Highlights

Java Decompiler Diversity and its Application to Meta-decompilation

Nicolas Harrand,César Soto-Valero,Martin Monperrus,Benoit Baudry

- A general framework to automatically evaluate decompilation results based on the concept of equivalence modulo inputs.
- An empirical assessment of eight Java decompilers, identifying the strengths and limitations of bytecode decompilation in practice.
- A novel approach to decompilation, called meta-decompilation, which outperforms the results of individual Java decompilers.

Java Decompiler Diversity and its Application to Meta-decompilation

Nicolas Harrand*, César Soto-Valero, Martin Monperrus and Benoit Baudry

KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Keywords:
Java bytecode
decompilation
reverse engineering
source code analysis

ABSTRACT

During compilation from Java source code to bytecode, some information is irreversibly lost. In other words, compilation and decompilation of Java code is not symmetric. Consequently, the decompilation process, which aims at producing source code from bytecode, must establish some strategies to reconstruct the information that has been lost. Modern Java decompilers tend to use distinct strategies to achieve proper decompilation. In this work, we hypothesize that the diverse ways in which bytecode can be decompiled has a direct impact on the quality of the source code produced by decompilers.

In this paper, we study the effectiveness of eight Java decompilers with respect to three quality indicators: syntactic correctness, syntactic distortion and semantic equivalence modulo inputs. Our results show that no single modern decompiler is able to correctly handle the variety of bytecode structures coming from real-world programs. Even the highest ranking decompiler in this study produces syntactically correct output for 84% of classes of our dataset and semantically equivalent code output for 78% of classes. Furthermore we demonstrate that each decompiler correctly handles a different set of bytecode classes.

Second, we build a new decompiler called Arlecchino that leverages the diversity of existing decompilers. To do so, we merge partial decompilation into a new one based on compilation errors. Arlecchino handles 37.6% of bytecode classes that were previously handled by no decompiler. We publish the sources of this new bytecode decompiler.

1. Introduction

In the Java programming language, source code is compiled into an intermediate stack-based representation known as bytecode, which is interpreted by the Java Virtual Machine (JVM). In the process of translating source code to bytecode, the compiler performs various analyses. Even if most optimizations are typically performed at runtime by the just-in-time (JIT) compiler, several pieces of information residing in the original source code are already not present in the bytecode anymore due to compiler optimization [20]. For example the structure of loops is altered and local variable names may be modified [11].

Decompilation is the inverse process, it consists in transforming the bytecode instructions into source code [23]. Decompilation can be done with several goals in mind. First, it can be used to help developers understand the code of the libraries they use. This is why Java IDEs such as IntelliJ and Eclipse include built-in decompilers to help developers analyze the third-party classes for which the source code is not available. In this case, the readability of the decompiled code is paramount. Second, decompilation may be a preliminary step before another compilation pass, for example with a different compiler. In this case, the main goal is that the decompiled code is syntactically and grammatically correct and can be recompiled. Some other applications of decompilation with slightly different criteria include clone detection [25], malware analysis [36, 4] and software archaeology [26].

Overall, the ideal decompiler is one that transforms all inputs into source code that faithfully reflects the original code: the decompiled code 1) can be recompiled with a Java compiler and 2) behaves the same as the original program. However, previous studies having compared Java decompilers [9, 17] found that this ideal Java decompiler does not exist, because of the irreversible data loss that happens during compilation. In this papers, we perform a comprehensive assessment of three aspects of decompilation: the syntactic correctness of the decompiled code (the decompiled code can recompile); the semantic equivalence with the original source (the decompiled code passes all tests); the syntactic similarity to the original source (the decompiled source looks like the original). We evaluate eight recent and notable decompilers on 2041 Java classes, making this study one order of magnitude larger than the related work [9, 17].

Next, we isolate a subset of 157 Java classes that no state-of-the-art decompiler can correctly handle. The reasons for which each decompiler fails are diverse, and the error contained in the decompiled sources for these classes are not always located at the same place. This raises the opportunity to merge the results of several incorrect decompiled sources to produce a version that can be recompiled. We coin this process **meta-decompilation**. Meta-decompilation is a novel approach for decompilation: 1) it leverages the natural diversity of existing decompilers by merging the results of different decompilers 2) it is able to provide decompiled sources for classes that no decompiler in isolation can handle.

Our results have important implications: 1) for all users of decompilation, our paper shows significant differences between decompilers and provide well-founded empirical evidence to choose the best ones; 2) for researchers in decompilation, our results shows that the problem is not solved;

*Corresponding author

harrand@kth.se (N. Harrand); cesarsv@kth.se (C. Soto-Valero); martin.monperrus@csc.kth.se (M. Monperrus); baudry@kth.se (B. Baudry)

ORCID(s): 0000-0002-2491-2771 (N. Harrand); 0000-0003-0541-6411 (C. Soto-Valero); 0000-0003-3505-3383 (M. Monperrus); 0000-0002-4015-4640 (B. Baudry)

3) for authors of decompilers, our experiments have identified bugs in their decompilers (3 have already been fixed, and counting) and our methodology of semantic equivalence modulo inputs can be embedded in the QA process of all decompilers in the world.

In summary, this paper makes the following contributions:

- an empirical comparison of eight Java decompilers based on 2041 real-world Java classes, tested by 25019 test cases, identifying the key strengths and limitations of bytecode decompilation;
- a novel approach to decompilation, called meta decompilation, that leverages decompilers diversity to improve decompilation effectiveness;
- a tool and a dataset for future research on Java decompilers publicly available at `https://github.com/castor-software/decompilercmp`

2. Motivating Example

In this section, we present an example drawn from the Apache `commons-codec` library. We wish to illustrate information loss during compilation of Java source code, as well as the different strategies that bytecode decompilers adopt to cope with this loss when they generate source code from bytecode. Listing 1 shows the original source code of the utility class `org.apache.commons.codec.net.Utils`, while Listing 2 shows an excerpt of the bytecode produced by the standard *javac* compiler.[1] Here, we omit the constant pool as well as the table of local variables and replace references towards these tables with comments to save space and make the bytecode more human readable.

As mentioned, the key challenge of decompilation resides in the many ways in which information is lost during compilation. Consequently, Java decompilers need to make several assumptions when interpreting bytecode instructions, which can also be generated in different ways. To illustrate this phenomenon, Listing 3 and Listing 4 show the Java sources produced by the Fernflower and Dava decompilers when interpreting the bytecode of Listing 2. In both cases, the decompilation produces correct Java code (i.e., recompilable) with the same functionality as the input bytecode. Notice that Fernflower guesses that the series of `StringBuilder` (bytecode instruction 23 to 27) calls is the compiler's way of translating string concatenation and is able to revert it. On the contrary, the Dava decompiler does not reverse this transformation. As we can notice, the decompiled sources are different from the original in at least three points:

- In the original sources, the local variable *i* was `final`, but *javac* lost this information during compilation.

[1] There are various Java compilers available, notably Oracle *javac* and Eclipse *ecj*, which can produce different bytecode for the same Java input.

```java
class Utils {
   private static final int RADIX = 16;
   static int digit16(final byte b) throws
        DecoderException {
      final int i = Character.digit((char) b, RADIX);
      if (i == -1) {
         throw new DecoderException("Invalid URL
              encoding: not a valid digit (radix " +
              RADIX + "): " + b);
      }
      return i;
   }
}
```

Listing 1: Source code of Java class correspondig to `org.apache.commons.codec.net.Utils`.

```
class org.apache.commons.codec.net.Utils {
  static int digit16(byte) throws
       org.apache.commons.codec.DecoderException;
        0: ILOAD_0 //Parameter byte b
        1: I2C
        2: BIPUSH 16
        4: INVOKESTATIC #19 //Character.digit:(CI)I
        7: ISTORE_1 //Variable int i
        8: ILOAD_1
        9: ICONST_m1
       10: IF_ICMPNE 37
//org/apache/commons/codec/DecoderException
       13: NEW #17
       16: DUP
       17: NEW #25 //java/lang/StringBuilder
       20: DUP
//"Invalid URL encoding: not a valid digit (radix
     16):"
       21: LDC #27
//StringBuilder."<init>":(Ljava/lang/String;)V
       23: INVOKESPECIAL #29
       26: ILOAD_0
//StringBuilder.append:(I)Ljava/lang/StringBuilder;
       27: INVOKEVIRTUAL #32
//StringBuilder.toString:()Ljava/lang/String;
       30: INVOKEVIRTUAL #36
//DecoderException."<init>":(Ljava/lang/String;)V
       33: INVOKESPECIAL #40
       36: ATHROW
       37: ILOAD_1
       38: IRETURN
}
```

Listing 2: Excerpt of disassembled bytecode from code in Listing 1.

```java
class Utils {
  private static final int RADIX = 16;
  static int digit16(byte b) throws DecoderException
       {
    int i = Character.digit((char)b, 16);
    if(i == -1) {
      throw new DecoderException("Invalid URL
           encoding: not a valid digit (radix 16):
           " + b);
    } else {
      return i;
    }
  }
}
```

Listing 3: Decompilation result of Listing 2 with Fernflower.

```
class Utils
{
   static int digit16(byte b)
      throws DecoderException
   {
      int i = Character.digit((char)b, 16);
      if(i == -1)
         throw new DecoderException((new
              StringBuilder()).append("Invalid URL
              encoding: not a valid digit (radix
              16): ").append(b).toString());
      else
         return i;
   }
   private static final int RADIX = 16;
}
```

Listing 4: Decompilation result of Listing 2 with Dava.

- The `if` statement had originally no `else` clause. Indeed, when an exception is thrown in a method that does not catch it, the execution of the method is interrupted. Therefore, leaving the `return` statement outside of the `if` is equivalent to putting it inside an `else` clause.
- In the original code the String `"Invalid URL encoding: not a valid digit (radix 16): "` was actually computed with `"Invalid URL encoding: not a valid digit (radix " + URLCodec.RADIX + "): "`. In this case, `URLCodec.RADIX` is actually a final static field that always contains the value 16 and cannot be changed. Thus it is safe for the compiler to perform this optimization, but the information is lost in the bytecode.

Besides, this does not include the different formatting choices made by the decompilers such as new lines placement and brackets usage for single instructions such as `if` and `else`.

3. Decompiler evaluation methodology

In this section, we introduce definitions, metrics and research questions. Next, we detail the framework to compare decompilers and we describe the Java projects that form the set of case studies for this work.

3.1. Definitions and Metrics

The value of the results produced by decompilation varies greatly depending on the intended use of the generated source code. In this work, we evaluate the decompilers capacity to produce a faithful retranscription of the original sources. Therefore, we collect the following metrics.

Definition 1. ***Syntactic correctness.*** *The output of a decompiler is syntactically correct if it contains a valid Java program, i.e. a Java program that is recompilable with a Java compiler without any error.*

When a bytecode decompiler generates source code that can be recompiled, this source code can still be syntactically different from the original. We introduce a metric to measure the scale of such a difference according to the abstract syntax tree (AST) dissimilarity[6] between the original and the decompiled results. This metric, called *syntactic distortion*, allows to measure the differences that go beyond variable names. The description of the metric is as follows:

Definition 2. ***Syntactic distortion.*** *The minimum number of atomic edits required to transform the AST of the original source code of a program into the AST of the corresponding decompiled version of it.*

In the general case, determining if two program are semantically equivalent is undecidable. For some cases, the decompiled sources can be recompiled into bytecode that is equivalent to the original, modulo reordering of the constant pool. We call these cases *strictly equivalent* programs. We measure this equivalence with a bytecode comparison tool named Jardiff.[2]

Inspired by the work of [19] and [37], we check if the decompiled and recompiled program is semantically equivalent modulo inputs. This means that for a given set of inputs, the two programs produce equivalent outputs. In our case, we select the set of relevant inputs and assess equivalence based on the existing test suite of the original program.

Definition 3. ***Semantic equivalence modulo inputs.*** *We call a decompiled program semantically equivalent modulo inputs to the original if it passes the set of tests from the original test suite.*

In the case where the decompiled and recompiled program produce non-equivalent outputs, that demonstrates that the sources generated by the decompiler express a different behavior than the original. As explained by Hamilton and colleagues [9], this is particularly problematic as it can mislead decompiler users in their attempt to understand the original behavior of the program. We refer to theses cases as *deceptive decompilation* results.

Definition 4. ***Deceptive decompilation:*** *Decompiler output that is syntactically correct but not semantically equivalent to the original input.*

3.2. Research Questions

We elaborated five research questions to guide our study on the characteristics of modern Java decompilers.

RQ1: To what extent is decompiled Java code syntactically correct? In this research question, we investigate the effectiveness of decompilers for producing syntactically correct and hence recompilable source code from bytecode produced by the *javac* and *ecj* compilers.

RQ2: To what extent is decompiled Java code semantically equivalent modulo inputs? In this research question, we investigate the semantic differences between the original source code and the outputs of the decompilers.

RQ3: To what extent do decompilers produce deceptive decompilation results? Le and colleagues [19] propose

[2] https://github.com/scala/jardiff

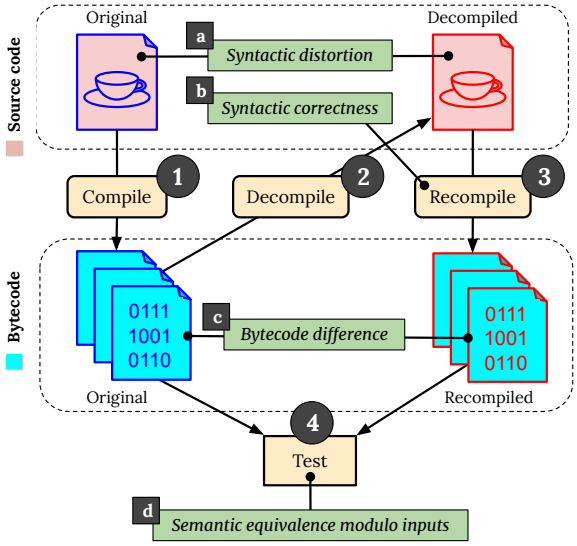

Figure 1: Java decompiler assessment pipeline with four evaluation layers: syntactic distortion, bytecode difference, syntactic correctness, and semantic equivalence modulo input.

to use equivalence modulo inputs assessment as a way to test transformations that are meant to be semantic preserving (in particular compilation). In this research question, we adapt this concept in the context of decompilation testing. In this paper we rely on the existing test suite instead of generating inputs.

RQ4: What is the syntactic distortion of decompiled code? Even if decompiled bytecode is ensured to be syntactically and semantically correct, syntactic differences may remain as an issue when the purpose of decompilation is human understanding. Keeping the decompiled source code free of syntactic distortions is essential during program comprehension, as many decompilers can produce human unreadable code structures. In this research question, we compare the syntactic distortions produced by decompilers.

RQ5: To what extent do the successes and failures of decompilers overlap? In this research question we investigate how the classes for which each decompiler produce both syntactically correct and semantically equivalent modulo input sources intersect.

3.3. Study Protocol

Figure 1 represents the pipeline of operations conducted on every Java source file in our dataset. For each triplet <*decompiler, compiler, project*>, we perform the following:

1. Compile the source files with a given compiler.

2. Decompile each class file with a decompiler (there might be several classes if the source defines internal classes). If the decompiler does not return any error, we mark the source file as decompilable. Then, (a) we measure syntactic distortion by comparing the AST of the original source with the AST of the decompiled source.

3. Recompile the class files with the given compiler. If the compilation is successful, we know that the decompiler produces (b) syntactically correct code. Then, we measure (c) the difference between the original and the recompiled bytecode.

4. Run the test cases on the recompiled bytecode. If the tests are successful, we mark the source as *passTests* for the given triplet, showing that the decompiler produces (d) semantically equivalent code modulo inputs.

If one of these steps fails we do not perform the following steps and consider all the resulting metrics not available. As decompilation can sometimes produce a program that does not stop, we set a 20 minutes timeout on the test execution (the original test suites run under a minute on the hardware used for this experiment, a Core i5-6600K with 16GB of RAM).

The tests used to assess the semantic equivalence modulo inputs are those of the original project that cover the given Java file.[3] We manually excluded the tests that fail on the original project (either flaky or because versioning issue). The list of excluded tests is available as part of our experiments.

3.4. Study Subjects

Decompilers. Table 1 shows the set of decompilers under study. We have selected Java decompilers that are (i) freely available, and (ii) have been active in the last two years. We add Jode in order to compare our results with a legacy decompiler, and because the previous survey by Hamilton and colleagues considers it to be one of the best decompilers [9].

The column VERSION shows the version used (some decompilers do not follow any versioning scheme). We choose the latest release if one exists, if not the last commit available the 09-05-2019. The column #COMMITS represents the number of commits in the decompiler project, in cases where the decompiler is a submodule of a bigger project (e.g. Dava and Fernflower) we count only commits affecting the submodule. The column #LOC is the number of lines of code in all Java files (and Python files for Krakatau) of the decompiler, including sources, test sources and resources counted with *cloc*.[4]

Projects. In order to get a set of real world Java projects to evaluate the eight decompilers, we reuse the set of projects of Pawlak and colleagues[24]. To these 13 projects we added a fourteenth one named `DcTest` made out of examples collected from previous decompiler evaluations [9, 17].[5] Table 2 shows a summary of this dataset: the Java version in which they are written, the number of Java source files, the number of unit tests as reported by Apache Maven, and the number of Java lines of code in their sources.

[3] Coverage was assessed using yajta `https://github.com/castor-software/yajta`

[4] `http://cloc.sourceforge.net/`

[5] `http://www.program-transformation.org/Transform/JavaDecompilerTests`

Table 1
Characteristics of the studied decompilers.

Decompiler	Version	Status	#Commits	#LOC
CFR [1]	0.141	Active	1433	52098
Dava [12]	3.3.0	2018-06-15*	14	22884
Fernflower [13]	NA*	Active	453	52118
JADX [29]	0.9.0	Active	970	55335
JD-Core [3]	1.0.0	Active	NA***	36730
Jode [10]	1.1.2-pre1	2004-02-25*	NA***	30161
Krakatau [30]	NA*	2018-05-13*	512	11301
Procyon [31]	0.5.34	Active	1080	122147

* Date of last update.
** Not following any versioning scheme.
*** CVS not available at the date of the present study.

Table 2
Characteristics of the projects used to evaluate decompilers.

Project name	Java version	#Classes	#Tests	#LOC
Bukkit	1.6	642	906	60800
Commons-codec	1.6	59	644	15087
Commons-collections	1.5	301	15067	62077
Commons-imaging	1.5	329	94	47396
Commons-lang	1.8	154	2581	79509
DiskLruCache	1.5	3	61	1206
JavaPoet*	1.6	2	60	934
Joda time	1.5	165	4133	70027
Jsoup	1.5	54	430	14801
JUnit4	1.5	195	867	17167
Mimecraft	1.6	4	14	523
Scribe Java	1.5	89	99	4294
Spark	1.8	34	54	4089
DcTest**	1.5 – 1.8	10	9	211
Total		2041	25019	378121

(*) Formerly named JavaWriter.
(**) Examples collected from previous decompilers evaluation.

As different java compilers may translate the same sources into different bytecode representations,[6] we employed the two most used Java compilers: *javac* and *ecj* (we use versions 1.8.0_17 and 13.13.100, respectively). We compiled all 14 projects with both compilers (except `commons-lang` which we failed to build with *ecj*). This represents 1887 class files for each compiler that we use to evaluate syntactic correctness of decompiler outputs in RQ1 and syntactic distortion in RQ4. We select only those that contain code executed by tests (2397 grouping files generated by the two compilers) to evaluate semantic correctness in RQ2 and RQ3.

4. Experimental Results

4.1. RQ1: (syntactic correctness) To what extent is decompiled Java code syntactically correct?

This research question investigates to what extent the source code produced by the different decompilers is syntactically correct, meaning that the decompiled code compiles. We also investigate the effect of the compiler that produces the bytecode on the decompilation results.

Figure 2 shows the ratio of decompiled classes that are syntactically correct per pair of compiler and decompiler. The horizontal axis shows the ratio of syntactically correct output in green, the ratio of syntactically incorrect output in blue, and the ratio of empty output in red (an empty output occurs, e.g. when the decompiler crashes). The vertical axis shows the compiler on the left and decompiler on the right. For example, Procyon, shown in the last row, is able to produce syntactically correct source code for 1609 (85.3%) class files compiled with *javac*, and produce a non empty syntactically incorrect output for 278 (14.7%) of them. On the other hand, when sources are compiled with *ecj*, Procyon generates syntactically correct sources for 1532 (82.2%) of class files and syntactically incorrect for 355 (18.8%) sources. In other words, Procyon is slightly more effective when used against code compiled with *javac*. It is interesting to notice that not all decompiler authors have decided to handle error the same way. Both Procyon and Jode's developers have decided to always return source files, even if incomplete (for our dataset). Additionally, when CFR and Procyon detect a method that they cannot decompile properly, they may replace the body of the method by a single `throw` statement and comment explaining the error. This leads to syntactically correct code, but not semantically equivalent.

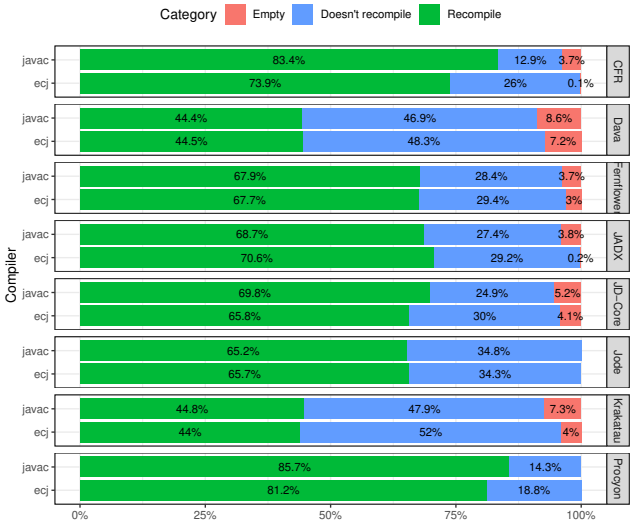

Figure 2: Successful recompilation ratio after decompilation for all considered decompilers.

The ratio of syntactically correct decompiled code ranges from 85.7% for Procyon on *javac* inputs (the best), down to 44% for Krakatau on *ecj* (the worst). Overall, no decompiler is capable of correctly handling the complete dataset. This illustrates the challenges of Java bytecode decompilation, even for bytecode that has not been obfuscated, as in the case of our experiments.

We note that syntactically incorrect decompilation can still be useful for reverse engineering. However, an empty output is useless: the ratio of class files for which the decompilation completely fails is never higher than 8.6% for Dava on *javac* bytecode.

Intuitively, it seems that the compiler has an impact on decompilation effectiveness. To verify this, we use a χ^2 test on the ratio of classfile decompiled into syntactically correct source code depending on the used compiler, *javac* versus

[6] https://www.benf.org/other/cfr/eclipse-differences.html

```
1    IFEQ L2
2    IFNE L2
3    GOTO L0
4   L2
5    ALOAD 5
6    INVOKESTATIC Lang$LangRule.access$100
         (LLang$LangRule;)Z
7    IFEQ L3
8    ALOAD 3
9    ALOAD 5
10   INVOKESTATIC Lang$LangRule.access$200
         (Lang$LangRule;)Ljava/util/Set;
11   INVOKEINTERFACE Set.retainAll (LCollection;)Z
12   @INVOKEVIRTUAL HashSet.retainAll (LCollection;)Z@
13   (itf)
14   POP
15   GOTO L2
16   GOTO L0
```

Listing 5: Excerpt of bytecode from class `org/apache/commons/codec/language/bm/Lang.class`, compiled with *javac* and decompiled with CFR: Lines in red are in the original bytecode, while lines in green are from the recompiled sources.

ecj. The compiler variable has an impact for three decompilers and no impact for the remaining five at 99% confidence level. The test rejects that the compiler has no impact on the decompilation syntactic correctness ratio for CFR, Procyon and JD-Core (p-value 10^{-14}, 0.00027 and 0.006444). For the five other decompilers we do not observe a significant difference between *javac* and *ecj* (p-values: Dava 0.15, Fernflower 0.47, JADX 0.17, Jode 0.50, and Krakatau 0.09). Note that beyond syntactic correctness, the compiler may impact the correctness of the decompiled code, this is discussed in more details in Section 4.3.

To sum up, Procyon and CFR are the decompilers that score the highest on syntactic correctness. The three decompilers ranking the lowest are Jode, Krakatau and Dava. It is interesting to note that those three are no longer actively maintained.

Answer to RQ1: No single decompiler is able to produce syntactically correct sources for more than 85.7% of class files in our dataset. The implication for decompiler users is that decompilation of Java bytecode cannot be blindly applied and does require some additional manual effort. Only few cases make all decompilers fail, which suggests that using several decompilers in conjunction could help to achieve better results.

4.2. RQ2: (semantic equivalence) To what extent is decompiled Java code semantically equivalent modulo inputs?

To answer this research question, we focus on the 2397 class files that are covered by at least one test case. When decompilers produce sources that compile, we investigate the semantic equivalence of the decompiled source and their original. To do so, we split recompilable outputs in three categories: (i) *semantically equivalent:* the code is recompiled into bytecode that is strictly identical to the original (modulo reordering of the constant pool, as explained in Section 3.1), (ii) *semantically equivalent modulo inputs:* the output is recompilable and passes the original project's test suite (i.e. we cannot prove that the decompiled code is semantically different), and (iii) *semantically different:* the output is recompilable but it does not pass the original test suite (deceptive decompilation, as explained in Definition 4).

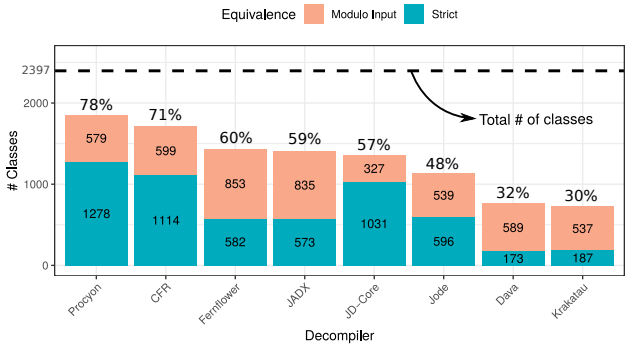

Figure 3: Equivalence results for each decompiler on all the classes of the studied projects covered by at least one test.

Let us first discuss an interesting example of semantic equivalence of decompiled code. Listing 5 shows an example of bytecode that is different when decompiled-recompiled but equivalent modulo inputs to the original. Indeed, we can spot two differences: the control flow blocks are not written in the same order (`L2` becomes `L0`) and the condition evaluated is reversed (`IFEQ` becomes `IFNEQ`), which leads to an equivalent control flow graph. The second difference is that the type of a variable originally typed as a `Set` and instantiated with an `HashSet` has been transformed into a variable typed as an `HashSet`, hence once `remainAll` is invoked on the variable `INVOKEINTERFACE` becomes directly `INVOKEVIRTUAL`. This is still equivalent code.

Now we discuss the results globally. Figure 3 shows the recompilation outcomes of decompilation regarding semantic equivalence for the 2397 classes under study. The horizontal axis shows the eight different decompilers. The vertical axis shows the number of classes decompiled successfully. Strictly equivalent output is shown in blue, equivalent classes modulo input are shown in orange. For example, CFR (second bar) is able to correctly decompile 1713 out of 2397 classes (71%), including 1114 classes that are recompilable into strictly equivalent bytecode, and 599 that are recompilable into equivalent bytecode modulo inputs.

The three decompilers that are not actively maintained anymore (Jode, Dava and Krakatau) handle less than 50% of the cases correctly (recompilable and pass tests). On the other hand, Procyon and CFR have the highest ratio of equivalence modulo inputs of 78% and 71%, respectively.

Answer to RQ2: The number of classes for which the decompiler produces semantically equivalent sources modulo input varies a lot from one decompiler to another. The source code generated by the decompilers is usually not strictly identical to the original, still many of the decompiled classes are semantically equivalent modulo inputs. For end users, it means that the state of the art of Java decompilation does not guarantee semantically correct decompilation, and care must be taken not to blindly trust in the decompiled code.

4.3. RQ3: (bug finding) To what extent do decompilers produce deceptive decompilation results?

As explained by Hamilton and colleagues [9], while a syntactically incorrect decompilation output may still be useful to the user, syntactically correct but semantically different output is more problematic. Indeed, this may mislead the user by making her believe in a different behavior than the original program has. We call this case *deceptive decompilation* (as explained in Definition 4). When such cases occur, since the decompiler produces an output that is semantically different from what is expected, they may be considered decompilation bugs.

Figure 4 shows the distribution of bytecode classes that are deceptively decompiled. Each horizontal bar groups deceptive decompilation per decompiler. The color indicates which compiler was used to produce the class file triggering the error. In blue is the number of classes leading to a decompilation error only when compiled with *javac*, in green only when compiled with *ecj*, and in pink is the number of classes triggering a decompilation error with both compilers. The sum of these classes is indicated by the total on the right side of each bar. Note that the bars in Figure 4 represent the number of bug manifestations, which are not necessarily distinct bugs: the same decompiler bug can be triggered by different class files from our benchmark.

Overall, Jode is the least reliable decompiler, with 83 decompilation bug instances in our benchmark. While Fernflower produces the least deceptive decompilations on our benchmark (13), it is interesting to note that CFR produces only one more deceptive decompilation (14) but that corresponds to less bugs per successful decompilation. This makes CFR the most reliable decompiler on our benchmark.

We manually inspected 10 of these bug manifestations. 2 of them were already reported by other users. We reported the other 8 to the authors of decompilers.[7] The sources of errors include incorrect cast operation, incorrect control-flow restitution, auto unboxing errors, and incorrect reference resolution. Below we detail two of these bugs.

4.3.1. Case study: incorrect reference resolution

We analyze the class `org.bukkit.Bukkit` from the Bukkit project. An excerpt of the original Java source code

[7] https://github.com/castor-software/decompilercmp/tree/master/funfacts

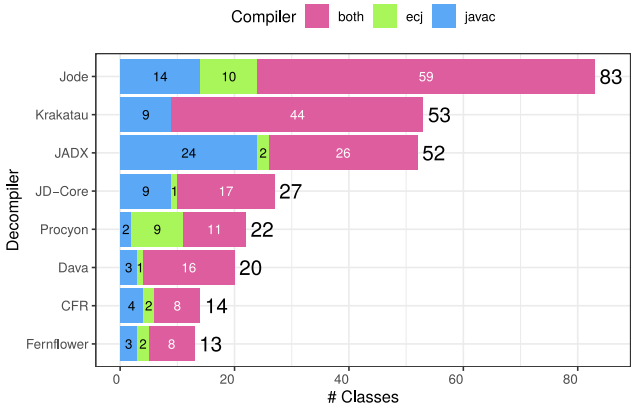

Figure 4: Deceptive decompilation results per decompiler.

```
1  public final class Bukkit {
2     private static Server server;
3        [...]
4     public static void setServer(Server server) {
5        if (Bukkit.server != null) {
6        if (server != null) {
7           throw new UnsupportedOperationException(
8           "Cannot redefine singleton Server");
9        }
10       Bukkit.server = server;
11       server = server;
12       [...]
13    }
```

Listing 6: Exerpt of differences in `org.bukkit.Bukkit` original (in red) and decompiled with JADX sources (in green).

is given in Listing 6. The method `setServer` implements a setter of the static field `Bukkit.server`. This is an implementation of the common Singleton design pattern. In the context of method `setServer`, `server` refers to the parameter of the method, while `Bukkit.server` refers to the static field of the class `Bukkit`.

When this source file is compiled with *javac*, it produces a file `org/bukkit/Bukkit.class` containing the bytecode translation of the original source. Listing 7 shows an excerpt of this bytecode corresponding to the `setServer` method (including lines are filled in red, while excluding lines are filled in green)

When using the JADX decompiler on `org/bukkit/Bukkit.class` it produces decompiled sources, of which an excerpt is shown in Listing 6. In this example, the decompiled code is not semantically equivalent to the original version. Indeed, inside the `setServer` method the references to the static field `Bukkit.server` have been simplified into `server` which is incorrect in this scope as the parameter `server` overrides the local scope. In the bytecode of the recompiled version (Listing 7, including lines are filled in green), we can observe that instructions accessing and writing the static field (`GETSTATIC`, `PUTSTATIC`) have been replaced by instructions accessing and writing the local variable instead (`ALOAD`, `ASTORE`).

```
public static setServer(Lorg/bukkit/Server;)V
    GETSTATIC org/bukkit/Bukkit.server :
    Lorg/bukkit/Server;
    ALOAD 0
    IFNULL L0
    NEW java/lang/UnsupportedOperationException
    DUP
    ATHROW
   L0
    ALOAD 0
    PUTSTATIC org/bukkit/Bukkit.server :
    Lorg/bukkit/Server;
    ASTORE 0
    ALOAD 0
    INVOKEINTERFACE org/bukkit/Server.getLogger
        ()Ljava/util/logging/Logger; (itf)
    NEW java/lang/StringBuilder
```

Listing 7: Exerpt of bytecode from class `org/bukkit/Bukkit.class` compiled with *javac*: Lines in red are in the original bytecode, while lines in green are from the recompiled sources (decompiled with JADX).

```
    protected StringBuffer applyRules(final Calendar
        calendar, final StringBuffer buf) {
        return (StringBuffer) applyRules(calendar,
                                (Appendable) buf);
        return this.applyRules(calendar, buf);
    }

    private <B extends Appendable> B applyRules(final
        Calendar calendar, final B buf) {...}
```

Listing 8: Excerpt of differences in `FastDatePrinter` original (in red) and decompiled with Procyon sources (in green).

When the test suite of *Bukkit* runs on the recompiled bytecode, the 11 test cases covering this code fail, as the first access to `setServer` will throw an exception instead of normally initializing the static field `Bukkit.server`. This is clearly a bug in JADX.

4.3.2. Case study: Down cast error

Listing 8 illustrates the differences between the original sources of `org/apache/commons/lang3/time/FastDatePrinter` and the decompiled sources produced by Procyon. The line in red is part of the original, while the line in green is from the decompiled version. In this example, method `applyRules` is overloaded, i.e. it has two implementations: one for a `StringBuffer` parameter and one for a generic `Appendable` parameter (`Appendable` is an interface that `StringBuffer` implements). The implementation for `StringBuffer` down casts `buf` into `Appendable`, calls the method handling `Appendable` and casts the result back to `StringBuffer`. In a non ambiguous context, it is perfectly valid to call a method which takes `Appendable` arguments on an instance of a class that implements that interface. But in this context, without the down cast to `Appendable`, the Java compiler will resolve the method call `applyRules` to the most concrete method. In this case, this will lead `applyRules` for `StringBuffer` to call itself instead of the other method. When executed, this will lead to an infinite recursion ending in a StackOverflowError. Therefore, in this example, Procyon changes the behavior of the decompiled program and introduces a bug in it.

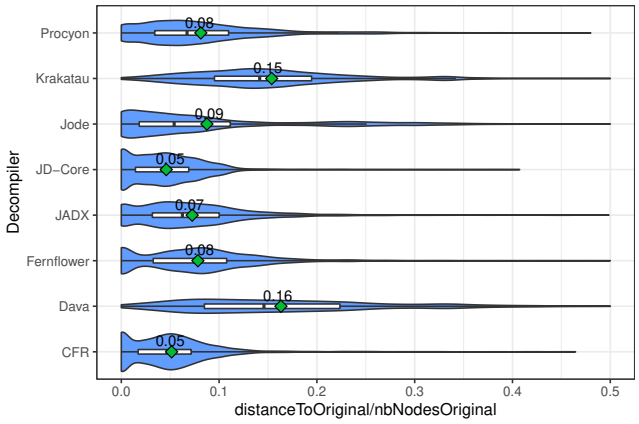

Figure 5: Distribution of ASTs differences between the original and the decompiled source code. Green diamonds indicate average.

Answer to RQ3: Our empirical results indicate that no decompiler is free of deceptive decompilation bugs. The developers of decompilers may benefit from the equivalent modulo input concept to find bugs in the wild and extend their test base. Two bugs found during our study have already been fixed by the decompiler authors, and three other have been acknowledged.

4.4. RQ4: (ASTs difference) What is the syntactic distortion of decompiled code?

The quality of decompilation depends not only on its syntactic compilability and semantic equivalence but also on how well a human can understand the behavior of the decompiled program. The code produced by a decompiler may be syntactically and semantically correct but yet hard to read for a human. In this research question, we evaluate how far the decompiled sources are from the original code. We measure the syntactic distortion between the original and the decompiled sources as captured by AST differences (Definition 2).

Figure 5 shows the distribution of syntactic distortion present in syntactically correct decompiled code, with one violin plot per decompiler. The green diamond marks the average syntactic distortion. For example, the syntactic distortion values of the Jode decompiler have a median of 0.05, average of 0.09, 1st-Q and 3rd-Q of 0.01 and 0.11, respectively. In this figure, lower is better: a lower syntactic distortion means that the decompiled sources are more similar to their original counterparts.

CFR and JD-Core introduce the least syntactic distortion, with high proportion of cases with no syntactic distortion at all (as we exclude renaming). Their median and average syntactic distortion are close to 0.05, which corresponds to 5 edits every 100 nodes in the AST of the source program. On the other extreme, Dava and Krakatau introduce the most

```
1  public class Foo {
2    public int foo(int i, int j) {
3      while (true) {
4        try {
5          while (i < j) i = j++ / i;
6          return j;
7        } catch (RuntimeException re) {
8          i = 10;
9          continue;
10       }
11       break;
12     }
13     return j;
14   }
15 }
```

Listing 9: Excerpt of differences in `Foo` original and decompiled with Fernflower sources.

syntactic distortion with average of 16 (resp. 15) edits per 100 nodes. They also have almost no cases for which they produce sources with no syntactic distortion. It is interesting to note that Dava makes no assumption on the provenance of the bytecode [21]. This partly explains the choice of its author to not reverse some of the optimizations made by Java compilers (See example introduced in Section 2.).

Listing 9 shows the differences on the resulting source code after decompiling the `Foo` class from `DcTest` with Fernflower. As we can observe, both Java programs represent a semantically equivalent program. Yet, their ASTs contain substantial differences. For this example, the edit distance is 3/104 as it contains three tree edits: `MOVE` the return node, and `DELETE` the break node and the continue node (the original source's AST contained 104 nodes).

Note that some decompilers perform some transformations on the sources they produce on purpose to increase readability. Therefore, it is perfectly normal to observe some minimal syntactic distortion, even for decompilers producing readable sources. But as our benchmark is composed of non obfuscated sources, it is expected that a readable output will not fall too far from the original.

Answer to RQ4: All decompilers present various degrees of syntactic distortion between the original source code and the decompiled bytecode. This reveals that all decompilers adopt different strategies to craft source code from bytecode. Our results suggest that syntactic distortion can be used by decompiler developers to improve the alignment between the decompiled sources and the original. Also, decompiler users can use this analysis when deciding which decompiler to employ.

4.5. RQ5: (Decompiler Diversity) To what extent do the successes and failures of decompilers overlap?

In the previous research questions, we observe that different decompilers produce source code that varies in terms of syntactic correctness, semantic equivalence and syntactic distortion. Now, we investigate the overlap in successes and failures of the different decompilers considered for this study.

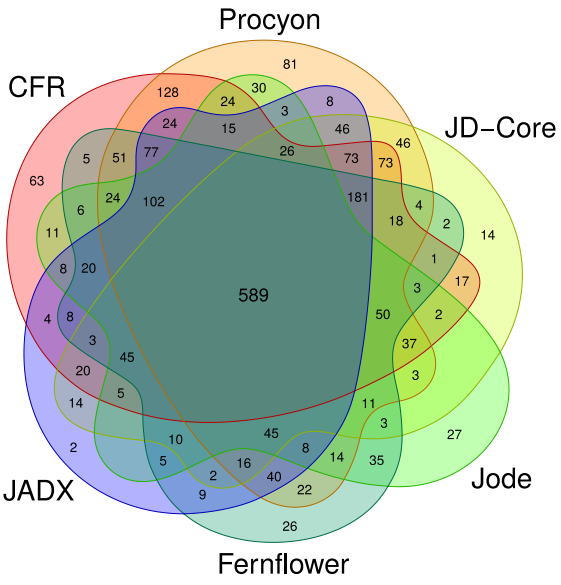

Figure 6: Venn diagram of syntactically and semantically equivalent modulo inputs decompilation results.

Table 3
Summary results of the studied decompilers

Decompiler	#Recompilable	#PassTest	#Deceptive
CFR	3097 (0.79)	1713 (0.71)	22
Dava	1747 (0.44)	762 (0.32)	36
Fernflower	2663 (0.68)	1435 (0.60)	21
JADX	2736 (0.70)	1408 (0.59)	78
JD-Core	2726 (0.69)	1375 (0.57)	82
Jode	2569 (0.65)	1161 (0.48)	142
Krakatau	1746 (0.44)	724 (0.30)	97
Procyon	3281 (0.84)	1869 (0.78)	33
Union	3734 (0.95)	2240 (0.93)	342

Figure 6 shows a Venn Diagram of syntactically and semantically equivalent classes modulo input for decompiled/recompiled classes. We exclude Dava and Krakatau because they do not handle correctly any class file unique to them. We see that 6/8 decompilers have cases for which they are the only decompiler able to handle it properly. These cases represent 276/2397 classes. Only 589/2397 classes are handled correctly by all of these 6 decompilers. Furthermore, 157/2397 classes are not correctly handled by any of the considered decompilers.

Let us now Table 3 summarize the quantitative results obtained from the previous research questions. Each line corresponds to a decompiler. Column `#Recompilable` shows the number of cases (and ratio) for which the decompiler produced a recompilable output; column `#PassTest` shows the number of cases where the decompiled code passes those tests; column `#Deceptive` indicate the number of cases that were recompilable but did not pass the test suite (i.e. a decompilation bug). The line 'Union' shows the number of classes for which at least 1 decompiler succeeds to produce `Recompilable` sources and respectively sources that `pass tests`. The column `#Deceptive` indicates the number of classes for which at least 1 decompiler produced a deceptive decompilation. This means that for 2240 out of the 2397 (93%) classes of our dataset, there is

at least 1 decompiler that produces semantically equivalent sources modulo inputs. This number must be taken with a grain of salt, as it does not mean that someone looking for a successful decompilation of one of these classes could find one trivially. Overall, 342 out of 2397 classes have at least 1 decompiler that produce a deceptive decompilation. Assuming that one can merge the successful decompilation results together, we would obtain a better decompiler overall, this is what we explore in Section 5.

Answer to RQ5: Overall, the classes for which each decompiler produce semantic equivalence modulo inputs sources do not overlap. For 6 out of 8 decompilers, there exist at least 1 classes for which the decompiler is the only one to produce semantic equivalence modulo inputs sources. In theory, a union of the best features of each decompiler would cover 2240 out of the 2397 (93%) classes of the dataset. This suggests to use multiple decompilers in conjunction to improve decompilation results.

5. Meta Decompilation

In this section, we present an original concept for decompilation.

5.1. Overview

In 1995, Selberg et al. [28] noticed that different web search engines produced different results for the same input query. Furthermore, there was no better tool over all queries: different types of query were better handled by different search engines. They exploited this finding in a tool called MetaCrawler[8], which delegates a user query to various search engines and merges the results.

In this paper, we apply a similar approach to improve Java decompilation. Each decompiler has its strengths and weaknesses, and the subset of JVM bytecode sequences they correctly handle is diverse (cf Section 4.5). Therefore, our idea is to combine decompilers in a meta-decompiler as MetaCrawler combines search engines. The idea is to have a tool that merges partially incorrect decompilation results in order to produce a correct one. In this paper, we propose a tool called Arlecchino, that implements such a 'meta-decompilation' approach.

5.2. Example

The class `org.bukkit.configuration.file.YamlConfiguration` of the project Bukkit is an example of a class file that is incorrectly handled by both JADX and Dava. While both decompilers produce syntactically incorrect Java code for this class, the error that prevents successful recompilation is not located at the same place in both decompiled classes.

Listing 10 shows an excerpt of the decompiled sources produced by Dava for `YamlConfiguration`. The static field `BLANK_CONFIG` is initialized with an incorrect string

```
1  public class YamlConfiguration extends
       FileConfiguration {
2     protected static final String COMMENT_PREFIX = "#
          ";
3     protected static final String BLANK_CONFIG = "{}
4  ";
5     private final DumperOptions yamlOptions;
6     private final Representer yamlRepresenter;
7     private final Yaml yaml;
8
9     public YamlConfiguration()
10    {
11       DumperOptions r7;
12       YamlRepresenter r8;
13       YamlConstructor r9;
14       Yaml r10;
15       BaseConstructor r11;
16       r7 = new DumperOptions();
17       yamlOptions = r7;
18       r8 = new YamlRepresenter();
19       yamlRepresenter = r8;
20       r9 = new YamlConstructor();
21       r11 = (BaseConstructor) r9;
22       r10 = new Yaml(r11, yamlRepresenter,
             yamlOptions);
23       yaml = r10;
24    }
25  [...]
26 }
```

Listing 10: Excerpt of `org.bukkit.configuration.file.YamlConfiguration` decompiled with Dava.

```
1  public class YamlConfiguration extends
       FileConfiguration {
2     protected static final String BLANK_CONFIG =
          "{}\n";
3     protected static final String COMMENT_PREFIX = "#
          ";
4     private final Yaml yaml =
5        new Yaml(new YamlConstructor(),
6        this.yamlRepresenter, this.yamlOptions);
7     private final DumperOptions yamlOptions = new
          DumperOptions();
8     private final Representer yamlRepresenter = new
          YamlRepresenter();
9   [...]
10 }
```

Listing 11: Excerpt of `org.bukkit.configuration.file.YamlConfiguration` decompiled with JADX.

literal that contains a non escaped line return. When attempting to recompile these sources, *javac* produces an `unclosed string literal` error for both line 3 and 4.

Listing 11 shows an excerpt of the decompiled sources produced by JADX for the same class. The static field `BLANK_CONFIG` is correctly initialized with `"{}\n"`, but the initialization of `yaml`, `yamlOptions` and `yamlRepresenter` are conducted out of order, which lead to a compilation error as `yamlOptions` and `yamlRepresenter` are still `null` when `yaml` is initialized. Intuitively, one can see that Dava's solution could be fixed by replacing lines 3 and 4 with line 2 from JADX's solution. This is an example of successful meta-decompilation, merging the output of two decompilers.

[8] https://www.metacrawler.com/

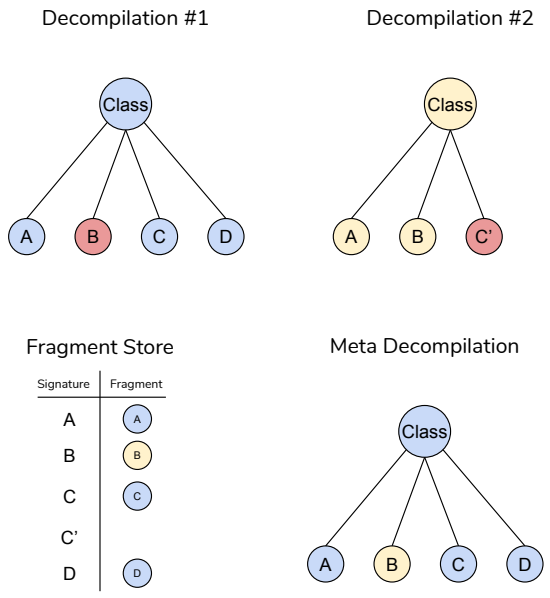

Figure 7: Meta decompilation: Merger of different partial decompilation. Node in blue originates from Decompiler #1, nodes in yellow originate from Decompiler #2. Nodes in red contains compilation errors.

Figure 7 illustrates how two erroneous decompilations can be merged into one that is correct. `Decompilation #1` represents the AST of the sources produced by one decompiler that contains 4 type members (A,B,C and D), and one compilation error located on B. `Decompilation #2` represents the sources produced by a different decompiler for the same class. It only contains 3 type member (A,B an C') and one compilation error located on C'. The fragment store is a dictionary containing an error free AST fragment for each type member when such a fragment exists. `Meta Decompilation` shows an example of error free AST that can be built based on `Decompilation #1` and the store that combines AST fragments from both decompilations. Note that different decompilers may produce sources that do not exactly contain the same type members. This is illustrated here by `Decompilation #2` not having a type member D and having a different signature for C.

5.3. Algorithm

Algorithm 1 describes the process of meta decompilation as implemented by Arlecchino. Arlecchino takes as input a bytecode file, and a list of bytecode decompilers. The process starts with an empty set of solutions and an empty fragment store of correct fragments. This fragment store is a dictionary that associates a type member signature to a fragment of AST free of compilation error corresponding to the type member in question.

For each decompiler, the meta-decompilation goes through the following steps.

The bytecode file is passed to the decompiler d. An AST is built from the decompiled sources (line 4). While building the AST, the compilation errors and their location are gathered (if any) and the type members containing errors are annotated as such. A class abstract syntax tree includes a node for the class itself as the root, as well as children representing class information (super class, super interfaces, formal type parameters) and type members. Type members include fields, methods, constructors, inner classes, enum values, and static blocks. These type members' source locations are recorded and compared with the compiler error locations. If an error is located between a type member start and end location, the type member is annotated as errored. For example, the element corresponding to the field `BLANK_CONFIG` is annotated as errored in Dava's solution for `YamlConfiguration`. This annotated AST, that we call solution, is added to the set of remaining solutions.

Additionally, for all type members in the current solution, if the fragment store does not already contain an error free fragment with the same signature, the type member is added to the fragment store (line 8). The signature of a type member is a character string that identifies it uniquely. For example, the signature of the field `BLANK_CONFIG` is `org.bukkit.configuration.file.YamlConfiguration#BLANK_CONFIG` and the signature of `YamlConfiguration`'s constructor is `org.bukkit.configuration.file.YamlConfiguration()`.

Data: $bytecode$ A bytecode file,
$Decompilers$ A set of decompilers
Result: The decompiled java sources corresponding

```
1  Solutions ← {}
2  FragmentStore ← {}
3  foreach dc ∈ Decompilers do
4      solution ← AST(decompile(dc, bytecode))
5      Fragments ← fragmentsOf(solution)
6      foreach f ∈ Fragments do
7          if ¬problem(f) ∧ signature(f) ∉ Store
             then
8              FragmentStore ←
                 FragmentStore ∪ {signature(f) → f}
9          end
10     end
11     Solutions ← Solutions ∪ {solution}
12     foreach s ∈ Solutions do
13         if completable(s, FragmentStore) then
14             if
                 recompile(complete(s, FragmentStore))
                 then
15                 return print(s);
16             else
17                 remove(s, FragmentStore)
18             end
19         end
20     end
21 end
```

Algorithm 1: Meta decompilation procedure.

Table 4
Arlecchino results on classes with no correct decompilation from state of the art decompilers.

#PassTest	59	37.6%
#Deceptive	11	7.0%
#!Recompile	87	55.4%
Total	157	100%

Each solution in the set of solutions is checked for completion with the current store (line 12). A solution is "completable" with the members in a given fragment store, if all the solution's type members annotated with an error are present in the fragment store. Indeed, these type members' AST can be replaced with an error free variant present in the fragment store. If a solution is completable with the current fragment store, all its type members annotated as errored are replaced with a fragment from the fragment store. The solution is then passed to the compiler to check if it compiles. If it does, it is printed, and the meta decompilation stops. If not, the solution is removed from the set of solutions.

By attempting to repair each solution and its given set of type members with a minimum of transplanted fragments from those available in the fragment store, Arlecchino does not favor any type member set. This allows Arlecchino to deal with cases where the different solutions do not contain the same type members. This occurs with implicit constructor declarations such as the one present in Listing 11 with `YamlConfiguration`. It also makes it possible to handle cases where element signatures might differ depending on how type erasure is dealt with by each decompiler. And finally, it handles cases where elements might not be in the same order (and the order of type members is meaningful as seen in Listing 11).

5.4. Experimental results about meta-decompilation

The following section evaluates the effectiveness of Arlecchino. It is organized as follows. First, we gather the 157 classes of our dataset for which no decompilers produced semantically equivalent modulo input sources and assess the results produced by Arlecchino. Second, we run Arlecchino on the complete dataset of classes in this study. We then evaluate the results in regards of semantic equivalence modulo inputs. Finally, we study the origin of fragments produced by Arlecchino and discuss the consequences on the number of deceptive decompilations.

Table 4 shows the results of meta decompilation on the 157 classes of our dataset that led to decompilation errors for all decompilers in the study. Arlecchino produces semantically equivalent results for 59 out of 157 (37.6%) classes. It produces deceptive decompilation for 11 (7.0%) classes and fails to produce recompilable results for 87 out of 157 (55.4%) classes. The success case where Arlecchino produces correct output is when: 1) at least one compiler is able to read the correct signature for all type members of a class and, 2) an error free decompilation exists for all of these type members. However, when no decompiler is able to decompile a specific type member or that no decompiler reads correctly the signature of all type members, no meta decompilation can be successful. These results demonstrate that successful decompilation (in terms of both syntactic correctness and semantic equivalence modulo inputs) can be found by Arlecchino for classes where no other decompilers can.

Table 5
Comparison of Arlecchino results with state of the art.

Decompiler	#Recompilable	#PassTest	#Deceptive	ASTDiff
CFR	3097 (79%)	1713 (71%)	22 (1.27%)	0.05
Procyon	3281 (84%)	1869 (78%)	33 (1.74%)	0.08
Arlecchino	3479 (89%)	2087 (87%)	30 (1.42%)	0.06
Total	3928 (100%)	2397 (100%)	-	-

Table 5 shows the results obtained when running Arlecchino on the whole dataset presented in Section 3 and compares it with Procyon and CFR. Procyon is the decompiler that scores the highest in terms of syntactic correctness as well as semantic equivalence modulo inputs, while CFR scores the lowest in deceptive decompilation rate and syntactic distortion. The first column indicates the number of classes for which each decompiler produced syntactically correct sources, among the 2928 from the dataset. The second column shows the number of classes for which each decompiler produced semantically correct modulo inputs sources among the 2397 classes covered by tests. The third column indicates the number of deceptive decompilations produced by each decompiler. The last column shows the median syntactic distortion in number of edits per nodes in the original AST. Arlecchino produced syntactically correct sources for 3479 classes (89%). It produces semantically equivalent modulo inputs sources for 2087 (87%) classes, and 30 deceptive decompilations. Compared with Procyon, Arlecchino produces syntactically correct sources for 198 more classes, semantically correct modulo inputs sources for 208 more. It also produces 3 less deceptive decompilations, and has a lower syntactic distortion. Compared with CFR, Arlecchino produces 8 more deceptive decompilations but it produces semantically correct modulo inputs sources for 374 more classes. In percentage of deceptive decompilation among recompilable decompilation, Arlecchino produces 1.42% of deceptive decompilation which is lower than Procyon's 1.74% but slightly higher than CFR's 1.27%. Overall, Arlecchino scores higher than all studied decompilers in terms of semantic correctness as well as semantic correctness modulo inputs, and ranks second in deceptive decompilation rate by a small margin. This is, to our knowledge, the first implementation of this meta-decompilation approach. It both demonstrates the validity of the approach and adds a new state of the art tool that practitioners can use to decompile java bytecode.

Note that Arlecchino also has its implementation flaws and may fail where other decompilers may succeed. But it may be used in conjunction of other decompilers. The union

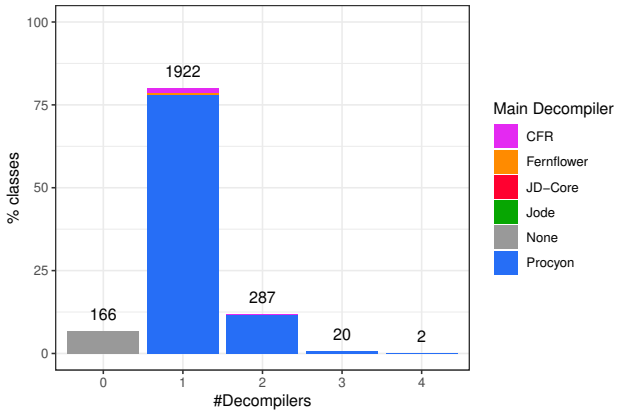

Figure 8: Distribution of the number of decompilers used by Arlecchino.

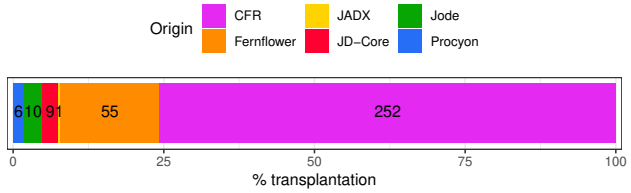

Figure 9: Distribution of the origin of transplanted fragments in Arlecchino results.

of classes for which at least one decompiler (including Arlecchino) produces both syntactically correct and semantically equivalent sources, presented in RQ5, now covers 2299 out of 2397 classes (96%) of our dataset.

5.4.1. Remaining deceptive decompilations

In order to investigate deceptive decompilations produced by Arlecchino, we need to investigate the origins of the AST fragments used in each decompilation.

Figure 8 shows the distribution of the number of decompilers used by Arlecchino for each of the 2397 classes of our dataset for which we have tests. Arlecchino finds no solution for 166 classes. For 1922 classes, only one decompiler was used, meaning that there is no need for meta-decompilation. For 287 classes, Arlecchino combines the output of 2 decompilers. It uses 3 and 4 decompilers for 20 classes and 2 classes respectively. The color indicates which decompiler's base solution was used. In the overwhelming majority, the Procyon solution is used.

Figure 9 shows the distribution of transplanted fragments' origin for the 309 classes where several decompilers are used. For 252 classes, one or more fragments from CFR's solution were transplanted to build Arlecchino's solution. 55 classes have fragments coming from Fernflower, and the rest of the distribution is negligible. Note that as Arlecchino stops as soon as it finds an admissible solution. Thus, the order of decompilers when building a solution largely impacts this distribution.

Arlecchino produces a deceptive decompilation either when the first recompilable solution of a given type member is a deceptive one, or the assembly of different fragments introduces an error.

```java
import org.junit.runners.model.*;
import org.junit.internal.runners.*;
public abstract class Request {
      [...]
   public static Request classes(final Computer
        computer, final Class<?>... classes) {
      try {
         final AllDefaultPossibilitiesBuilder
              builder = new
              AllDefaultPossibilitiesBuilder(true);
         final Runner suite =
              computer.getSuite(builder, classes);
         return runner(suite);
      }
      catch ( InitializationError e) {
         throw new RuntimeException("Bug in saff's
              brain: Suite constructor, called as
              above, should always complete");
      }
   }

   public static Request runner(final Runner runner)
        {
      return new Request() {
         @Override
         public Runner getRunner() {
            return runner;
         }
      };
   }
}
```

Listing 12: Excerpt of `org.junit.runner.Request` decompiled with Procyon.

In order to minimize these problems, Arlecchino uses Procyon as the first decompiler and orders the other decompilers by their deceptive decompilation rate.

Therefore, most of the decompilers' deceptive decompilations are for the same classes as Procyon's one. In a lesser way, deceptive decompilation originating from type members decompiled with CFR affect Arlecchino when those type member are decompiled with syntactic errors by Procyon. Note that, as no software is free of bugs, the implementation of Arlecchino could also add new sources of error. In practice, as shown by Table 5, the number of deceptive decompilations (30) corresponds to a better deceptive decompilation rate than all decompilers of this study except CFR.

5.4.2. Case studies

Here we discuss two examples in details: one successful and one failed meta decompilation.

Success: Request Listing 12 shows the decompiled sources for `org.junit.runner.Request` produced by Procyon. In this example, there are ambiguous references because two types share the same simply qualified name: both `org.junit.runners.model` and `org.junit.internal.runners` contain a type named `InitializationError`, therefore the decompiled sources generated by Procyon lead to a compilation error.

```
import org.junit.runners.model.InitializationError;

public abstract class Request {
      [...]

   public static Request classes(Computer computer,
        Class<?> ... classes) {
      try {
         AllDefaultPossibilitiesBuilder builder =
              new
              AllDefaultPossibilitiesBuilder(true);
         Runner suite = computer.getSuite(builder,
              classes);
         return Request.runner(suite);
      }
      catch (InitializationError e) {
         throw new RuntimeException("Bug in saff's
              brain: Suite constructor, called as
              above, should always complete");
      }
   }

   public static Request runner(Runner runner) {
      return new Request(){

         public Runner getRunner() {
            return Runner.this;
         }
      };
   }
}
```

Listing 13: Excerpt of `org.junit.runner.Request` decompiled with CFR.

Listing 13 shows the decompiled sources for `org.junit.runner.Request` produced by CFR. These sources contains an error in the body of the static method `runner(Runner)`. Since this method contains an anonymous class, when the original sources are compiled, a synthetic field `runner` is created, by the compiler, for the anonymous class. This field contains the parameter `runner` from the enclosing method. When CFR decompiles the bytecode, it incorrectly replaces the statement that returns the parameter of the enclosing method by a statement that returns a field that does not exist in the sources. This leads to a compilation error when attempting to recompile. Since our report, CFR's author has fixed this bug.[9]

While both Procyon and CFR's solutions contain an error, these errors are not located on the same type member. Hence, CFR's fragment for the method `classes(Computer, Class<?>[])` is transplanted on Procyon's solution. This combined solution is recompilable and semantically equivalent modulo input.

Failure: SwitchClosure There are Java constructs for which all decompilers struggle. In these cases, all decompilers may produce an error on the same type member, and this leads to a failed meta-decompilation. The following example illustrates the problem of generic type lower bounds, which all decompilers considered in this study struggle with.

Listing 14 shows an excerpt of the original sources

[9] https://github.com/leibnitz27/cfr/issues/50

```
   private final Closure<? super E> iDefault;

   private SwitchClosure(final boolean clone,
      final Predicate<? super E>[] predicates,
      final Closure<? super E>[] closures,
      final Closure<? super E> defaultClosure) {
      super();
      iPredicates = clone ?
           FunctorUtils.copy(predicates) :
           predicates;
      iClosures = clone ?
           FunctorUtils.copy(closures) : closures;
      iDefault = (Closure<? super E>) (defaultClosure
           == null ? NOPClosure.<E>nopClosure() :
           defaultClosure);
      this.iDefault = (defaultClosure == null ?
           NOPClosure.nopClosure() : defaultClosure);
   }
```

Listing 14: Excerpt of `org.apache.commons.collections4.functors.SwitchClosure`, original and decompiled.

for `org.apache.commons.collections4.functors.SwitchClosure`. The line highlighted in green is the original line. The line highlighted in red is the corresponding line as decompiled by Procyon, CFR and JD-Core. None of them is able to correctly reproduce the cast to `Closure<? super E>`. This leads to a compilation error as the method `NOPClosure.<E>nopClosure()` return type is *Closure<E>*, which is not a subtype of `Closure<? super E>`.

As the decompiled sources for `SwitchClosure` produced by all decompilers contain at least one error on this constructor, no solution is completable with the fragment store at the end of Algorithm 1. Therefore, the meta decompilation fails to produce recompilable sources.

Highlights about meta-decompilation: To summarize, we have devised and implemented a novel approach to merge results from different decompilers, called meta-decompilation. This tool is able to handle 59 of the 157 cases (37.6%) previously not handled by any decompiler. Meta-decompilation is, to our knowledge, a radically new idea that has never been explored before. Our experiments demonstrate the feasibility and effectiveness of the idea.

6. Threats to Validity

In this section, we report about internal, external and reliability threats against the validity of our results.

Internal validity The internal threats are related to the metrics employed, especially those used to compare the syntactic distortion and semantic equivalence modulo inputs between the original and decompiled source code. Moreover, the coverage and quality of the test suite of the projects under study influences our observations about the semantic equivalence of the decompiled bytecode. To mitigate this threat, we select a set of mature open-source projects with good test

suites as study subjects, and rely on state-of-the-art AST and bytecode differencing tools.

External validity The external threats refer to what extent the results obtained with the studied decompilers can be generalized to other Java projects. To mitigate this threat, we reuse an existing dataset of Java programs which we believe is representative of the Java world. Moreover, we added a handmade project which is a collection of classes used in previous decompilers evaluations as a baseline for further comparisons.

Reliability validity Our results are reproducible, the experimental pipeline presented in this study is publicly available online. We provide all necessary code to replicate our analysis, including AST metric calculations and statistical analysis via R notebooks.[10]

7. Related work

This paper is related to previous works on bytecode analysis, decompilation and program transformations. In this section, we present the related work on Java bytecode decompilers along these lines.

The evaluation of decompilers is closely related to the assessment of compilers. In particular, Le et al. [19] introduce the concept of semantic equivalence modulo inputs to validate compilers by analyzing the interplay between dynamic execution on a subset of inputs and statically compiling a program to work on all kind of inputs. Naeem et al. [22] propose a set of software quality metrics aimed at measuring the effectiveness of decompilers and obfuscators. In 2009, Hamilton et al. [9] show that decompilation is possible for Java, though not perfect. In 2017, Kostelansky et al. [17] perform a similar study on updated decompilers. In 2018, Gusarovs [8] performed a study on five Java decompilers by analyzing their performance according to different handcrafted test cases. All those works demonstrate that Java bytecode decompilation is far from perfect.

The objectives of decompilers are similar to disassemblers. However, instead of translating machine language into assembly language for different architectures [34, 16], decompilers work at the high level of source code [15, 32, 33]. A focus on reassembling disassembled binary and its applications already exists in the literature[35, 5, 7]. Miecznikowski and Hendren [21] report about the problems and solutions found during the development of the Dava decompiler. They highlight particular issues related to expression evaluation on the Java stack, exceptions and synchronized blocks and type assignments. Disassemblers can sometimes be much more effective than decompilers, especially when the decompilation process goes wrong[16].

Recently, Katz et al. [14] present a technique for decompiling binary code snippets using a model based on Recurrent Neural Networks, which produces source code that is more similar to human-written code and therefore more easy for humans to understand. This a remarkable attempt at driving decompilation towards a specific goal. Lacomis et al. [18] propose a probabilistic technique for variable name recovery. Schulte et al. [27] use evolutionary search to improve and recombine a large population of candidate decompilations by applying source-to-source transformations gathered from a database of human-written sources. As an example of a multi-tool that exploits diversity, Chen et al. [2] rely on various fuzzers to build an ensemble based fuzzer that gets better performance and generalization ability than that of any constituent fuzzer alone.

[10] https://github.com/castor-software/decompilercmp/tree/master/notebooks

8. Conclusion

In this work, we presented a fully automated pipeline to evaluate the Java bytecode decompilers' capacity to produce compilable, semantically equivalent, and readable code. We proposed the concept of semantic equivalence modulo inputs to compare decompiled sources to their original counterpart. We applied this approach on eight available decompilers through a set of 2041 classes from 14 open-source projects compiled with two different decompilers. The results of our analysis show that bytecode decompilation is a nontrivial task that still requires human work. Indeed, even the highest ranking decompiler in this study produces syntactically correct output for 84% of classes of our dataset and semantically equivalent modulo inputs output for 78%. Meanwhile, the diversity of implementation of these decompilers allows to merge their different results to bypass the shortcomings of single decompilers. We called this approach 'meta decompilation' and implements it in a tool called Arlecchino. Our experimental results show that Arlecchino can produce semantic equivalence modulo inputs sources for 37.6% of classes for which, previously, no single decompiler could.

Acknowledgments

This work has been partially supported by the Wallenberg Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation and by the TrustFull project financed by the Swedish Foundation for Strategic Research.

References

[1] Benfield, L., 2019. CFR. https://www.benf.org/other/cfr/. [Online; accessed 19-July-2019].

[2] Chen, Y., Jiang, Y., Ma, F., Liang, J., Wang, M., Zhou, C., Su, Z., Jiao, X., 2018. EnFuzz: Ensemble Fuzzing with Seed Synchronization among Diverse Fuzzers. arXiv e-prints , arXiv:1807.00182arXiv:1807.00182.

[3] Dupuy, E., 2019. Java Decompiler. https://http://java-decompiler.github.io/. [Online; accessed 19-July-2019].

[4] Ďurfina, L., Křoustek, J., Zemek, P., 2013. PsybOt Malware: A Step-By-Step Decompilation Case Study, in: 20th Working Conference on Reverse Engineering (WCRE), pp. 449–456. doi:10.1109/WCRE.2013.6671321.

[5] Emamdoost, N., Sharma, V., Byun, T., McCamant, S., 2019. Binary mutation analysis of tests using reassembleable disassembly. doi:10.14722/bar.2019.23058.

[6] Falleri, J.R., Morandat, F., Blanc, X., Martinez, M., Monperrus, M., 2014. Fine-grained and Accurate Source Code Differencing, in: 29th International Conference on Automated Software Engineering (ASE), ACM, New York, NY, USA. pp. 313–324. URL: http://doi.acm.org/10.1145/2642937.2642982, doi:10.1145/2642937.2642982.

[7] Flores-Montoya, A., Schulte, E.M., 2019. Datalog disassembly. CoRR abs/1906.03969. URL: http://arxiv.org/abs/1906.03969, arXiv:1906.03969.

[8] Gusarovs, K., 2018. An Analysis on Java Programming Language Decompiler Capabilities. Applied Computer Systems 23, 109–117.

[9] Hamilton, J., Danicic, S., 2009. An Evaluation of Current Java Bytecode Decompilers, in: 9th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM), pp. 129–136. doi:10.1109/SCAM.2009.24.

[10] Hoenicke, J., 2019. JODE. http://jode.sourceforge.net/. [Online; accessed 19-July-2019].

[11] Jaffe, A., Lacomis, J., Schwartz, E.J., Goues, C.L., Vasilescu, B., 2018. Meaningful Variable Names for Decompiled Code: A Machine Translation Approach, in: 26th Conference on Program Comprehension (ICPC), ACM, New York, NY, USA. pp. 20–30. URL: http://doi.acm.org/10.1145/3196321.3196330, doi:10.1145/3196321.3196330.

[12] Jerome Miecznikowski, Nomair A. Naeem, L.J.H., 2019. Dava. http://www.sable.mcgill.ca/dava/. [Online; accessed 19-July-2019].

[13] JetBrains, 2019. Fernflower. https://github.com/JetBrains/intellij-community/tree/master/plugins/java-decompiler/engine. [Online; accessed 19-July-2019].

[14] Katz, D.S., Ruchti, J., Schulte, E., 2018. Using Recurrent Neural Networks for Decompilation, in: 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), pp. 346–356. doi:10.1109/SANER.2018.8330222.

[15] Katz, O., Olshaker, Y., Goldberg, Y., Yahav, E., 2019. Towards Neural Decompilation. arXiv e-prints , arXiv:1905.08325arXiv:1905.08325.

[16] Khadra, M.A.B., Stoffel, D., Kunz, W., 2016. Speculative disassembly of binary code, in: International Conference on Compliers, Architectures, and Sythesis of Embedded Systems (CASES), pp. 1–10. doi:10.1145/2968455.2968505.

[17] Kostelanský, J., Dedera, L., 2017. An Evaluation of Output from Current Java Bytecode Decompilers: Is it Android Which is Responsible for Such Quality Boost?, in: Communication and Information Technologies (KIT), pp. 1–6. doi:10.23919/KIT.2017.8109451.

[18] Lacomis, J., Yin, P., Schwartz, E.J., Allamanis, M., Goues, C.L., Neubig, G., Vasilescu, B., 2019. Dire: A neural approach to decompiled identifier naming. arXiv:1909.09029.

[19] Le, V., Afshari, M., Su, Z., 2014. Compiler Validation via Equivalence Modulo Inputs, in: 35th Conference on Programming Language Design and Implementation (PLDI), ACM, New York, NY, USA. pp. 216–226. URL: http://doi.acm.org/10.1145/2594291.2594334, doi:10.1145/2594291.2594334.

[20] Lindholm, T., Yellin, F., Bracha, G., Buckley, A., 2014. The Java Virtual Machine Specification. Pearson Education.

[21] Miecznikowski, J., Hendren, L., 2002. Decompiling Java Bytecode: Problems, Traps and Pitfalls, in: Horspool, R.N. (Ed.), Compiler Construction, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 111–127.

[22] Naeem, N.A., Batchelder, M., Hendren, L., 2007. Metrics for Measuring the Effectiveness of Decompilers and Obfuscators, in: 15th IEEE International Conference on Program Comprehension (ICPC), pp. 253–258. doi:10.1109/ICPC.2007.27.

[23] Nolan, G., 2004. Decompiler Design. Apress, Berkeley, CA. pp. 121–157. URL: https://doi.org/10.1007/978-1-4302-0739-9_5, doi:10.1007/978-1-4302-0739-9_5.

[24] Pawlak, R., Monperrus, M., Petitprez, N., Noguera, C., Seinturier, L., 2015. Spoon: A Library for Implementing Analyses and Transformations of Java Source Code. Software: Practice and Experience 46, 1155–1179. URL: https://hal.archives-ouvertes.fr/hal-01078532/document, doi:10.1002/spe.2346.

[25] Ragkhitwetsagul, C., Krinke, J., 2017. Using Compilation/Decompilation to Enhance Clone Detection, in: 11th International Workshop on Software Clones (IWSC), pp. 1–7. doi:10.1109/IWSC.2017.7880502.

[26] Robles, G., Gonzalez-Barahona, J.M., Herraiz, I., 2005. An Empirical Approach to Software Archaeology, in: 21st International Conference on Software Maintenance (ICSM), pp. 47–50.

[27] Schulte, E., Ruchti, J., Noonan, M., Ciarletta, D., Loginov, A., 2018. Evolving exact decompilation, in: Shoshitaishvili, Y., Wang, R.F. (Eds.), Workshop on Binary Analysis Research, San Diego, CA, USA. URL: http://www.cs.unm.edu/~eschulte/data/bed.pdf.

[28] Selberg, E., Etzioni, O., 1997. The metacrawler architecture for resource aggregation on the web. IEEE Expert 12, 11–14. doi:10.1109/64.577468.

[29] skylot, 2019. JADX. https://github.com/skylot/jadx. [Online; accessed 19-July-2019].

[30] Storyyeller, 2019. Krakatau. https://github.com/Storyyeller/Krakatau. [Online; accessed 19-July-2019].

[31] Strobel, M., 2019. Procyon. https://bitbucket.org/mstrobel/procyon. [Online; accessed 19-July-2019].

[32] Troshina, K., Derevenets, Y., Chernov, A., 2010. Reconstruction of Composite Types for Decompilation, in: 10th IEEE Working Conference on Source Code Analysis and Manipulation (SCAM), pp. 179–188. doi:10.1109/SCAM.2010.24.

[33] Van Emmerik, M.J., 2007. Static Single Assignment for Decompilation. University of Queensland.

[34] Vinciguerra, L., Wills, L., Kejriwal, N., Martino, P., Vinciguerra, R., 2003. An Experimentation Framework for Evaluating Disassembly and Decompilation Tools for C++ and Java, in: 10th Working Conference on Reverse Engineering (WCRE), IEEE Computer Society, Washington, DC, USA. pp. 14–. URL: http://dl.acm.org/citation.cfm?id=950792.951361.

[35] Wang, S., Wang, P., Wu, D., 2015. Reassembleable disassembling, in: 24th USENIX Security Symposium (USENIX Security 15), USENIX Association, Washington, D.C.. pp. 627–642. URL: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/wang-shuai.

[36] Yakdan, K., Dechand, S., Gerhards-Padilla, E., Smith, M., 2016. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study, in: IEEE Symposium on Security and Privacy (SP), pp. 158–177. doi:10.1109/SP.2016.18.

[37] Yang, Y., Zhou, Y., Sun, H., Su, Z., Zuo, Z., Xu, L., Xu, B., 2019. Hunting for Bugs in Code Coverage Tools via Randomized Differential Testing, in: 41st International Conference on Software Engineering (ICSE), ACM.