IBM Research Report

Directed Test Generation for Improved Fault Localization

Shay Artzi, Julian Dolby, Frank Tip, Marco Pistoia
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
Directed Test Generation for Improved Fault Localization

Shay Artzi    Julian Dolby    Frank Tip    Marco Pistoia

IBM T.J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598, USA
{artzi,dolby,ftip,pistoia}@us.ibm.com

ABSTRACT
Fault-localization techniques that apply statistical analyses to execution data gathered from multiple tests are quite effective when a large test suite is available. However, if no test suite is available, what is the best approach to generate one? This paper investigates the fault-localization effectiveness of test suites generated according to several test-generation techniques based on combined concrete and symbolic (concolic) execution. We evaluate these techniques by applying the Ochiai fault-localization technique to generated test suites in order to localize 35 faults in four PHP Web applications. Our results show that the test-generation techniques under consideration produce test suites with similar high fault-localization effectiveness, when given a large time budget. However, a new, "directed" test-generation technique, which aims to maximize the similarity between the path constraints of the generated tests and those of faulty executions, reaches this level of effectiveness with much smaller test suites. On average, when compared to test generation based on standard concolic execution techniques that aims to maximize code coverage, the new directed technique preserves fault-localization effectiveness while reducing test-suite size by 86.1% and test-suite generation time by 88.6%.

1. INTRODUCTION
When a test fails, developers need to find the location of the fault in the source code before they can fix the problem. In recent years, a number of automated techniques have been proposed to assist programmers with this task, which is usually called fault localization. Many fault-localization techniques attempt to predict the location of a fault by applying statistical analyses to data obtained from the execution of multiple tests (see, for example, [18, 19, 20, 17, 2, 26]). The basic idea is that a statement [18], control-flow predicate [19], or def-use pair [26] is more suspicious (or more likely to be responsible for a test failure) if it correlates strongly with failing tests, and less suspicious if it correlates strongly with passing tests.

For a given application that contains a fault and a given fault-localization algorithm, one way to measure a test suite’s suitability for fault localization is the number of statements that must be explored until the fault is found, assuming that statements are explored in order of decreasing suspiciousness. However, as in previous work on fault localization [17, 4], we instead choose to concentrate on the percentage of faults that are well-localized—these are faults for which less than 1% of all executed statements need to be examined until the fault is found, assuming that statements are examined in order of decreasing suspiciousness. In this paper, we explore a number of strategies for generating test suites and measure their fault-localization effectiveness, defined as the percentage of well-localized faults for that suite. Our goal is to determine which test-generation strategy achieves the best tradeoff between size and effectiveness for the test suites that it generates.

The research presented in this paper was conducted in the context of Apollo [5, 6, 4], a tool that uses combined concrete and symbolic (concolic) execution [12, 27, 9, 13, 30] to generate failure-inducing inputs for PHP Web applications. Apollo currently targets two kinds of failures: (i) HTML failures, in which the application generates malformed HTML, and (ii) execution failures, which manifest themselves by a program crash or an obtrusive error message. In recent work [4], we incorporated several variations of the Tarantula fault-localization technique [18] in Apollo, and demonstrated that these can localize faults quite well using a test suite that was generated using concolic execution. However, in those previous experiments, the test suites were not generated with fault localization in mind, but with the dual objective of maximizing code coverage and finding as many failures as possible. Therefore, it is not clear whether the test suites used in [4] have maximal effectiveness for fault localization. In particular, the question is if even better fault localization can be achieved using test suites that are generated with other test-generation strategies. Furthermore, it is conceivable that the use of other test-generation strategies might enable equally effective fault localization using much smaller test suites.

In this paper, we assume a scenario where a user has just encountered a failure and where no test suite is available. For such situations, we want to answer the following research question: What is the best strategy for generating a test suite that has maximal effectiveness for fault localization? Notice that the effectiveness of fault-localization techniques is premised on the availability of a high-quality test suite that provides good coverage of the application’s statements. However, even 100% coverage does not guarantee good fault localization.

Consider, for example, a scenario where two statements $s_1$ and $s_2$ are executed by exactly the same tests and where Tarantula [18] or Ochiai [2] is used for fault localization. These fault-localization techniques compute a program construct’s suspiciousness rating from the number of passing tests and failing tests that execute it. Hence, if $s_1$ and $s_2$ are executed by the same tests, those techniques will report them as being equally suspicious. In such cases, creating an additional test that executes $s_1$ but not $s_2$ (or vice versa) may
enable better fault localization. In this paper, we present a number of strategies for generating tests that are parameterized by a similarity criterion, and evaluate their fault-localization effectiveness. The similarity criterion measures how similar the execution characteristics associated with two tests are, and is used to direct concolic execution towards generating tests whose execution characteristics are similar to those of a given failing test.

We implemented the techniques in Apollo, an automated tool that detects and localizes faults in PHP Web applications. We evaluated the test-generation techniques by localizing 35 faults in four PHP applications. The results that we present in Section 5 show that a new, directed test-generation technique based on path-constraint similarity yields the smallest test suites with the same excellent fault-localization characteristics as test suites generated by other techniques. In particular, when compared to test generation based on the concolic execution algorithm of [4], which aims to maximize code coverage, our directed technique reduces test-suite size by 86.1% and test-suite generation time by 88.6%.

The remainder of this paper is organized as follows. Section 2 presents a motivating example. Section 3 provides details on our techniques for concolic execution and fault localization. Section 4 presents the similarity criteria with which our directed test generation algorithms are parameterized. Section 5 presents the implementation and experimental results. Finally, related work is discussed in Section 6, followed by conclusions in Section 7.

2. EXAMPLE

Figure 1 shows a simple PHP script that scales one point to have the same magnitude as another. Object-oriented code in PHP resembles code in C++ or Java. The point class in Figure 1 declares two fields, \( x \) and \( y \). Also declared in class point are methods \( \text{magnitude}() \), which computes the distance of a point from the origin, and \( \text{scale}() \), which scales a point to have the same magnitude as an argument point. Two PHP features of particular note are the \( \text{isset}() \) construct for checking whether a variable has been defined, and the \$REQUEST associative array, which is used to access input values that were provided to a PHP script.

The script code that follows the class definition in Figure 1 creates a point object and then initializes its \( x \) and \( y \) fields with values \( x \) and \( y \) that were provided as inputs to the script by accessing \$REQUEST. The program then calls \( \text{isset}() \) to check if input values \( \text{scale}_x \) and \( \text{scale}_y \) have been provided to the script. If so, the program creates and initializes the point \$scale by which to scale the point \$x. The intended use of this script is to be given two points, and scale one of them according to the other. There are two issues with this code:

1. This scaling is not well defined for the origin, which causes a division by zero in the \( \text{scale}() \) method in that case. We will assume that the desired fix for this is to leave the origin point unscaled. Note that the division by zero is a warning rather than an error in PHP, and so execution continues using a default value—0 in this case—as the “result” of the division.

2. There is also an inconsistency if the scaling parameters are not provided: In this case, the \( \text{scale} \) point will not be created, but it will be used anyway, resulting in attempting to call \( \text{magnitude}() \) on a non-object. This error aborts the PHP script, and execution ceases at that point. We will assume that, in the absence of the \( \text{scale}_x \) and \( \text{scale}_y \) parameters, the intended behavior is not to scale the point.

Our fault-localization procedure assumes that we have some test—either written by hand or generated by some technique, possibly

```php
<?php
1 error_reporting( E_ALL );
2
class point {
3 var $x;
4 var $y;
5
6 function magnitude() {
7 return sqrt($this->x*$this->x+$this->y*$this->y);
8 }
9
10 function scale($p) {
11 $factor = $p->magnitude() / $this->magnitude();
12 $this->x = $this->x * $factor;
13 $this->y = $this->y * $factor;
14 }
15
16
17 $x = new point();
18 $x->x = $REQUEST['x'];
19 $x->y = $REQUEST['y'];
20
21 if (isset($REQUEST['scale_x']) &
22 isset($REQUEST['scale_y']) ) {
23 $scale = new point();
24 $scale->x = $REQUEST['scale_x'];
25 $scale->y = $REQUEST['scale_y'];
26
27 }
28
29 $x->scale($scale);
30
31 print "$x->x." "$x->y." "$x->magnitude()."\n";
?>
```

Figure 1: Example PHP Program with Bugs

Apollo—that exposes a given failure. There are many tests that could reveal these failures, and so we choose the following ones arbitrarily.

1. The first fault can be revealed by any choice of inputs that defines the \( \text{scale}_x \) and \( \text{scale}_y \) parameters and uses 0 for both \( x \) and \( y \). In this case, the if test will succeed, and so the script will call \( \text{scale}() \). Then the call to \( \text{magnitude}() \) on \$this in \( \text{scale}() \) will return 0, triggering the divide-by-zero failure.

2. The second fault can be revealed by any input that does not define either \( \text{scale}_x \) or \( \text{scale}_y \) or both. In this case, the if test will fail, causing the code that creates and initializes \$scale not to execute. In this case, the script will try to scale \$x by the undefined \$scale, resulting in an error when it tries to access a field of the undefined value.

3. APPROACH

This section describes in detail our solution for directed test generation for fault localization.

3.1 Concolic Testing

We begin by briefly reviewing the combined concrete and symbolic execution algorithm as embodied in Apollo [4]. The idea of this algorithm is to execute an application on some initial input
The full algorithm can be found in [6], Figure 6.
tests generate different tuples for that method when it is called on points with different magnitudes. Note also that 22 is the line of the if statement, so its statements are augmented with the conditional outcome. This is illustrated in Figure 4.

Consider the calls to p.magnitude() in Figure 4; recall that we are recording both the statement itself and its return value. There are two of the four tests that execute this statement and get a result of 0. Looking at the Ochiai formula, we see that it is of suspiciousness 1, since \(a_{i1} = 1\) and \(d_{i0} = 0\). Indeed, one way to fix this issue is to handle the case when this call returns 0. This result and the lower suspiciousness for all other statements is shown in Figure 5.

On the other hand, for the second test suite, only the given test F exhibits the bug. And also only F executes the call to scale(), so we see many more statements that are correlated with the bug. This is shown in Figure 6 and Figure 7, where statements 23, 24, 25, 26, 12, 13, 14, 15 and 31 all correspond exactly to the failing test.

Thus, we observe that, especially for small test suites, the choice of tests can make a big difference. Our work focuses on selection strategies that allow us to focus fault localization quickly.

### 4. SIMILARITY METRICS

This paper evaluates various strategies for automatic test generation for the purpose of fault localization. Given a failing execution, the general intuition behind our techniques is that localizing the corresponding fault is more effective if a passing test is generated whose characteristics are “similar” to those of the failing execution, because that maximizes the chances that the fault is correlated with the difference between the path constraints of the generated passing test and those of the faulty execution; the smaller the difference, the higher the precision with which the fault can be localized.

For this to be more precise, we need to formalize the concept of “similarity” between two executions. This leads us to introducing a similarity criterion, which is a function that takes as an input two executions, and produces as output a percentage index that indicates how similar the two executions are. More formally, if \(E\) is the set of all the executions of a program, a similarity criterion is a function \(\sigma_a : E \times E \rightarrow [0, 100]\), where \(a\) is itself a function that abstracts executions. Specifically, \(a\) maps each execution \(e \in E\) to a set of characteristics of \(e\) that depend on the particular similarity metric under consideration. There can be multiple similarity criteria, each based on what characteristics are considered when measuring similarity and, consequently, what abstraction function \(a\) is being considered.

A similarity criterion \(\sigma_a\) can be extended to a function \(\sigma_a' : E \times 2^E \rightarrow [0, 100]\), defined as follows:

\[
\sigma_a'(e, S) := \max_{e' \subseteq S} \sigma_a(e, e')
\]

which can be used to compare a passing execution with a set of failing executions.

In order to guide our test generation technique (Section 3.1) towards generating similar executions, a similarity function is used as the selection methodology (Figure 2, line 4). The selection methodology is responsible for selecting the next input to explore, thus directing the generation to explore similar executions.

In this paper, we consider two different similarity metrics: path constraints and inputs. These two approaches and the relevant similarity criteria are described in the next subsections.

#### 4.1 Path-Constraint Similarity

In general, any execution is generated by a set of inputs to the program. This defines a function \(f : 2^I \rightarrow E\), where \(I\) is the set of inputs to the program. Function \(f\) maps any set of program inputs to one program execution. Furthermore, given a particular set

<table>
<thead>
<tr>
<th>test</th>
<th>args</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>x=0, y=0, scale_x=1, scale_y=3</td>
</tr>
<tr>
<td>A1</td>
<td>x=0, y=0, scale_x=3, scale_y=2</td>
</tr>
<tr>
<td>A2</td>
<td>x=5, y=0, scale_x=1, scale_y=0</td>
</tr>
<tr>
<td>A3</td>
<td>x=3, y=2, scale_x=2, scale_y=4</td>
</tr>
</tbody>
</table>

**Figure 3: Example test suites**

<table>
<thead>
<tr>
<th>statement</th>
<th>executions</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>13</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>14</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>18</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>19</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>20</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>23</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>24</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>25</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>26</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>29</td>
<td>F, A1, A2, A3</td>
</tr>
<tr>
<td>31</td>
<td>F, A1, A2, A3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>suspiciousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>.71</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 5: Example test suite 1 suspiciousness**

<table>
<thead>
<tr>
<th>test</th>
<th>statements executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>18, 19, 20, &lt;22, true&gt;, 23, 24, 25, 26, 29, 12, &lt;8, 3.1&gt;, &lt;8, 0&gt;, 13, 14, 31, &lt;8, 0&gt;</td>
</tr>
<tr>
<td>A1</td>
<td>18, 19, 20, &lt;22, true&gt;, 23, 24, 25, 26, 29, 12, &lt;8, 3.6&gt;, &lt;8, 0&gt;, 13, 14, 31, &lt;8, 0&gt;</td>
</tr>
<tr>
<td>A2</td>
<td>18, 19, 20, &lt;22, true&gt;, 23, 24, 25, 26, 29, 12, &lt;8, 1&gt;, &lt;8, 8.6&gt;, 13, 14, 31, &lt;8, 1&gt;</td>
</tr>
<tr>
<td>A3</td>
<td>18, 19, 20, &lt;22, true&gt;, 23, 24, 25, 26, 29, 12, &lt;8, 3.6&gt;, &lt;8, 3.6&gt;, 13, 14, 31, &lt;8, 3.6&gt;</td>
</tr>
</tbody>
</table>

**Figure 4: Example test suite 1 executions**

<table>
<thead>
<tr>
<th>test</th>
<th>statements executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>x=3</td>
</tr>
<tr>
<td>B2</td>
<td>x=5, y=4</td>
</tr>
<tr>
<td>B3</td>
<td>x=4, y=0, scale_x=3</td>
</tr>
</tbody>
</table>

**Figure 6: Example test suite 2 executions**

In this paper, we consider two different similarity metrics: path constraints and inputs. These two approaches and the relevant similarity criteria are described in the next subsections.

#### 4.1 Path-Constraint Similarity

In general, any execution is generated by a set of inputs to the program. This defines a function \(f : 2^I \rightarrow E\), where \(I\) is the set of inputs to the program. Function \(f\) maps any set of program inputs to one program execution. Furthermore, given a particular set
of program inputs ε ∈ I, a heuristic function can compute an input from a given path constraint. This defines a function $g : 2^P \rightarrow 2^I$, where $P$ is the set of path constraints that can arise during the execution of the program. The composition function $f \circ g : 2^P \rightarrow E$ can be used to base a similarity criterion on path constraints instead of actual executions, which is a very useful property in test generation. Specifically, given a set $\pi$ of path constraints and the corresponding execution $e = f(g(\pi))$, we define $\alpha(e) = \pi$, and we use the resulting function $\alpha : E \rightarrow 2^P$ to parameterize the similarity criterion $\sigma$.

We have implemented two techniques for path-constraint similarity: subset comparison and subsequence comparison. With subset comparison, execution similarity is computed based on the cardinality of the largest subset of identically evaluating conditional statements that are traversed in the two executions; with subsequence comparison, execution similarity is computed based on the cardinality of the largest subsequences of conditions that evaluate to the same value in both executions.

To better understand the difference between these two metrics, consider for example two program executions $e_1, e_2 \in E$ that evaluate conditions $(C_1, C_2, C_3, C_4, C_5, C_6)$, and assume that condition $C_3$ evaluates to true in $e_1$ and false in $e_2$, but $C_1, C_2, C_4, C_5, C_6$ evaluate to the same boolean value in both executions. In this case, $\sigma_{\text{subset}}(e_1, e_2) = 83.3\%$ if the similarity criterion is based on subset comparison, and $\sigma_{\text{subsequence}}(e_1, e_2) = 50\%$ if the similarity criterion is based on subsequence comparison. In practice, we observed that these two similarity metrics lead to very similar results. Therefore, in the remainder of this paper, we concentrate only on path-constraint similarity based on subset comparison.

### 4.2 Input Similarity

With this approach, we compare the inputs to different executions. Each execution $e \in E$ is reduced to only its inputs, as follows. Given a set $\pi$ of path constraints, we consider the corresponding set of execution inputs $g(\pi)$, and we define $\alpha(e) = g(\pi)$. We then use the resulting function $\alpha : E \rightarrow 2^I$ to parameterize the similarity criterion $\sigma$.

Input similarity is based on subset comparison: the similarity between two executions is computed based on the number of inputs that are identical for both executions. For example, consider two executions $e_1$ and $e_2$ with inputs $(S_1, S_2, S_3, S_4, S_5, S_6)$ and $(T_1, T_2, T_3, T_4, T_5, T_6)$, respectively, such that $S_3 \neq T_3$, but $S_i = T_i, \forall i \neq 3$. In this case, $\sigma_{\text{input}}(e_1, e_2) = 83.3\%$.

### Example.

The example in Figure 1 motivates the use of different similarity criteria for fault-localization-oriented test generation. Since the sample program only exhibits three different paths, a test-generation technique based on input similarity is more effective for fault localization.

Let us assume that the program in Figure 1 fails with inputs given in $F$, where $F$ is defined as in Figure 3. By looking at test suites $A$ and $B$ in Figure 3, we observe that test suite $A$, generated with the input-similarity technique, allows for quick fault localization. This is due to the fact that the faulty statement in the program is executed by multiple failing tests in suite $A$, namely $F$ and $A1$. Even more importantly, suite $A$ has the advantage of presenting a passing test, $A2$, that is similar to the failing execution $F$.

If all the tests in the $A$ and $B$ suites are available, $A1$ will be the first test to be selected with the input-similarity strategy since it is the most similar to the faulty execution $F$ that is given as input to the algorithm. That will be followed by test $A2$, which is the second test to be the most similar to the faulty execution. $A2$ will be followed by $A3$. Notice that both $A2$ and $A3$ are passing tests. This makes it possible to localize and isolate the faulty statement with a test-suite size of only 4. In contrast, the coverage strategy would get full coverage with $F$, and then would select random tests, with a potential test-suite size of 7, before making the fault localizable.

### 5. IMPLEMENTATION AND EVALUATION

We implemented several test-generation strategies in Apollo [5, 4], a tool for automatically finding and localizing faults in PHP web applications. This section reports on the implementation, and on experiments that measure the effectiveness of the different test-generation strategies.

#### 5.1 Implementation

For the purpose of test generation, we use Apollo [5, 6], which employs a shadow interpreter based on the Zend PHP Interpreter VS.2.2.2. Apollo simultaneously performs concrete program execution using concrete values, and a symbolic execution that uses symbolic values that are associated with variables. Furthermore, Apollo uses the choco solver to solve path constraints during the concolic generation. The process of concolic execution is orchestrated by a standard Apache Web server that uses the instrumented PHP interpreter.

Our fault-localization technique performs conditional [4] and return-value modeling. These two enhancements where implemented on top of Apollo’s shadow interpreter. For the conditional modeling, Apollo records all comparisons in the executed PHP script. For each comparison, the shadow interpreter stores the statement’s line number and the relevant boolean result. For a switch statement, the shadow interpreter stores the line number of the call, and an abstract model of the value. The model allows the fault localization technique to distinguish between null and non-null values, zero and non-zero int and double values, true and false boolean values, constant and non-constant values, as well as empty and non-empty arrays, strings, and resources.

1[http://www.php.net/]
2[http://choco-solver.net]
3[http://www.apache.org/]
Table 1: Characteristics of subject programs. The #files column lists the number of .php and .inc files in the program. The PHP LOC column lists the number of lines that contain executable PHP code. The #downloads column lists the number of downloads from http://sourceforge.net.

<table>
<thead>
<tr>
<th>subject</th>
<th># faults</th>
</tr>
</thead>
<tbody>
<tr>
<td>webchess</td>
<td>14</td>
</tr>
<tr>
<td>faqforge</td>
<td>7</td>
</tr>
<tr>
<td>schoolmate</td>
<td>2</td>
</tr>
<tr>
<td>total</td>
<td>35</td>
</tr>
</tbody>
</table>

Figure 8: Number of faults used in the localization experiments. Each fault manifested itself as an execution failure.

Figure 6: Number of faults used in the localization experiments. Each fault manifested itself as an execution failure.

Figure 5: Number of faults used in the localization experiments. Each fault manifested itself as an execution failure.

Table 2: Number of faults used in the localization experiments. Each fault manifested itself as an execution failure.

<table>
<thead>
<tr>
<th>subject</th>
<th># faults</th>
</tr>
</thead>
<tbody>
<tr>
<td>webchess</td>
<td>14</td>
</tr>
<tr>
<td>faqforge</td>
<td>7</td>
</tr>
<tr>
<td>schoolmate</td>
<td>2</td>
</tr>
<tr>
<td>total</td>
<td>35</td>
</tr>
</tbody>
</table>

Figure 8: Number of faults used in the localization experiments. Each fault manifested itself as an execution failure.

5.2 Research Questions

For each of the test-generation strategies under consideration, we are interested in determining the maximal fault-localization effectiveness that can be achieved using test suites generated according to that strategy. As in previous work on fault localization, we will concentrate on the percentage of well-localized faults for which less than 1% of all executed statements need to be examined until the fault is found, assuming that statements are examined in order of decreasing suspiciousness. It is reasonable to expect that a limited amount of time will be available for test generation. Therefore, we are also interested in determining how quickly each of the test-generation strategies under consideration converges towards its maximal effectiveness. This leads us to formulate the following research questions:

RQ1. What is the maximal fault-localization effectiveness of test suites, measured as the percentage of well-localized faults, generated by each of the test-generation strategies?

RQ2. How many tests need to be generated by each test-generation strategy in order to reach its maximal fault-localization effectiveness?

5.3 Subject Programs

For the evaluation, we selected four open-source PHP programs from http://sourceforge.net:

- faqforge is a tool for creating and managing documents
- webchess is an online chess game
- schoolmate is a PHP/MySQL solution for administering elementary, middle, and high schools
- phpsysinfo is a utility for displaying system information, such as uptime, CPU, memory, etc.

Figure 1 presents some characteristics of these programs.

5.4 Methodology

In order to answer our research questions, we needed localized faults. We used actual faults that were discovered by Apollo [5, 6]. In our previous work, Apollo was used to discover two types of failures: HTML failures that occur when malformed HTML is generated, and execution failures when an input causes a crash or obtrusive error message. In this paper, we restrict our attention to execution failures for which the location of the fault is not immediately obvious from an error message. The reason for restricting our attention to these cases is that, in our opinion, these are the situations where the use of an automated fault localization is most warranted.

We manually localized all faults. For each fault, we devised a patch and ensured that applying this patch fixed the problem, by running the tests again, and making sure that the associated failures had been corrected. Figure 8 summarizes the number of faults for each subject program.

We used the following four test-generation strategies to generate test suites used for fault localization:

- Base Test generation using the concolic execution algorithm of [6], which starts from an empty input, and aims to maximize branch coverage. We call this algorithm Base because we will use it as the baseline for comparison with the new similarity-based directed generation algorithms.
- Coverage Test generation using the concolic execution algorithm of [6], but starting test generation from the failing test.
- PCS (Path Constraint Similarity) Test generation using the subset-based path-constraint similarity metric that was described in Section 4.1
- IS (Input Similarity) Test generation using the input similarity that was described in Section 4.2.

For each strategy and for each fault we used Apollo to generate test suites. Then, for each test suite and each localized fault, we computed suspiciousness ratings for all executed statements according to the Ochiai technique [2] with the improvements described in Section 3.2. Similar to previous fault-localization studies [18, 11, 17, 26], we measured the effectiveness of a fault localization algorithm as the minimal number of statements that need to be inspected until the first faulty line is detected, assuming that statements are examined in order of decreasing suspiciousness. We then computed the number of statements to be inspected as a percentage of the number of executed statements. Finally, we computed the percentage of faults that are “well-localized”, meaning that they require the inspection of less than 1% of all executed statements.

5.5 RQ1

We first discuss the “maximal” fault-localization effectiveness of the test suites generated by the four test-generation techniques above, as measured by the percentage of well-localized faults, assuming each technique is given an infinite amount of time to construct a test suite. In practice, we found that it sufficed to have each technique generate 100 tests for each fault, with the exception of schoolmate, which required 252 tests to reach a plateau. Generating more tests beyond this point resulted in larger test suites, but not in an increased number of well-localized faults. Table 2 shows three columns for each subject program and each technique. These columns show, from left to right: (i) on average, for each subject program, the percentage of faults that is well-localized, (ii) on average, the absolute number of statements that needs to be inspected to localize each fault, and (iii) on average, the percentage of executed statements that are inspected until the first faulty line is detected, assuming that statements are examined in order of decreasing suspiciousness. We then computed the number of statements to be inspected as a percentage of the number of executed statements. Finally, we computed the percentage of faults that are “well-localized”, meaning that they require the inspection of less than 1% of all executed statements.

It is theoretically possible that some minor further gains could be achieved by generating many additional tests, but we consider this to be very unlikely.
statements that needs to be inspected to localize each fault. For example, for faqforge, both the Base and PCS techniques eventually localize 100% of the faults to within 1% of all executed statements. Furthermore, on average, each of these faults is localized by these techniques to 4.6 statements, which corresponds to 0.6% of all executed statements. The Coverage and IS generation techniques also reach 100% well-localized faults on faqforge eventually, albeit a slightly higher plateau of 5 and 5.1 statements, respectively, that need to be inspected, which corresponds to 0.7% of all executed statements.

In summary, the test-generation strategies are capable of generating test suites with nearly identical maximal fault-localization effectiveness when given an infinite amount of time. In particular, for faqforge, schoolmate and phpsysinfo, 100% of all faults was eventually well-localized by each technique. However, for webchess, only 77% of all faults was eventually well-localized by each technique.

<table>
<thead>
<tr>
<th>program</th>
<th>% wl</th>
<th>% stmts</th>
<th>% stmts</th>
<th>% wl</th>
<th>% stmts</th>
<th>% stmts</th>
<th>% wl</th>
<th>% stmts</th>
<th>% stmts</th>
<th>% wl</th>
<th>% stmts</th>
<th>% stmts</th>
</tr>
</thead>
<tbody>
<tr>
<td>webchess</td>
<td>77</td>
<td>11.3</td>
<td>1.3</td>
<td>77</td>
<td>16.4</td>
<td>1.9</td>
<td>77</td>
<td>16.9</td>
<td>1.9</td>
<td>77</td>
<td>16.8</td>
<td>1.9</td>
</tr>
<tr>
<td>faqforge</td>
<td>100</td>
<td>4.6</td>
<td>0.6</td>
<td>100</td>
<td>5</td>
<td>0.7</td>
<td>100</td>
<td>4.6</td>
<td>0.6</td>
<td>100</td>
<td>5.1</td>
<td>0.7</td>
</tr>
<tr>
<td>schoolmate</td>
<td>100</td>
<td>11.4</td>
<td>0.4</td>
<td>100</td>
<td>6.8</td>
<td>0.2</td>
<td>100</td>
<td>4.6</td>
<td>0.2</td>
<td>100</td>
<td>4.6</td>
<td>0.2</td>
</tr>
<tr>
<td>phpsysinfo</td>
<td>100</td>
<td>17</td>
<td>0.7</td>
<td>100</td>
<td>17</td>
<td>0.7</td>
<td>100</td>
<td>14</td>
<td>0.6</td>
<td>100</td>
<td>14</td>
<td>0.6</td>
</tr>
<tr>
<td>Average</td>
<td>100</td>
<td>11.1</td>
<td>0.75</td>
<td>100</td>
<td>11.3</td>
<td>0.9</td>
<td>100</td>
<td>10</td>
<td>0.83</td>
<td>100</td>
<td>10.1</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 2: Summary, for each test generation technique and subject program, of the percentage of faults that is well-localized, and the absolute number (# stmts) and percentage (% stmts) of executed statements that need to be inspected on average until the fault is localized.

As we have seen, the different test generation techniques eventually achieve very similar effectiveness. However, the question remains to what extent the test generation techniques require a different number of tests to reach this plateau. Table 3 shows two columns for each subject program and each test generation technique. These columns show, from left to right: (i) the number of tests that is needed to reach the maximal percentage of well-localized faults as reported in Table 2, and (ii) the time required to generate these tests. Here, it should be noted that the time reported in (ii) is an average over all faults for the Coverage, PCS, and IS techniques. For the Base technique, there is just one test suite that is used for all faults, and the time reported is the time needed to generate that test suite.

As can be seen in Table 3, there are significant difference in how quickly the different test generation techniques converge on the optimal result. For faqforge, the Base test generation technique that we used in [4] requires 60 tests to reach the maximal percentage of well-localized faults, whereas the PCS technique requires only 5 tests. The amount of time required to generate a test suite differs similarly, with 63.6 seconds for the Base technique and only 7.3 seconds for the PCS technique. The graphs in Figure 9 provide some more detail on how quickly the test generation strategies converge towards their maximal effectiveness. Each graph shows the percentage of well-localized faults plotted against the number of generated tests, for each of the generation techniques. By examining the graphs, we can observe that the directed strategies (IS and PCS) converge much faster than the undirected strategies (Coverage and Base). In three of the four subject programs (webchess, faqforge, and phpsysinfo), the PCS strategy is superior. In the case of schoolmate, however, the IS strategy (7 tests) is slightly better than PCS (11 tests).

On the whole, we conclude that the PCS strategy is the preferred technique. On average, PCS requires only 6.5 tests to achieve the optimal number of well-localized faults, versus 46.8 tests for the Base strategy that we used in our previous work [4]. This can be viewed an an improvement of (46.8 − 6.5) * 100)/46.8 ≈ 86.1%. Similarly, we notice that, on average, the Base strategy takes 131.2 seconds for test generation, compared to only 14.9 seconds required by PCS, for an improvement of 88.6%.

5.7 Threats to Validity

There are several objections a critical reviewer might raise to the evaluation presented in this section. First, one might argue that the benchmarks are not representative of real-world PHP programs. To minimize this issue, we selected open-source PHP applications that are widely used, as is evidenced by the significant number of downloads reported in Figure 1. Second, it could be the case that the faults we exposed and localized are not representative. However, these bugs represent a range of execution errors, and all of them were exposed by automatic and systematic means. A potentially more serious issue is that any given fault may be fixed in multiple different ways. The fixes we devised were mostly simple one-line code changes, for which we attempted to come up with the simplest possible solution. And we verified that our fixes indeed resolved the issues, so to that extent our fixes are correct. The most serious criticism to our evaluation, in our own opinion, is the assumption that programmers would inspect the statements strictly in decreasing order of suspiciousness. In practice, it is very likely that programmers who try to follow this discipline would automatically look at adjacent statements, so the assumption is probably not completely realistic. However, we ourselves made use of the fault localization results while debugging. Also, this approach to measuring the effectiveness of fault localization methods has been used in previous research in the area (e.g., [18, 17, 26]), which makes our results more comparable with related work.

6. RELATED WORK

Fault localization algorithms have been studied extensively in recent years. In this section, we concentrate on fault localization techniques that predict the location of faults based on the analysis of data from multiple tests or executions. A discussion of other fault localization techniques can be found in [4].

Early work. Early work on fault localization relied on the use of program slicing [29]. Lyle and Weiser [22] introduce program dicing, a method for combining the information of different program slices. The basic idea is that, when a program computes a correct value for variable x and an incorrect value for variable y, the fault is likely to be found in statements that are in the slice w.r.t. y, but not in the slice w.r.t. x. Variations on this idea technique were later explored by Pan and Spafford [23], and by Agrawal, et al. [3]. In the spirit of this early work, Reniers and Reiss [25] use set-union and set-intersection methods for fault localization. The
<table>
<thead>
<tr>
<th>program</th>
<th>Base</th>
<th>Coverage</th>
<th>PCS</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># tests</td>
<td>time(s)</td>
<td># tests</td>
<td>avg time(s)</td>
</tr>
<tr>
<td>webchess</td>
<td>69</td>
<td>78.9</td>
<td>20</td>
<td>22.7</td>
</tr>
<tr>
<td>faqforge</td>
<td>60</td>
<td>63.6</td>
<td>37</td>
<td>47.5</td>
</tr>
<tr>
<td>schoolmate</td>
<td>18</td>
<td>20.1</td>
<td>253</td>
<td>386.2</td>
</tr>
<tr>
<td>phpsysinfo</td>
<td>40</td>
<td>362.3</td>
<td>87</td>
<td>818</td>
</tr>
<tr>
<td>Average</td>
<td>46.8</td>
<td>131.2</td>
<td>99.2</td>
<td>318.6</td>
</tr>
</tbody>
</table>

Table 3: Summary, for each test generation technique and subject program, of the time (time(s) for Base and avg time(s) for Coverage, PCS and IS) and number of tests (# tests) required to achieve the maximal percentage of well-localized faults, as reported in Table 2.

Figure 9: Average number of statements to inspect for all the execution failure found in each subject program. (a) faqforge, (b) schoolmate, (c) phpsysinfo, (d) webchess (e) aggregated
set-union technique computes the union of all statements executed by passing test cases and subtracts these from the set of statements executed by a failing test case. The resulting set contains the suspicious statements that the programmer should explore first. The set-intersection technique identifies statements that are executed by all passing test cases, but not by the failing test case, and attempts to address errors of omission, where the failing test case neglects to execute a statement.

Nearest Neighbors. The nearest neighbors fault localization technique by Reniers and Reiss [25] assumes the existence of a failing test case and many passing test cases. The technique selects the passing test case whose execution spectrum most closely resembles that of the failing test case according to one of two distance criteria, and reports the set of statements that are executed by the failing test case but not by the selected passing test case. In the event that the report does not contain the faulty statement, Reniers and Reiss use a ranking technique in which additional statements are considered based on their distance to previously reported statements along edges in the System Dependence Graph (SDG) [14]. Nearest Neighbor was evaluated on the Siemens suite [15], and was found to be superior to the set-union and set-intersection techniques.

Tarantula and Ochiai. Tarantula [18] is a fault-localization technique that associates with each statement a suspiciousness rating that indicates the likelihood for that statement to contribute to a failure. Jones and Harrold [17] conducted a detailed empirical evaluation in which they apply Tarantula to faulty versions of the Siemens suite [15], and compare its effectiveness to that of several other fault-localization techniques. In the fault localization literature, the effectiveness of a fault-localization technique is customarily measured by reporting the percentage of the program that needs to be examined by the programmer, assuming that statements are inspected in decreasing order of suspiciousness [11, 2, 25, 17]. Santelices, et al. [26] investigate the tradeoffs of applying the Tarantula algorithm to different types of program entities: statements, branches, and def-use pairs. The outcome of this study is that the branch-based algorithm is more precise than the statement-based one, and that the def-use-based variant is more precise still.

Recent papers by Jones and Harrold [17] and by Abreu, et al. [2] present empirical evaluations of several fault localization techniques, including several of the techniques discussed above, using the Siemens suite. The emerging consensus appears to be that the Ochiai similarity metric advocated by Abreu et al. [2] is more effective than Tarantula [26, 1]. Our own experiments with Apollo confirm this, which is why the work presented in this paper uses the Ochiai technique.

Other Statistical Techniques. Other fault localization techniques analyze statistical correlations between control flow predicates [20, 21] or path profiles [10] and failures, time spectra [31], and correlations between changes made by programmers and test failures [28, 24].

Fault Localization using Generated Test Suites Our work differs from the previous approaches for fault localization described above in that we rely on dynamic test generation techniques to create test suites. In a previous paper [4], we relied on concolic execution [12, 27, 9, 13, 30] to generate test suites. In that work, the generation of tests was guided by the objective of maximizing code coverage, by directing the solver to solutions that cover additional branches of condition statements. We found that, given a time budget of 20 minutes, 87.7% of the faults under consideration are localized to within 1% of all executed statements. The present paper uses the same technique for test generation, but the generation of tests is now guided by various similarity metrics in order to produce tests that are similar to a given failing execution. This is the same spirit as the Nearest Neighbors algorithm [25], but instead of selecting tests that are similar to a given failing test, we are attempting to generate such tests.

The Impact of Test Suite Composition on Fault Localization. Several other projects have explored the relationship between the composition of a test suite and its effectiveness for fault localization. Baudry et al. [7] study how the fault localization effectiveness of a test suite can be improved by adding tests. They propose the notion of a dynamic basic block, which is a set of statements that is covered by the same tests, and a related testing criterion that aims to maximize the number of dynamic basic blocks. Baudry et al. use a genetic algorithm for deriving new tests from existing ones by a series of mutation operations. Our research also aims to improve fault localization effectiveness by creating tests, but our starting point is a situation where no test suite is available. In such cases, it is not clear how mutation-based approaches, which generate new tests from existing ones, could be applied.

Other researchers have focused on the opposite problem: determining how reducing the size of a test suite impacts fault localization effectiveness. Yu et al. [32] study the impact of several test suite reduction strategies on fault localization. They conclude that statement-based reduction approaches negatively affect fault localization effectiveness, but that vector-based reduction approaches, which aim to preserve the set of statement vectors exercised by a test suite, have negligible effects on effectiveness. Jiang et al. [16] also study the impact of test suite reduction strategies on fault localization effectiveness. One of the strategies they consider (AS) prefers those test cases that maximally increase the number of additional statements covered. They report that reducing a test suite to half its original size according to this strategy only has minimal impact on fault localization effectiveness.

Other Directed Concolic Testing. In [8], the authors investigate 3 strategies to direct concolic testing to improve coverage. Heuristics based on the Control Flow Graphs (CFGs) of functions are evaluated in the context of real C programs. The most effective CFG-based heuristic is able to improve branch coverage substantially, more than a factor of 2 for their largest program. The mechanism of using some explicit metric to direct concolic search resembles ours, but the goal is very different, since they are trying to increase coverage by generating new different inputs whereas we are trying to improve localization by generating new similar inputs.

7. CONCLUSIONS

Statistical fault-localization techniques that analyze execution data from multiple tests [18, 19, 20, 2, 26] have been shown to be quite effective when a sufficiently large test suite is available. In this paper, we address the scenario where such a test suite is not available, and the user relies on test-generation techniques to create one. In particular, we consider several test-generation techniques based on combined concrete and symbolic (concolic) execution and compare the fault-localization effectiveness of the generated test suites. The techniques under consideration include a standard, undirected concolic execution algorithm, used in our previous work [4], which attempts to maximize branch coverage. Furthermore, we consider several variants of concolic execution that aim to maximize similarity between the execution characteristics of the generated tests and those of a given fault-exposing execution, according to a number of similarity criteria.

We implemented the techniques in Apollo, an automated tool that detects and localizes faults in PHP Web applications. We evaluated the test generation techniques by localizing 35 faults in four PHP
applications. The results that we presented in this paper show that a new, directed test-generation technique based on path-constraint similarity yields the smallest test suites with the same excellent fault-localization characteristics as test suites generated by other techniques. In particular, when compared to test generation based on the concolic execution algorithm of [4], which aims to maximize code coverage, our directed technique reduces test-suite size by 86.1% and test-suite generation time by 88.6%.

As part of future work, we plan to explore additional similarity metrics, and understand how they compare to the path-constraint and input similarity metrics presented in this paper.

8. REFERENCES


