Noname manuscript No.
(will be inserted by the editor)

Examining the Effectiveness of Using Concolic Analysis to Detect Code Clones

Daniel E. Krutz · Emad Shihab · Samuel A. Malachowsky

the date of receipt and acceptance should be inserted later

Abstract During the initial construction and subsequent maintenance of an application, duplication of functionality is common, whether intentional or otherwise. This replicated functionality, known as a code clone, has a diverse set of causes and can have moderate to severe adverse effects on a software project in a variety of ways. A code clone is simply defined as multiple code fragments that produce similar results when provided the same input. While there are an array of powerful clone detection tools, most suffer from a variety of drawbacks including the inability to accurately and reliably detect all four types of clones.

This paper presents a new method for detecting code clones based on concolic analysis, which uses a mixture of concrete and symbolic values to traverse a large and diverse portion of the source code. By performing concolic analysis on the targeted source code and then examining the holistic output for similarities, code clone candidates can be consistently identified. In order to measure the effectiveness of the technique, we performed a case study and found that concolic analysis was able to detect 92% of known clones in a controlled environment, including a significant number of harder-to-find type-4 clones. Concolic analysis was also able to consistently and effectively locate existing clones in several open source applications with an average precision of 83% and recall of 93%, both of which were significantly higher than existing clone detection tools to which they were compared to.

[rewrite abstract based on findings] [Clean up format based upon new publication]

Keywords Code Clones · Concolic Analysis · Software Engineering

D. Krutz
Department of Software Engineering, Rochester Institute of Technology, NY USA
E-mail: dxkvse@rit.edu

E. Shihab
E-mail: emad.shihab@rit.edu

S. Malachowsky
E-mail: samvse@rit.edu

1 Introduction

Software must continually change in order to keep up with user requirements, enhance its functionality, fix bugs, or repair security vulnerabilities. Prior work has shown that these code changes often results in cloned code for a variety of reasons. In many instances, developers knowingly duplicate functionality across the software system because of laziness or an unwillingness to refactor and retest the modified portion of the application. Careful developers, who know to avoid code clones, may not be aware that identical functionality exists in the system, unintentionally injecting clones into the application [2, 10, 21, 31]. Whatever the reason, clones continue to be extremely widespread in software development; estimates have shown that clones typically amount to between 5% and 30% of an application's source code [4, 24, 37].

Many previous works have stated that code clones are undesirable since they often lead to more bugs and make their remediation process more difficult and expensive [2, 4, 10, 32]. Clones may also substantially raise the maintenance costs associated with an application [18], the importance of which is highlighted by the fact that the maintenance phase of a project has been found to encompass between 40% and 90% of the total cost of a software project [6, 11, 12, 38, 40, 43]. Ultimately, unintentionally making inconsistently applied bug fixes to cloned code across a software system increases the likeliness of further system faults [9].

There are four types of code clones generally recognized by the research community. Type-1 clones are the simplest, representing identical code except for variations in whitespace, comments, and layout [5]. Type-2 clones are syntactically similar except for variations in identifiers and types. Type-3 clones are two code segments which are syntactically different due to altered or removed statements. Type-4 clones, the most difficult to detect, are two code segments which have considerable differences syntactically, but produce identical results when executed [8, 14].

To assist software practitioners in detecting and managing code clones, clone detection tools have been indispensable in detecting clone-related bugs and even security vulnerabilities in software systems [8]. Of the numerous clone detection tools, most have only been able to detect the simpler clones: type-1, type-2, and type-3. Type-4 clones, the most difficult to detect [35, 47], have, to the best of our knowledge, only two processes able to reliably detect them. MeCC, capable of reliably detecting type-4 clones, suffers from several drawbacks, including the ability to only analyze pre-processed C programs [23].

In this paper, we examine the effectiveness of using concolic analysis to detect code clones. Concolic analysis combines concrete and symbolic values in order to traverse all possible paths of an application (up to a given length). Traditionally used in software testing to find application faults [22, 25], concolic analysis forms the basis of a powerful clone detection tool because it only considers the functionality of the source code and not its syntactic properties. Because of this, elements that are challenging for existing clone detection systems such as comments and naming conventions do not affect concolic analysis and its detection of clones.

This research is innovative because, to our knowledge, no previous attempts have been made in using concolic analysis in clone discovery. Any technique which can

effectively discover all four types of code clones is important, since, at the present time, so few clone detection processes are able to do so.

Concolic Code Clone Detection (CCCD) is a fully functional tool that uses concolic analysis for clone detection. Concolic analysis is performed on the target application using CREST [1]. First, a Java component uses CTAGS [2] to break up the concolic output at the method level. Next, a comparison process uses the developed Levenshtein-distance-based measurement to evaluate the similarity between the concolic output files. Finally, a report displays the detected code clone candidates. This tool, installation instructions, and further details may be found on the project website [1] and was published in a previous work [28].

Our study will answer the following research questions:

RQ1: *What types of clones is concolic analysis effective at detecting?*
We find concolic analysis is able to detect all types of clones, in both, a controlled environment and in several open source applications. In a controlled environment using clones identified by previous research, concolic analysis was able to detect x% of type-1, type-2, and type-3 clones, and x% of type-4 clones. **[update values]**

RQ2: *How does concolic analysis based clone detection compare to other leading clone detection tools?*
While several existing methods are very innovative and successful at detecting a variety of code clones, we find that concolic analysis compares very favorably to these tools. Using manually identified clones in several open source systems, concolic analysis consistently discovered clones with a higher rate of precision, recall and F-score when compared to several leading existing clone detection tools. Concolic analysis averaged x% for precision, x% for recall and an x% F-Score while the next leading tool, Nicad, was only able to achieve scores of x%, x%, and x%, respectively. **[update values]**

The remainder of the paper is organized as follows. Section 2 describes how concolic analysis may be used to to detect software clones. Section 3 evaluates the ability of concolic analysis in identifying clones in relation to existing tools. Section 4 conveys interesting results from the research. Section 5 discusses related works in clone detection and concolic analysis. Section 6 details some threats to the findings of this work. Section 8 provides concluding remarks and future research directions for this work. **[update section]**

2 How Concolic Clone Detection Works

In explaining how concolic code clone detection works, a breakdown of the concept and illustration of it in action is needed. First, we will describe how concolic analysis is performed on two cloned methods. This will include the use of the Levenshtein distance algorithm in measuring the similarity of two sets of concolic output. Next, we will provide a motivating example using a cloned method which has been analyzed by several leading clone detection tools. Finally, we will briefly explain some of the

[1] https://github.com/jburnim/crest/
[2] http://ctags.sourceforge.net

Code Segment #1	Code Segment #2

`void sumProd(int n) {`
`  double sum=0.0;`
`  double prod =1.0;`
`  int i;`
`  for (i=1; i<=n; i++){`
`  sum=sum + i;`
`  prod = prod * i;`
`  foo2(sum, prod);`
`  }`
`}`
	`void sumProd2(int n) {`
	`  int sum=0; //C1`
	`  int prod =1;`
	`  int i;`
	`  for (i=1; i<=n; i++){`
	`  sum=sum + i;`
	`  prod = prod * i;`
	`  foo2(sum, prod);`
	`  }`
	`}`

Table 1: An Example of Type-2 clones from Roy

shortcomings of these tools and why concolic analysis was successful in finding the clones.

2.1 Concolic Clone Detection Technique

We will first provide an example of two code clones and then describe how concolic analysis is able to detect these clones. Two type-2 clones are shown in Table 1. These are derived from clones presented by Royet al. [36].

Concolic code clone detection is comprised of two primary phases. The first step is the generation of the concolic output on the target application. This may be done using an existing concolic analysis tool such as CREST [3], Java Path Finder (JPF) [4], or CATG [5]. An abbreviated example segment of concolic output is shown in Listing 1; the complete output is may be viewed on the project website [1]. The generated concolic output represents all executable paths that the software may take, and is broken into several *path conditions*. These conditions, which are specific to code segments, must be true in order for the application to follow a specified path. For example, if in order to follow a specific path of an *if* statement a boolean variable must be *true*, the contingency of the path condition would be that the variable be *true*. Otherwise, this path will not be traversed [39].

[3] http://code.google.com/p/crest/

[4] http://babelfish.arc.nasa.gov/trac/jpf/

[5] https://github.com/ksen007/janala2

Listing 1: Example Concolic Output

```
PC#=3
CONST_3>a_1_SYMINT[2]&&
CONST_2<=a_1_SYMINT[2]&&
CONST_1<=a_1_SYMINT[2]

PC#=2
CONST_2>a_1_SYMINT[1]&&
CONST_1<=a_1_SYMINT[1]

PC#=1
CONST_1>a_1_SYMINT[2]
```

In Listing 1, constant variable types are represented generically by "CONST" while the variable type integer is represented by a generic tag "SYMINT." Though not present above, other variable types are represented in a similar fashion in concolic output such as this. Actual variable names do not appear anywhere in the output and are irrelevant to the concolic analysis technique. When comparing the output from the type-2 clones in Table 1, one of the primary differences would be that *sum* would be defined as a *double* variable type in the first method while it would be an *int* for the second. This would create a small variation in the compared output. Once this concolic output is created, similar sections of output will be searched for and noted to be code clone candidates. Concolic analysis explores the possible paths that an application can take, with similar execution paths signifying analogous functionality and is thus are indicative of a code clone candidate. Clones in *dead code* or code that is unreachable via excecution paths will not be analyzed by concolic analysis, and therefore are not discoverable via concolic analysis. To demonstrate this functionality, clones as defined by Roy et al. [36] in Table 1 were analyzed using concolic analysis. A portion of the generated concolic output, demonstrated in Table 2, was then compared for variations. The only difference between the two sets of concolic output is the method name displayed in the first line of each file (an abbreviated listing is shown; full results are available on the project website).

In order to measure the similarity between sets of the concolic output, the Levenshtein distance measurement is used. This is the minimal number of characters that would need to be replaced to convert one string to another [3, 15]. As an example, if the strings "ABCD" and "BCDE" are measured, the Levenshtein distance would be 2, because "A" would need to be removed and "E" inserted into the first string to make them identical. This technique was selected for several reasons, but the main motivation related to the impracticality of other string similarity techniques in clone detection using concolic analysis. The Hamming technique, for example, may only be used with strings which are the same length [17, 34], and concolic output of even two very similar methods rarely yields output of identical length. Another possibility, the longest common subsequence technique, does not account for the substitution of values, only the addition and deletion of characters [30], which also proves problematic.

Because of the relative flexibility of the Levenshtein distance metric, it has proven to be especially well suited for this particular task in clone detection. Due in part to

Concolic Segment #1	Concolic Segment #2
sumProd1(a);	sumProd2(a)
PC # 3 = 3	PC # = 3
CONST_3 > a_1_SYMINT[2] &&	CONST_3 > a_1_SYMINT[2] &&
CONST_2 <= a_1_SYMINT[2] &&	CONST_2 <= a_1_SYMINT[2] &&
CONST_1 <= a_1_SYMINT[2]	CONST_1 <= a_1_SYMINT[2]
SPC # = 0	SPC # = 0
PC # = 2	PC # = 2
CONST_2 > a_1_SYMINT[1] &&	CONST_2 > a_1_SYMINT[1] &&
CONST_1 <= a_1_SYMINT[1]	CONST_1 <= a_1_SYMINT[1]
SPC # = 0	SPC # = 0
PC # = 1	PC # = 1
CONST_1 > a_1_SYMINT[-2]	CONST_1 > a_1_SYMINT[-2]
SPC # = 0	SPC # = 0

Table 2: Diff of Concolic Output

its ability to work with strings of different lengths and its restriction of upper and lower bounds in the calculated distances, normalization is achieved via Equation 1 below. The final Levenshtein score (ALV) is computed by dividing the Levenshtein distance between two files (LD) by the longest string length of the two strings being compared (LSL) and then multiplying by 100.

$$ALV = (LD/LSL) \times 100 \tag{1}$$

In order to evaluate the effectiveness of concolic analysis in detecting clones, we created two prototypes. The first was able to analyze source code written in C and utilized CREST in generating the necessary concolic output. The second was created using JPF and examines source code written in Java, though it was not as robust. This Java-based prototype was important because it demonstrated several key concepts, including the assertion that this technique of detecting clones is language agnostic, and that other concolic analysis tools may be used in support of this process. When properly implemented using an existing concolic analysis tool for java, we have no reason to doubt the results will not differ from those found using a c-based technology such as crest.*[Dan says: should this sentence be left in here?]* An overview of the entire clone detection technique is shown in Figure 1.

2.2 Preliminary Experiments

A simple initial analysis to determine the effectiveness of concolic analysis in code clone detection was done using a type-3 clone as defined by Roy et al. [36] and is shown in Table 3. The tools used for our analysis are described below.

[add more tools] [only compare tools that work at the method level?]*[Dan says: It seems like we are often compare tools at different levels, and this makes things apples to oranges]* **[talk about settings used or break this up into its own section]**

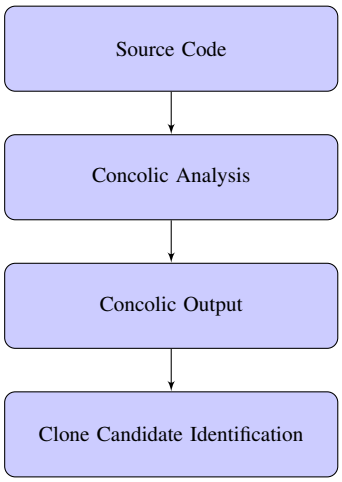

Fig. 1: Concolic Analysis

Code Segment #1

```
void sumProd1(int n) {
    double sum=0.0;
    double prod=1.0;
    int i;
    for (i=1; i<=n; i++){
      sum=sum + i;
      prod = prod * i;
      foo2(sum, prod);
    }
}
```

Code Segment #2

```
void sumProd3(int n) {
  double sum=0.0;
  double prod =1.0;
  int i;
  for (i=1; i<=n; i++){
    if (i %2 == 0){
      sum+= i;
    }
    prod = prod * i;
    foo2(sum, prod);
  }
}
```

Table 3: An Example of Type 3 Clones from Roy et al.

- Iclones, Nicad, Simcad

[Add in further description of the other tools used and if they could find this clone, and if they could not]

[Talk about how each tool did in finding this simple clone]

[update everything above with the current results]

3 Evaluation

[Provide Brief introduction here] In the following sections, concolic analysis for clone detection will be evaluated individually and against other leading existing clone detection tools.

3.1 Oracle

– Talk about the Murk/Bellon oracle used here. – Use [41] ”Evaluating Modern Clone Detection Tools” As a guide

3.2 Technique Comparisons

[Provide a better introduction to this]

The C-based applications were analyzed by concolic analysis using the Concolic Code Clone Detection (CCCD) tool, published in a previous work by the authors [28]. In the previous paper, we demonstrated the ability of concolic analysis to effectively discover all types of code clones in a small, controlled environment*[Dan says: Has this already been said enough?]*. We will build on these results and further compare concolic analysis against several leading existing clone detection tools using clones as defined in previous research by Krawitz [26] and Roy et al. [36].

RQ1: ***What types of clones is concolic analysis effective at detecting?*** **[check all of this data to make sure that it is up to date]**

The initial step of evaluating concolic analysis for code clone detection was to evaluate it against 4 clones defined by Krawitz, and 16 by Roy et al. These 20 defined clones were added to a single Java and C file. The results in Table 4 indicate the ability of concolic analysis to find a wide range of clones in a small, controlled environment in both C and Java applications. We used CCCD for the C code, and developed a small prototype based upon JPF for the Java code. *[Dan says: add more to this?]* Due to the small scale of this analysis, we did not evaluate concolic analysis for code clone detection against any existing tools and only wish to demonstrate that concolic analysis is capable of detecting all four types of clones in both C and Java.

Table 4: Concolic Analysis Finding Clones in Single Java Class

Language	T1	T2	T3	T4	Total
Java	5	6	6	6	23 (96%)
C	5	6	7	4	22 (92%)
Total Possible	5	6	7	6	24

Within the limited Java implementation, the concolic analysis based technique was able to detect 96% of all clones and the only clone which concolic analysis was unable to detect was a type-3 clone as defined by Roy et al. JPF, the implementation used by concolic code clone detection for Java, was unable to traverse all paths of this method for technical reasons, including its inability to perform analysis on several unsupported variable types (float, byte, and short). This limitation ultimately affects the concolic analysis clone identification process specifically when applied to Java. A similar C file containing the clones of Krawitz and Roy et al. was then examined

for clones. Concolic analysis was able to detect 92% of all clones. *[Dan says: state why this clone was not found?]*

The size of the examined functions did not have a significant impact on the ability of any of the examined processes in detecting clones. The only clones that concolic analysis was unable to detect were the type-4 clones as defined by Krawitz. In this clone example, a method has been refactored into two functionally similar methods. Two different concolic paths were generated for these methods, and thus the generated concolic output was not similar, so no clone code candidate was detected.

This small, controlled example demonstrates that concolic analysis for code clone detection is capable of detecting all four types of clones in both Java and C in a small, controlled environment.

RQ2: ***How does concolic analysis based clone detection compare to other leading clone detection tools?*** **[update all of this]**

[Talk about and use existing oracles, Bellon etc.....]

In order to evaluate the effectiveness of concolic analysis for code clone detection, we will use an oracle created by Murakami et al. [33]. We selected this oracle since it is an updated extension of of the oracle created by Bellon et al. [5], which has been used in a substantial amount of clone detection research **[cite]**. Unfortunately, this oracle does not contain any identified type-4 code clones, so we will not be able to evaluate these clones in this research.**[is this the best way to address this shortcoming?]** We chose this oracle over one which we created for a previous publication [27] since we wanted to remove as much potential bias as possible for this analysis.

[13]

3.3 Accuracy, Precision & Recall

In addition to their ability to find clones, precision and recall are important factors in evaluating clone detection tools [48]. The tool should not return too high of a rate of false positives, while it should also not miss a significant portion of code clones - striking an ideal balance between the two. In order to calculate accuracy, precision and recall, we used the data previously attained from running concolic analysis and existing tools against our oracle. To study the prediction accuracy, we built two multivariate logistic regression models: one that uses all of the metrics and one that uses a smaller set of statistically and minimally collinear metrics. The logistic regression models are designed predict the likelihood of a file being defect prone or otherwise. The output is given as a value between 0 and 1; we classified values above 0.5 as defect prone, with the remainder classified defect free. The classification results of the prediction models were stored in a confusion matrix, as shown in Table 5.

The performance of the prediction model is measured in four different ways. Values for each will range from 0 to 1, with a 1 being favorable:

1. **Precision:** Relates the number of files predicted *and* observed as defect prone to the number of files predicted as defect prone. It is calculated as $\frac{a}{a+b}$.
2. **Recall:** Relates the number of files predicted *and* observed as defect prone to the number of files that actually had defects. It is calculated as $\frac{a}{a+c}$.

Table 5: Confusion matrix

		True Class	
		Yes	No
Predicted	Yes	a	b
	No	c	d

3. **F-Score:** Considers precision *and* recall to measure the accuracy of a system. It is calculated as $2 \times (\frac{precision \times recall}{precision + recall})$.
4. **Accuracy:** Percentage of elements classified correctly. The highest attainable value is 1.0. It is calculated as $\frac{a+d}{a+b+c+d}$.

– Explain how I did the study – If I am running the tools myself, make sure to state why, how and what the configuration settings for the other tools were. - Take information from [45] –

Table 6 displays the precision, recall, F-score and accuracy for Nicad, MeCC, Simcad**[update the tools]**, and CCCD using a Levenshtein distance of 30. A complete listing of all results and tool types may be found in Table 8 in the appendix. *[Dan says: not sure if this table is still needed]*

Table 6: Precision, Recall, F-Score & Accuracy for Compared Techniques

Tool	Source Example	Precision	Recall	F-Score	Accuracy
xxxx	**xxxxxx**	x	x	x	x
	x	x	x	x	x
	x	x	x	x	x
	Avg.	**x**	**x**	**x**	**x**
xxxx	**xxxxxx**	x	x	x	x
	x	x	x	x	x
	x	x	x	x	x
	Avg.	**x**	**x**	**x**	**x**

All of the analyzed clone detection techniques were able to achieve a high rate of accuracy when examining open source applications. This is due to the relatively small number of clones that existed in these applications. While some of the tools fared better in specific areas, CCCD achieved overall better results overall. Nicad, MeCC, and Simcad all presented a lower precision and F-Score when searching for clones in P-SQL, possibly due the relatively large size of the analyzed subset of this application. When a statistically significant portion of this subset was created, many of the discovered clones were not selected for analysis, which helped lead to the smaller detection scores.

Concolic analysis has been shown to be a powerful clone detection method which is not only able to discover a wide range of clone types (including type-4), but is also able to find them with a high rate of precision, recall, F-Score, and accuracy. **[update this entire section with new results & tools]**

4 Discussion

During our analysis of concolic analysis for clone detection, we found several interesting areas that warrant further discussion. First, a discussion of the amount of time required to required to run the examined clone detection tools. Second, the effects that using various Levenshtein distance values have in determining clones has on the precision, recall, F-score, and accuracy values of concolic analysis for clone detection. Finally, a discussion on how the Levenshtein score between compared methods are relate to the likelihood of different types of clones found by concolic analysis.

4.1 Execution Times

One aspect of note in concolic analysis for clone detection is the amount of time required to search for clones on an application. Table 7 compares the execution time for xxxxxxxxxx, and concolic analysis in detection clones implemented using CCCD, on Murakami et al.'s oracle*[Dan says: correct format?]*. All comparisons were conducted on a Fedora 32-bit machine with a 2.5 GHz Intel Core 2 Duo Processor and 4 GB ram.

Table 7: Execution Times

Tool	Time (Seconds)
Tool1	x
Tool2	x

[Analyze these values here]

4.2 Levenshtein Distance

[This entire subsection will need to be re-written with future findings] Concolic analysis for clone detection uses the Levenshtein distance algorithm to measure the similarity of two sets of concolic output, with sets of concolic output with a specific similarity scores marked as potential clones. Our first step in determining the most appropriate Levenshtein value was to use was to produce concolic output from the control, Apache, Python, and PostgreSQL applications, compare them against one another using a round robin methodology, then to record the Levenshtein distance scores. We then evaluated this against our clone oracle using Levenshtein scores of 0-40 with 5 point increments as a basis for determining clones. To obtain the optimal number, we compared the precision, recall, F-score, and accuracy scores of each increment and found that for all of the codebases, the Levenshtein value of 30 produced the highest rates.

We combined the precision, recall, F-score and accuracy values of all four codebases and placed them into a chart to better visualize the effects of using the different

Levenshtein scores to determine clones. Figure 2 displays the results of various Levenshtein values in discovering clones in a single class as defined by Krawitz and Roy et al. Figure 3 shows a similar analysis using our generated clone oracle using an aggregate of values from Apache, Python, and PostgreSQL.

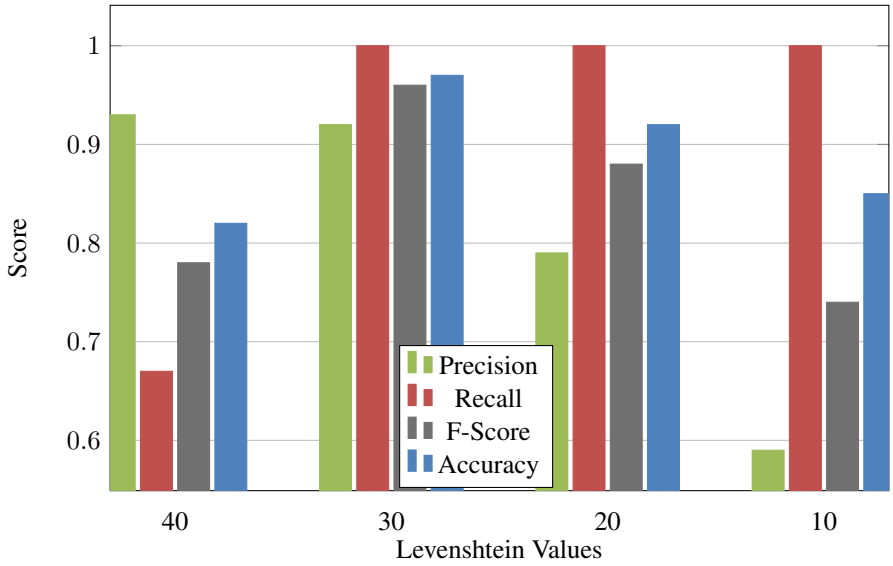

Fig. 2: Levenshtein Impact In Control Environment

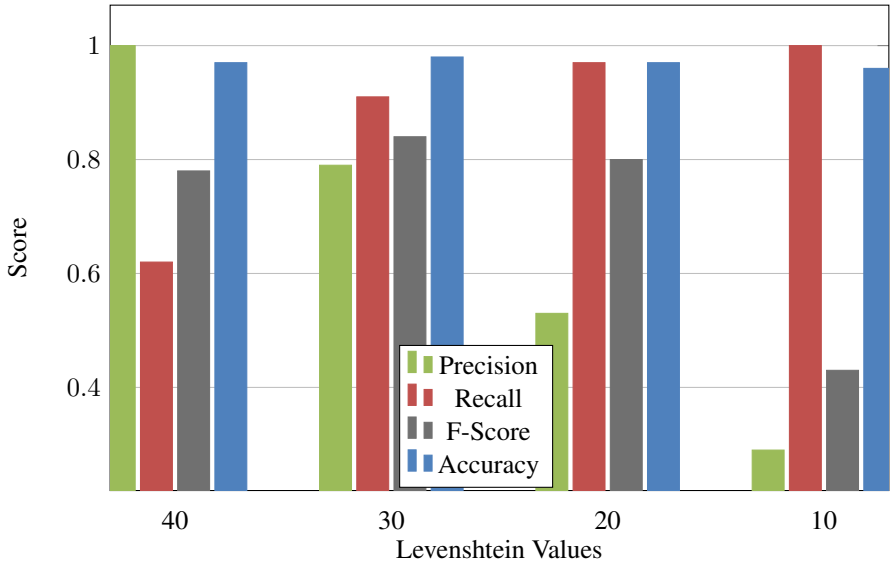

Fig. 3: Levenshtein Impact In Open Source Applications

In the controlled environment with the clones from Krawitz and Roy et al., accuracy, recall, and precision values were collectively highest when a Levenshtein score of 30 was used. Similar results were observed when analyzing the open source applications for clones. A higher Levenshtein score is likely to discover more clones, but will also lead to more false positives, creating high precision but lower recall. Conversely, a low Levenshtein score will find fewer actual clones, but also have less false positives leading to high recall, but low precision. This is because a higher Levenshtein score means that the similarity threshold for noting cloned items will be reduced. A user could select different Levenshtein values depending on their desired levels of precision, recall, F-Score and accuracy.

These findings are important for several reasons. In both examples, the most appropriate Levenshtein score for attaining the highest accuracy, recall and precision was found to be 30, and this value has been used throughout our analysis. Additionally, these findings are indicative of those that future researchers may expect when using concolic analysis to find clones in their respective applications. In certain situations, researchers may wish to increase the recall or precision of their clone discovery technique, using resulting data to seek the most appropriate values.

4.3 Calculated Levenshtein Distance & Clone Types

[This entire subsection will need to be re-written with future findings] An interesting discovery is how calculated Levenshtein distance may show indication of clone type. In general, a lower calculated distance score is indicative of a closer level of similarity between two compared items. Overall, the average similarity score for all types of clones was 21.77, while non-clones averaged 71.19. A smaller variation would likely lead to many more errors in the clone detection process. What is most interesting is that the average Levenshtein score between two compared methods may not only indicate if they are clones, but also may help to indicate what type of clone they are as well. These values were calculated by recording the Levenshtein distance for each type of identified clone in the oracle, along with items which were not clones. The scores were then averaged together; type-1 clones were found to have the lowest average, with more complicated clones showing progressively higher values. Final results are displayed in Figure 4.

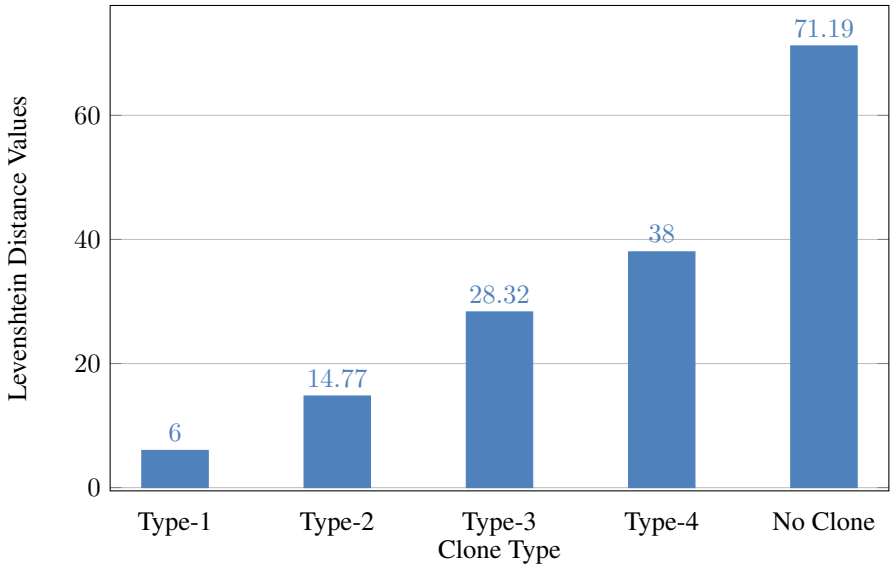

Fig. 4: Levenshtein Impact on Clone Types

5 Related Works

[Beef up the related works section]

There are numerous clone detection tools which utilize a variety of methods for discovering clones including text, lexical, semantic, symbolic and behavioral based approaches [36]. Only two , however, are known to be able to reliably detect type-4 clones.

MeCC discovers clones based on the ability to compare a program's abstract memory states. While this work was successful in finding type-4 clones, there are several areas for improvement such as its limitation in analyzing pre-processed C programs and an excessive clone detection time, likely caused by the exploration of an unreasonably large number of possible program paths [23]. Krawitz [26] proposed a clone discovery technique based on functional analysis which was shown to detect clones of all types, but was never implemented into a reasonably functional tool. This technique also requires a substantial amount of random data, which may be a difficult and time consuming process to produce.

CCFinder is a powerful clone detection tool which has been extensively referenced in existing clone detection research [7, 16, 20]. The process used first transforms the input source text and then performs a token-by-token comparison in order to discover clones [19]. While this tool has been evaluated in a significant amount of previous research, all links to download the application from its website appear to be dead, so it was unable to be evaluated in our work.*[Dan says: Updated this sentence]*

The most prominent area that concolic analysis has been applied to thus far is software testing, specifically for dynamic test input generation, test case generation, and bug detection [25, 39, 46]. Several tools exist for performing concolic analysis, including Crest [6], Java Path Finder [7], CUTE [39], and Pex [8].

Tempero [42] described a collection of 1.3M method-level-clone-pairs from 109 different systems. The goal of this work was to create a similar data set for clone research. While this work was profound, much of the data has a low level of confidence and requires further work and analysis. Additionally, the clones are only from Java-based systems.

Lavoie and Merlo [29] created an clone oracle set containing type-3 clones using the Levensthein metric. There was no mention of type-4 clones being created as part of this oracle, and the provided oracle only contained Java code. Krawitz [26] and Roy et al. [36] both defined clones of all four types in a small controlled environment. However, these works only specified a small number of clones which were artificially created.

SeByte Scorpio CCFinderX, ConQat, CPD, ctCompare, Deckard, Duplo, iClones, Simian - mention that these other tools exist

6 Threats to validity

There are certain threats to the validity of our results. First, our results were only run on Java and C. We do not believe the results would significantly differ if concolic clone detection was run in different languages, but without verification it is impossible to tell for certain. Concolic analysis only executes the functional aspects of an application, meaning that it will not be able to detect clones in non-functional portions of the software. Second, this technique is limited by the concolic analysis tools available for use, and while these tools continue to improve and are robust, they are not perfect. In some cases they are unable to traverse various portions of an application or are incapable of recognizing segments of the application for technical reasons. This inhibits the clone detection process for these portions of the application. Finally, the followed path conditions depend upon the control flow graph and its predicates, meaning that concolic analysis for clone detection is still dependent upon its implementation. While it is less dependent than syntax or token based clone detectors, many code instances of identical semantics or different implementations will not be detected by concolic analysis for clone detection.

A significant portion of this study was based off previous research by Krawitz [26], Roy et al. [36] and Kim et al. [23]. Therefore, our results depend to a certain extent on the benchmarks provided by the aforementioned prior work. Manually finding type-4 clones in source code is extremely difficult and there is only one existing method known to reliably find type-4 clones. This makes it very difficult to test out a new mechanism in finding these clones specifically because there are very few benchmarks to be evaluated against. We are confident that concolic analysis is able

[6] http://code.google.com/p/crest/

[7] http://babelfish.arc.nasa.gov/trac/jpf/

[8] http://research.microsoft.com/en-us/projects/pex/

to discover type-4 clones as is exemplified by our evaluation using the small sample oracle largely derived from Krawitz [26] and Roy et al. [36]. Unfortunately, since type-4 clones are very hard to manually identify and are only found by one existing tool, generating an accurate evaluation of a new technique in its ability to accurately identify type-4 clones is very difficult.

While we did our best to manually identify and classify clones using several people, and previous research has demonstrated the difficulty and problems with manually identifying and classifying code clones [44]. This indicates that other researchers may disagree with many of the clones identified and how they were classified in our work. This is a problem which is not at all unique to our work and is one that hinders other research as well [29].

There are also numerous clone detection tools that detect clones in numerous different ways. While we were able to compare concolic analysis to several other leading detection processes, it is unreasonable to attempt to compare them to all known techniques. Many clone detection tools have adjustable inputs which may be altered to determine the size of the methods examined for clones, along with the similarity score needed to determine if two methods are defined as clones. While we did our best to use the most appropriate input settings for each tool, it is quite possible that more appropriate settings could have been selected to yield more accurate results. When we were unsure of the most appropriate setting, we chose to use the defaults for each tool.

7 Future Work

[Dan says: Show that concolic analysis COULD detect clones at the sub method level, but technique needs more pruning work]

[Add in other work to be done]

8 Conclusion

[Work on the conclusion] In the future, we plan on applying the techniques described in this paper to other areas of computing research. One area we will research is how type-4 clones affect software development including how problematic they actually are. While existing research has examined many of the effects that simpler clones have on the software development lifecyle [18], to our knowledge no work has been done to analyze the effect of type-4 clones specifically in that context.

Concolic analysis has only been evaluated in finding clones at the method level. However, many clones occur as only portions of methods, or across numerous methods. Future work is needed to determine the ability of the proposed technique in discovering clones at a more granular level or across methods.

Concolic Code Clone Detection represents a new and powerful clone detection technique. Concolic analysis executes various paths of an application. Similar application paths represents functional similarity, and thus a code clone candidate. When compared to leading existing clone detection tools, concolic analysis was able to

more accurately and reliably identify all types of clones. The proposed clone detection technique is innovative because it not only represents the first known concolic-based clone detection technique, but is also one of only two known processes which are able to reliably detect type-4 clones.

Project Website: A complete implementation and more in depth results regarding this study may be found at the project website [9]

References

1. Cccd concolic code clone detection. URL http://www.se.rit.edu/~dkrutz/CCCD/
2. Baker, B.S.: On finding duplication and near-duplication in large software systems. In: Proceedings of the Second Working Conference on Reverse Engineering, WCRE '95, pp. 86–. IEEE Computer Society, Washington, DC, USA (1995). URL http://dl.acm.org/citation.cfm?id=832303.836911
3. Bard, G.V.: Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. In: L. Brankovic, C. Steketee (eds.) Fifth Australasian Information Security Workshop (Privacy Enhancing Technologies) (AISW 2007), *CRPIT*, vol. 68, pp. 117–124. ACS, Ballarat, Australia (2007)
4. Baxter, I.D., Yahin, A., Moura, L., Sant'Anna, M., Bier, L.: Clone detection using abstract syntax trees. In: Proceedings of the International Conference on Software Maintenance, ICSM '98, pp. 368–. IEEE Computer Society, Washington, DC, USA (1998). URL http://dl.acm.org/citation.cfm?id=850947.853341
5. Bellon, S., Koschke, R., Antoniol, G., Krinke, J., Merlo, E.: Comparison and evaluation of clone detection tools. Software Engineering, IEEE Transactions on **33**(9), 577–591 (2007). DOI 10.1109/TSE.2007.70725
6. Boehm, B., Basili, V.R.: Software defect reduction top 10 list. Computer **34**(1), 135–137 (2001). DOI 10.1109/2.962984. URL http://dx.doi.org/10.1109/2.962984
7. Choi, E., Yoshida, N., Ishio, T., Inoue, K., Sano, T.: Extracting code clones for refactoring using combinations of clone metrics. In: Proceedings of the 5th International Workshop on Software Clones, IWSC '11, pp. 7–13. ACM, New York, NY, USA (2011). DOI 10.1145/1985404.1985407. URL http://doi.acm.org/10.1145/1985404.1985407
8. Dang, Y., Zhang, D., Ge, S., Chu, C., Qiu, Y., Xie, T.: Xiao: tuning code clones at hands of engineers in practice. In: Proceedings of the 28th Annual Computer Security Applications Conference, ACSAC '12, pp. 369–378. ACM, New York, NY, USA (2012). DOI 10.1145/2420950.2421004. URL http://doi.acm.org/10.1145/2420950.2421004
9. Deissenboeck, F., Hummel, B., Juergens, E.: Code clone detection in practice. In: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 2, ICSE '10, pp. 499–500. ACM, New York, NY, USA (2010). DOI 10.1145/1810295.1810449. URL http://doi.acm.org/10.1145/1810295.1810449

[9] http://www.se.rit.edu/~dkrutz/CCCD/

10. Duala-Ekoko, E., Robillard, M.P.: Clone region descriptors: Representing and tracking duplication in source code. ACM Trans. Softw. Eng. Methodol. **20**(1), 3:1–3:31 (2010). DOI 10.1145/1767751.1767754. URL http://doi.acm.org/10.1145/1767751.1767754
11. Ducasse, S., Rieger, M., Demeyer, S.: A language independent approach for detecting duplicated code. In: Proceedings of the IEEE International Conference on Software Maintenance, ICSM '99, pp. 109–. IEEE Computer Society, Washington, DC, USA (1999). URL http://dl.acm.org/citation.cfm?id=519621.853389
12. Erlikh, L.: Leveraging legacy system dollars for e-business. IT Professional **2**(3), 17–23 (2000). DOI 10.1109/6294.846201. URL http://dx.doi.org/10.1109/6294.846201
13. Falke, R., Frenzel, P., Koschke, R.: Empirical evaluation of clone detection using syntax suffix trees. Empirical Softw. Engg. **13**(6), 601–643 (2008). DOI 10.1007/s10664-008-9073-9. URL http://dx.doi.org/10.1007/s10664-008-9073-9
14. Gold, N., Krinke, J., Harman, M., Binkley, D.: Issues in clone classification for dataflow languages. In: Proceedings of the 4th International Workshop on Software Clones, IWSC '10, pp. 83–84. ACM, New York, NY, USA (2010). DOI 10.1145/1808901.1808916. URL http://doi.acm.org/10.1145/1808901.1808916
15. Greenhill, S.J.: Levenshtein distances fail to identify language relationships accurately. Comput. Linguist. **37**(4), 689–698 (2011). DOI 10.1162/COLI_a_00073. URL http://dx.doi.org/10.1162/COLI_a_00073
16. Hotta, K., Sano, Y., Higo, Y., Kusumoto, S.: Is duplicate code more frequently modified than non-duplicate code in software evolution?: an empirical study on open source software. In: Proceedings of the Joint ERCIM Workshop on Software Evolution (EVOL) and International Workshop on Principles of Software Evolution (IWPSE), IWPSE-EVOL '10, pp. 73–82. ACM, New York, NY, USA (2010). DOI 10.1145/1862372.1862390. URL http://doi.acm.org/10.1145/1862372.1862390
17. Jain, M., Benmokhtar, R., Jégou, H., Gros, P.: Hamming embedding similarity-based image classification. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12, pp. 19:1–19:8. ACM, New York, NY, USA (2012). DOI 10.1145/2324796.2324820. URL http://doi.acm.org/10.1145/2324796.2324820
18. Juergens, E., Deissenboeck, F., Hummel, B., Wagner, S.: Do code clones matter? In: Proceedings of the 31st International Conference on Software Engineering, ICSE '09, pp. 485–495. IEEE Computer Society, Washington, DC, USA (2009). DOI 10.1109/ICSE.2009.5070547. URL http://dx.doi.org/10.1109/ICSE.2009.5070547
19. Kamiya, T., Kusumoto, S., Inoue, K.: Ccfinder: a multilinguistic token-based code clone detection system for large scale source code. IEEE Trans. Softw. Eng. **28**(7), 654–670 (2002). DOI 10.1109/TSE.2002.1019480. URL http://dx.doi.org/10.1109/TSE.2002.1019480
20. Kamiya, T., Ohata, F., Kondou, K., Kusumoto, S., Inoue, K.: Maintenance support tools for java programs: Ccfinder and jaat. In: Proceedings of the 23rd International Conference on Software Engineering, ICSE '01, pp. 837–838. IEEE Computer Society, Washington, DC, USA (2001). URL http://dl.acm.org/

citation.cfm?id=381473.381749
21. Kapser, C.J., Godfrey, M.W.: Supporting the analysis of clones in software systems: Research articles. J. Softw. Maint. Evol. **18**(2), 61–82 (2006). DOI 10.1002/smr.v18:2. URL http://dx.doi.org/10.1002/smr.v18:2
22. Kiezun, A., Ganesh, V., Artzi, S., Guo, P.J., Hooimeijer, P., Ernst, M.D.: Hampi: A solver for word equations over strings, regular expressions, and context-free grammars. ACM Trans. Softw. Eng. Methodol. **21**(4), 25:1–25:28 (2013). DOI 10.1145/2377656.2377662. URL http://doi.acm.org/10.1145/2377656.2377662
23. Kim, H., Jung, Y., Kim, S., Yi, K.: Mecc: memory comparison-based clone detector. In: Proceedings of the 33rd International Conference on Software Engineering, ICSE '11, pp. 301–310. ACM, New York, NY, USA (2011). DOI 10.1145/1985793.1985835. URL http://doi.acm.org/10.1145/1985793.1985835
24. Kim, M., Sazawal, V., Notkin, D., Murphy, G.: An empirical study of code clone genealogies. SIGSOFT Softw. Eng. Notes **30**(5), 187–196 (2005). DOI 10.1145/1095430.1081737. URL http://doi.acm.org/10.1145/1095430.1081737
25. Kim, Y., Kim, M., Kim, Y., Jang, Y.: Industrial application of concolic testing approach: a case study on libexif by using crest-bv and klee. In: Proceedings of the 2012 International Conference on Software Engineering, ICSE 2012, pp. 1143–1152. IEEE Press, Piscataway, NJ, USA (2012). URL http://dl.acm.org/citation.cfm?id=2337223.2337373
26. Krawitz, R.M.: Code clone discovery based on functional behavior. Ph.D. thesis, Nova Southeastern University (2012)
27. Krutz, D.E., Le, W.: A code clone oracle. In: Proceedings of the 11th Working Conference on Mining Software Repositories, MSR 2014, pp. 388–391. ACM, New York, NY, USA (2014). DOI 10.1145/2597073.2597127. URL http://doi.acm.org/10.1145/2597073.2597127
28. Krutz, D.E., Shihab, E.: Cccd: Concolic code clone detection. In: Reverse Engineering (WCRE), 2013 20th Working Conference on (2013). DOI 10.1109/WCRE.2012.60
29. Lavoie, T., Merlo, E.: Automated type-3 clone oracle using levenshtein metric. In: Proceedings of the 5th International Workshop on Software Clones, IWSC '11, pp. 34–40. ACM, New York, NY, USA (2011). DOI 10.1145/1985404.1985411. URL http://doi.acm.org/10.1145/1985404.1985411
30. Li, R.: A space efficient algorithm for the constrained heaviest common subsequence problem. In: Proceedings of the 46th Annual Southeast Regional Conference on XX, ACM-SE 46, pp. 226–230. ACM, New York, NY, USA (2008). DOI 10.1145/1593105.1593164. URL http://doi.acm.org/10.1145/1593105.1593164
31. Li, Z., Lu, S., Myagmar, S., Zhou, Y.: Cp-miner: Finding copy-paste and related bugs in large-scale software code. IEEE Trans. Softw. Eng. **32**(3), 176–192 (2006). DOI 10.1109/TSE.2006.28. URL http://dx.doi.org/10.1109/TSE.2006.28
32. Mondal, M., Roy, C.K., Schneider, K.A.: An empirical study on clone stability. SIGAPP Appl. Comput. Rev. **12**(3), 20–36 (2012). DOI 10.1145/2387358.2387360. URL http://doi.acm.org/10.1145/2387358.2387360
33. Murakami, H., Higo, Y., Kusumoto, S.: A dataset of clone references with gaps. In: Proceedings of the 11th Working Conference on Mining Software Repos-

itories, MSR 2014, pp. 412–415. ACM, New York, NY, USA (2014). DOI 10.1145/2597073.2597133. URL http://doi.acm.org/10.1145/2597073.2597133
34. Ros, M., Sutton, P.: A post-compilation register reassignment technique for improving hamming distance code compression. In: Proceedings of the 2005 international conference on Compilers, architectures and synthesis for embedded systems, CASES '05, pp. 97–104. ACM, New York, NY, USA (2005). DOI 10.1145/1086297.1086311. URL http://doi.acm.org/10.1145/1086297.1086311
35. Roy, C.K., Cordy, J.R.: A survey on software clone detection research. SCHOOL OF COMPUTING TR 2007-541, QUEENS UNIVERSITY **115** (2007)
36. Roy, C.K., Cordy, J.R., Koschke, R.: Comparison and evaluation of code clone detection techniques and tools: A qualitative approach. Sci. Comput. Program. **74**(7), 470–495 (2009). DOI 10.1016/j.scico.2009.02.007. URL http://dx.doi.org/10.1016/j.scico.2009.02.007
37. Schulze, S., Apel, S., Kästner, C.: Code clones in feature-oriented software product lines. SIGPLAN Not. **46**(2), 103–112 (2010). DOI 10.1145/1942788.1868310. URL http://doi.acm.org/10.1145/1942788.1868310
38. Seaman, C.B.: Software maintenance: Concepts and practice. Journal of Software Maintenance and Evolution: Research and Practice **13**(2), 143–147 (2001). DOI 10.1002/smr.225. URL http://dx.doi.org/10.1002/smr.225
39. Sen, K., Marinov, D., Agha, G.: Cute: a concolic unit testing engine for c. In: Proceedings of the 10th European software engineering conference held jointly with 13th ACM SIGSOFT international symposium on Foundations of software engineering, ESEC/FSE-13, pp. 263–272. ACM, New York, NY, USA (2005). DOI 10.1145/1081706.1081750. URL http://doi.acm.org/10.1145/1081706.1081750
40. Shukla, R., Misra, A.K.: Estimating software maintenance effort: a neural network approach. In: Proceedings of the 1st India software engineering conference, ISEC '08, pp. 107–112. ACM, New York, NY, USA (2008). DOI 10.1145/1342211.1342232. URL http://doi.acm.org/10.1145/1342211.1342232
41. Svajlenko, J., Roy, C.K.: Evaluating modern clone detection tools. Proc. ICSME (2014)
42. Tempero, E.: Towards a curated collection of code clones. In: Proc. IWSC, pp. 53–59. IEEE (2013)
43. Ueda, Y., Kamiya, T., Kusumoto, S., Inoue, K.: Gemini: Maintenance support environment based on code clone analysis. In: Proceedings of the 8th International Symposium on Software Metrics, METRICS '02, pp. 67–. IEEE Computer Society, Washington, DC, USA (2002). URL http://dl.acm.org/citation.cfm?id=823457.824039
44. Walenstein, A., Jyoti, N., Li, J., Yang, Y., Lakhotia, A.: Problems creating task-relevant clone detection reference data. In: Proceedings of the 10th Working Conference on Reverse Engineering, WCRE '03, pp. 285–. IEEE Computer Society, Washington, DC, USA (2003). URL http://dl.acm.org/citation.cfm?id=950792.951349
45. Wang, T., Harman, M., Jia, Y., Krinke, J.: Searching for better configurations: A rigorous approach to clone evaluation. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2013, pp. 455–465. ACM, New York, NY, USA (2013). DOI 10.1145/2491411.2491420. URL

http://doi.acm.org/10.1145/2491411.2491420
46. Wassermann, G., Yu, D., Chander, A., Dhurjati, D., Inamura, H., Su, Z.: Dynamic test input generation for web applications. In: Proceedings of the 2008 international symposium on Software testing and analysis, ISSTA '08, pp. 249–260. ACM, New York, NY, USA (2008). DOI 10.1145/1390630.1390661. URL http://doi.acm.org/10.1145/1390630.1390661
47. Yuan, Y., Guo, Y.: Cmcd: Count matrix based code clone detection. In: Proceedings of the 2011 18th Asia-Pacific Software Engineering Conference, APSEC '11, pp. 250–257. IEEE Computer Society, Washington, DC, USA (2011). DOI 10.1109/APSEC.2011.13. URL http://dx.doi.org/10.1109/APSEC.2011.13
48. Zibran, M.F., Roy, C.K.: Ide-based real-time focused search for near-miss clones. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12, pp. 1235–1242. ACM, New York, NY, USA (2012). DOI 10.1145/2231936.2231970. URL http://doi.acm.org/10.1145/2231936.2231970

A Appendix

We discussed precision, recall, F-score and accuracy in Section 3.3, where we omitted the results for several clone detection tools due to the inability to perform many of the calculations for specific tools since they were unable to find any clones. We present the complete results in Table 8, with incalculable values represented with an *x*.

Table 8: Precision, Recall, F-Score & Accuracy for Each Tool

Tool	Source Example	Precision	Recall	F-Score	Accuracy
Mecc-4	**Control**	0	x	x	.64
	Apache	x	.3	.46	.88
	P-SQL	x	x	x	.94
	Python	x	x	x	.97
CloneDR	**Control**	.78	1	.88	.67
	Apache	0	x	x	.95
	P-SQL	0	x	x	.95
	Python	0	x	x	.97
Simian	**Control**	.14	1	.26	.69
	Apache	.06	.33	.1	.94
	P-SQL	0	x	x	.94
	Python	0	x	?	.97
Nicad	**Control**	.59	1	.74	.85
	Apache	.77	1	.87	.96
	P-SQL	.03	1	.06	.95
	Python	.67	1	.8	.97
CCCD	**Control**	.92	1	.96	.97
	Apache	1	.9	.95	.99
	P-SQL	.73	.96	.83	.98
	Python	.67	.84	.75	.98

[entire table needs to be updated]