mentioned in Section 4.2, reproducing a specific bug is complex and time-consuming.

6.2 Single Point of Repair

As all previous works in test-suite based repair, the program under repair must be repaired at one single point. In the current implementation of NOPOL, we do not target programs with multiple faults, or bugs which require patches at multiple locations.

6.3 Test Case Modification

In our work, we aim to repair bugs with buggy IF conditions and missing preconditions. Test cases are employed to validate the generated patch. In our experiment, several test cases are modified to facilitate repair. As mentioned in Section 4.4, such test case modification consists of test case addition, transformation, and deletion. The answer to RQ3 analyzes the root causes of test case modification. All test case modifications are listed in our project website [1].

6.4 Dataset Construction

We describe how to construct our dataset in Section 4.2. The manually-written patches of conditional statements are extracted from commits in the version control system. However, it is common that a commit contains more code than the patch in buggy IF conditions and missing preconditions. In our work, we manually separate these patch fragments. In particular, the fixing commit of Bug PM2 contains two nested preconditions within a complex code snippet. We manually separate the patch of this bug according to the code context and keep only one precondition. Hence, there exists a potential bias in the dataset construction.

7 Related Work

We list related work in four categories: approaches to test-suite based repair, repair besides test-suite based repair, empirical foundations of test-suite based repair, and related techniques in NOPOL.

7.1 Test-Suite Based Repair

GenProg. Test-suite based repair generates and validates a patch with a given test suite. Le Goues et al. [27] propose GenProg, an approach to test-suite based repair using genetic programming for C programs. In GenProg, a program is viewed as an Abstract Syntax Tree (AST) while a patch is a newly-generated AST by weighting statements in the program. Based on genetic programming, candidate patches are generated via multiple trials. The role of genetic programming is to obtain new ASTs by copying and replacing nodes in the original AST. A systematic study by Le Goues et al. [25] shows that GenProg can fix 55 out of 105 bugs in C programs. The difference between NOPOL and GenProg are as follows. NOPOL targets a specific defect class while GenProg is generic; NOPOL uses component-based program synthesis while GenProg only copies existing code from the same code base; NOPOL uses a four-phase repair approach (fault localization, angelic fix localization, runtime trace collection, and patch synthesis) while GenProg uses a different two-phase approach (fault localization and trial); NOPOL is designed for object-oriented Java programs while GenProg is for C.

AE. Weimer et al. [49] report an adaptive repair method based on program equivalence, called AE. This method can fix 54 out of the same 105 bugs as in the work [25] while evaluating fewer test cases than GenProg.

PAR. Kim et al. [24] propose PAR, a repair approach using fix patterns representing common ways of fixing bugs in Java. These fix patterns can avoid nonsensical patches, which are caused by the randomness of some operators in genetic programming. Based on the fix patterns, 119 bugs are examined for patch generation. In this work, the evaluation of patches is contributed by 253 human subjects, including 89 students and 164 developers.

RSRepair. Qi et al. [44] design RSRepair, a random search based technique for navigating the search space. This work indicates that random search performs more efficiently than genetic programming in GenProg [27]. RSRepair can fix 24 bugs, which are
derived from a subset of 55 fixed bugs by GenProg [25]. Another work by Qi et al. [43] reduces the time cost of patch generation via test case prioritization.

**SemFix.** Nguyen et al. [38] propose SemFix, a constraint based repair approach. This approach generates patches for assignments and conditions by semantic analysis via SMT encoding. Program components are synthesized into one patch via translating the solution of the SMT instance. Our proposed approach, NOPOL, is motivated by the design of SemFix. The major differences between NOPOL and SemFix were discussed in Section 5.1.

**Mutation-based repair.** Debroy & Wong [13] develop a mutation-based repair method, which is inspired by the concept of mutation testing. Their method integrates program mutants with fault localization to explore the search space of patches.

**DirectFix.** Mechtaev et al. [35] propose DirectFix, a repair method for simplifying patch generation. Potential program components in patches are encoded into a Maximum Satisfiability (MaxSAT) problem, i.e. an optimization problem; the solution to the MaxSAT instance is converted into the final concise patch.

**SearchRepair.** Ke et al. [23] develop SearchRepair, a repair method with semantic code search, which encodes human-written code fragments as SMT constraints on input-output behavior. This method reveals 20% newly repaired defects, comparing with GenProg, AE, or RSRepair.

**SPR.** After the original publication presenting NOPOL [14], Long & Rinard [29] have proposed a repair technique called SPR using condition synthesis. SPR addresses repairing conditional bugs, as well as other types of bugs, like missing non-IF statements. The differences are as follows. First, a major difference between NOPOL and SPR is that NOPOL synthesizes a condition via component-based program synthesis while SPR is based on multiple trials of pre-defined program transformation schemas. For instance, in SPR, a transformation schema for conditional bugs is called condition refinement, which updates an existing condition in an IF via tightening or loosening the condition. To repair a bug, SPR tries a potential patch with the transformation schemas one by one and validates the patch with the test suite; the technique of NOPOL is entirely different, based on runtime data collection during test execution. Second, another difference is that SPR is for repairing C programs. Patches by SPR only contain primitive values while patches by NOPOL contain both primitive values and object-oriented expressions (e.g., fields and unary method calls). Third, in SPR, the technique of collecting angelic values is based on NOPOL’s, yet extends it. It finds sequences of values rather than one simplified trace during collecting angelic values in NOPOL (Section 3.2). As mentioned in Section 3.2.3, the simplified trace in NOPOL reduces the search space of patch synthesis, but may result in failed repair attempts for specific bugs, where a condition is executed more than once by a test case. Examples of these bugs can be found in the SPR evaluation [29]. The simplification in NOPOL can be viewed as a trade-off between repairability and time cost.

**Prophet.** Also by Long & Rinard, Prophet is an extension of SPR that uses a probability model for prioritizing candidate patches. Based on historical patches, Prophet learns model parameters via maximum likelihood estimation. Experiments show that this method can generate correct patches for 15 out of 69 real-world defects of the GenProg benchmark. We have also noticed that in NOPOL, it is possible to synthesize more than one patch with our SMT-based synthesis implementation. Hence, the probability model in Prophet can be leveraged to direct the synthesis of more correct patches by NOPOL.

7.2 Other Kinds of Repair

Besides test-suite based repair, other approaches are designed for fixing software bugs and improving software quality. Dallmeier et al. [11] propose Pachika, a fix generation approach via object behavior anomaly detection. This approach identifies the difference between program behaviors by the execution of passing and failing test cases; then fixes are generated by inserting or deleting method calls. Carzaniga et al. [8] develop an automatic technique to avoid failures by a faulty web application. This technique is referred as an automatic workaround, which aims to find and execute a correct program variant. AutoFix by Pei et al. [41], employs a contract-based strategy to generate fixes. This approach requires simple specifications in contracts, e.g., pre-conditions and post-conditions of a function, to enhance the debugging and fixing process. Experiments on Eiffel programs show that this approach can fix 42% of over 200 faults.

7.3 Empirical Foundations of Repair

Applying automatic repair to real-world programs is limited by complex program structures and semantics. We list existing work on the investigation of empirical foundations of test-suite based repair.

Martinez & Monperrus [32] mine historical repair actions to reason about future actions with a probabilistic model. Based on a fine granularity of ASTs, this work analyzes over 62 thousands versioning transactions in 14 repositories of open-source Java projects to collect probabilistic distributions of repair actions. Such distributions can be used as prior knowledge to guide program repairing.

Fry et al. [16] design a human study of patch maintainability with 150 participants and 32 real-world defects. This work indicates that machine-generated patches are slightly less maintainable than human-written ones; hence, patches by automatic repair could be used as the patches written by humans. Another
case study is conducted by Tao et al. [48]. They investigate the possibility of leveraging patches by automatic repair to assist the process of debugging by humans.

Barr et al. [5] address the “plastic surgery hypothesis” of genetic-programming based repair, such as GenProg. Their work presents evidences of patches based on reusable code, which make patch reconstitution from existing code possible. Martinez et al. [33] conduct empirical investigation to the redundancy assumption of automatic repair; this work indicates that code extracted from buggy programs could form a patch that passes the test suite.

Monperrus [36] details the problem statement and the evaluation of automatic software repair. This work systematically describes the pitfalls in software repair research and the importance of explicit defect classes; meanwhile, this paper identifies the evaluation criteria in the field: understandability, correctness, and completeness. Zhong & Su [56] examine over 9,000 real-world patches and summarize 15 findings in two key ingredients of automatic repair: fault localization and faulty code fix. This work provides empirical foundations for localization and patch generation of buggy statements.

Qi et al. [45] propose Kali, an efficient repair approach based on simple actions, such as statement removal. Their work presents the repair results via simple methods; meanwhile, their work checks previous empirical results by GenProg [27], AE [49], and RSRepair [43]. Empirical studies show that only two bugs by GenProg, three bugs by AE, and two bugs by RSRepair are correctly patched. All the reported patches for the other bugs are incorrect due to improper experimental configurations or semantic issues; an incorrect patch either fails to produce expected outputs for the inputs in the test suite, or fails to implement functionality that is expected by developers. As the latest result in test-suite based repair, the work by Qi et al. [45] shows that repairing real-world bugs is complex and difficult. Hence, it is worth investigating the empirical results on fixing real bugs.

Recent work by Smith et al. [46] investigates the overfitting patches on test cases in automatic repair. They report a controlled experiment on a set of programs written by novice developers with bugs and patches; two typical repair methods, GenProg [27] and RSRepair [44], are evaluated to explore the factors that affect the output quality of automatic repair.

Recent work by Le Goues et al. [26] presents two datasets of bugs in C programs to support comparative evaluation of automatic repair algorithms. The detailed description of these datasets is introduced and a quantified empirical study is conducted on the datasets. Defects4J by Just et al. [22] is a bug database that consists of 357 real-world bugs from five widely-used open-source Java projects. It has recently been shown [31] that NOPOL is capable of fixing 35 bugs of this benchmark.

7.4 Related Techniques: Program Synthesis and Fault Localization

Our approach, NOPOL, relies on two important techniques, program synthesis and fault localization.

Program synthesis aims to form a new program by synthesizing existing program components. Jha et al. [19] mine program oracles based on examples and employ SMT solvers to synthesize constraints. In this work, manual or formal specifications are replaced by input-output oracles. They evaluate this work on 25 benchmark examples in program deobfuscation. Their follow-up work [17] addresses the same problem by encoding the synthesis constraint with a first-order logic formula. In general, any advance in program synthesis can benefit program repair by enabling either more complex or bigger expressions to be synthesized.

In our work, fault localization is used as a step of ranking suspicious statements to find out locations of bugs. A general framework of fault localization is to collect program spectra (a matrix of testing results based on a given test suite) and to sort statements in the spectra with specific metrics (e.g., Tarantula [21] and Ochiai [4]). Among existing metrics in fault localization, Ochiai [4] has been evaluated as one of the most effective ones. In Ochiai, statements are ranked according to their suspiciousness scores, which are values of the Ochiai index between the number of failed test cases and the number of covered test cases. Fault localization techniques are further improved recently, for example, the diagnosis model by Naish et al. [37], the localization prioritization by Yoo et al. [54], and the test purification by Xuan & Monperrus [53].

8 Conclusion

In this paper, we have proposed NOPOL, a test-suite based repair approach using SMT. NOPOL targets two kinds of bugs: buggy IF conditions and missing preconditions. Given a buggy program and its test suite, NOPOL employs angelic fix localization to identify potential locations of patches and expected values of IF conditions. For each identified location, NOPOL collects test execution traces of the program. Those traces are then encoded as an SMT problem and the solution to this SMT is converted into a patch for the buggy program. We conduct an empirical evaluation on 22 real-world programs with buggy IF conditions and missing preconditions. We have presented four case studies to show the benefits of generating patches with NOPOL as well as the limitations.

NOPOL is publicly-available to support further replication and research on automatic software repair: http://github.com/SpoonLabs/nopol/.
In future work, we plan to evaluate our approach on more real-world bugs. Our future work also includes addressing the current limitations, e.g., designing better strategy for angelic fix localization, collecting more method calls, and improving the SMT encoding.

ACKNOWLEDGMENT

The authors would like to thank David Cok for giving us full access to jSMTLIB. This work is partly supported by the INRIA Internship program, the INRIA postdoctoral research fellowship, the CNRS delegation program, the National Natural Science Foundation of China (under grant 61502345), and the Young Talent Development Program of the China Computer Federation.

REFERENCES


Jifeng Xuan received the BSc degree in computer engineering from UNICEN, Argentina, and the PhD degree for University of Lille, France, in 2014. He is currently a postdoctoral researcher at University of Lugano (USI), Switzerland. His current research focuses on automatic software repair, software testing, and mining software repositories.

Matias Martinez received the Master degree in computer engineering from UNICEN, Argentina, and the PhD degree for University of Lille, France, in 2014. He is currently a postdoctoral researcher at University of Technology, China. His research interests include software testing and debugging, software data analysis, and search based software engineering. He is a member of the IEEE.

Favio DeMarco received the MSc degree in Computer Science from Universidad de Buenos Aires, Argentina. His research interests focus on software repair.

Maxime Clément is the MSc degree student in computer engineering from Lille 1 University, France. He is currently working in an IT services company as part of an internship program.

Sebastian Lamelas Marcote received the MSc degree in Computer Science from Universidad de Buenos Aires (FCEN), Argentina in 2015. He is currently working in the industry, but still has a research spirit focused on the integration of computer power with other scientific disciplines.
Thomas Durieux received the master degree in computer engineering from University of Lille, France in 2015. He is working toward the PhD degree in software engineering at the University of Lille. His current research focuses on automatic repair at runtime.

Daniel Le Berre is a professor of computer science at Artois University, Lens, France, and a member of the CRIL research lab, a research center specialized in artificial intelligence. His main research interests lie around Boolean constraint programming and its applications to software engineering.

Martin Monperrus has been an associate professor at the University of Lille, France, since 2011. He is an adjunct researcher at Inria. He was previously with the Darmstadt University of Technology, Germany, as a research associate. He received a Ph.D. from the University of Rennes in 2008. His research lies in the field of software engineering with a current focus on automatic software repair. He is a member of the IEEE.