

# Will Dependency Conflicts Affect My Program's Semantics?

Ying Wang, Rongxin Wu, Chao Wang, Ming Wen, Yepang Liu, Shing-Chi Cheung *Senior Member, IEEE*,  
 Hai Yu, Chang Xu *Senior Member, IEEE*, Zhiliang Zhu *Member, IEEE*

**Abstract**—Java projects are often built on top of various third-party libraries. If multiple versions of a library exist on the classpath, JVM will only load one version and shadow the others, which we refer to as *dependency conflicts*. This would give rise to *semantic conflict* (SC) issues, if the library APIs referenced by a project have identical method signatures but inconsistent semantics across the loaded and shadowed versions of libraries. SC issues are difficult for developers to diagnose in practice, since understanding them typically requires domain knowledge. Although adapting the existing test generation technique for dependency conflict issues, RIDDLE, to detect SC issues is feasible, its effectiveness is greatly compromised. This is mainly because RIDDLE randomly generates test inputs, while the SC issues typically require specific arguments in the tests to be exposed. To address that, we conducted an empirical study of 75 real SC issues to understand the characteristics of such specific arguments in the test cases that can capture the SC issues. Inspired by our empirical findings, we propose an automated testing technique SENSOR, which synthesizes test cases using ingredients from the project under test to trigger inconsistent behaviors of the APIs with the same signatures in conflicting library versions. Our evaluation results show that SENSOR is effective and useful: it achieved a *Precision* of 0.803 and a *Recall* of 0.760 on open-source projects and a *Precision* of 0.821 on industrial projects; it detected 150 semantic conflict issues in 29 projects, 81.8% of which had been confirmed as real bugs.

**Index Terms**—Third-party Libraries, Test Generation, Empirical Study.

## I. INTRODUCTION

BUILDING software projects on top of third-party libraries is a common practice to save development cost and improve software quality [1], [2], [3], [4]. However, the heavy dependencies on third-party libraries often induce dependency conflict issues [5]. When multiple versions of the same library class are present on the classpath, the Java class loader will load only one version and shadow the others [6]. If the loaded

Ying Wang, Chao Wang, Hai Yu, and Zhiliang Zhu are with the Software College, Northeastern University, China. E-mail: wangying@swc.neu.edu.cn, wangc\_neu@163.com, {yuhai, zzl}@mail.neu.edu.cn.

Rongxin Wu is with Department of Cyber Space Security, Xiamen University, China. E-mail: wurongxin@xmu.edu.cn.

Ming Wen is with School of Cyber Science and Engineering, Huazhong University of Science and Technology, China. E-mail: mwena@hust.edu.cn.

Yepang Liu is with Department of Computer Science and Engineering, Southern University of Science and Technology, China. E-mail: liuyup1@sustech.edu.cn.

Shing-Chi Cheung is with Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, China. E-mail: scc@cse.ust.hk.

Chang Xu is with State Key Lab for Novel Software Technology and Department of Computer Science and Technology, Nanjing University, China. E-mail: changxu@nju.edu.cn.

Manuscript received May 9, 2020; revised August 26, 2020.

```
// Method in loaded library Jackson-core-asl.jar - 1.9.4
102: public boolean validate(ClientResponse<T> response){
103:     return true;
104: }

// Method in loaded library Jackson-core-asl.jar - 1.9.13
102: public boolean validate(ClientResponse<T> response){
103:     return response.getStatus();
104: }

// Entry method in class connector.RESTEasyConnector of the client project
134: public ClientExecutor createClientExecutor(){
135:     if(request.status())
136:         return request.getDefaultExecutor();
137:     else
138:         return null;
139: }

// Method referenced by entry method (defined in jackson-jaxrs:jar - 1.9.4)
56: public boolean status(){
57:     return (Paster.validate(response) && getAccesss())
58: }
```

**Fig. 1:** Issue #214 [9] in the project Openstack-java-sdk 3.2.5 version has inconsistent implementations with the intended but shadowed versions, dependency conflict issues will occur, inducing risks of runtime exceptions or unexpected program behaviors.

The state-of-the-art techniques [5], [7] for detecting dependency conflict issues mainly focus on specific categories of the issues, such as `ClassNotFoundException` and `NoSuchMethodError`, which happen when the loaded library versions do not cover all the APIs referenced by the client projects. One limitation of these techniques is that they cannot identify the dependency conflict issues that arise from referencing those APIs with identical method signatures but inconsistent behaviors across multiple library versions [8], [7]. We refer to such issues as *Semantic Conflict* issues (SC issues for short). Figure 1 gives a real example of SC issues. On the classpath of the project Openstack-java-sdk 3.2.5, there are two versions of the library `Jackson-core-asl`, namely, *Version 1.9.4* and *Version 1.9.13*. In the example, Java class loader loads *Version 1.9.13* but shadows *Version 1.9.4*. As shown in the code snippet, the method `createClientExecutor()` in the project will transitively invoke `validate(ClientResponse)` of the library `Jackson-core-asl`. However, the implementations of `validate(ClientResponse)` are semantically inconsistent between the two versions. The project was originally designed to use *Version 1.9.4* of `Jackson-core-asl`, which is unfortunately shadowed. Although there will be no runtime exceptions in such cases, the semantic inconsistency of library method implementations will inappropriately affect the variable states of the client project via the invocation of the concerned methods, leading to unexpected program behaviors.

SC issues arise from code changes of API implementations, which are common in popular libraries. Many of these changes are too subtle for developers to understand their effects on program semantics [2]. Existing test suites may not help effectively expose such differences neither. As such, SC issues are difficult to diagnose. For example, a developer left the following comment in the pull request [10] of the aforementioned issue:

*"I have encountered these types of semantic inconsistency issues lots of times when dealing with dependency conflicts. When such issues happen, signature changes can be detected by static analyzers. However, semantic changes would be more difficult to detect. Empirically, developers diagnose them by reading the git history of the library or dynamic testing."*

Detecting SC issues typically requires rich domain knowledge to discern the subtle differences in API implementations, which is a non-trivial task. Therefore, an automated technique to detect SC issues is highly desirable. We note that the most relevant and recent technique is RIDDLE [7]. It was designed to verify dependency conflict issues caused by the missing of classes or methods. RIDDLE can generate tests to drive the execution of a client project towards the target call sites that could induce dependency conflict issues. While it seems possible to adapt RIDDLE using the idea of differential testing (i.e., comparing the runtime behavior of a target API across multiple library versions) to detect SC issues, our trial on 70 open-source projects shows that this technique is not that effective as anticipated (see Section II-C). It is mainly because merely reaching the call sites of a target API and invoking it with random arguments can hardly trigger the inconsistent behaviors of the API across different versions. As such, many SC issues, whose manifestation requires specific arguments (referred to as *divergence arguments* in this paper), cannot be effectively exposed. This motivates us to design a more effective testing technique to detect SC issues.

As discussed above, an obvious challenge in detecting SC issues via testing is to generate divergence arguments to trigger inconsistent API behaviors across different library versions. To address this challenge, we performed an in-depth study of 75 real SC issues collected from open-source Java projects to understand the characteristics of divergence arguments in the test cases that could expose these issues. The study revealed several interesting findings. First, to generate class instances as test inputs for detecting SC issues, almost all (98.5%) the object constructors take at least one argument, and, for 97.8% of these constructors, at least one of their arguments has specific values that can hardly be generated by random techniques such as RIDDLE. Second, we observed three common patterns to produce divergence arguments for object constructors in the test cases. Third, we found that, for 56.9% of our analyzed object constructors, their divergence arguments can be directly obtained from the source code of the client project. Fourth, for the constructors in the test cases whose arguments cannot be found in the source code of the client project, we replaced the arguments with other compatible values that can be found in the client project's source code and discovered that 37 out of 58 (63.8%) such

revised test cases could still capture SC issues.

Inspired by our empirical findings, our idea of generating divergence arguments for triggering SC issues is to synthesize these arguments from the source code of client projects. Specifically, we synthesize an object constructor of a divergence argument by distilling the set of legitimate API usages and the values of its arguments from the source code. We refer to the set as the constructor's *invocation context*. We implemented our idea into an automated testing technique, SENSOR. Given a client project to analyze, SENSOR first extracts the invocation contexts of each object constructor from the source code and leverages them to construct a pool of class instances. Combining a seeding strategy of class instances with EVO-SUITE, it then generates tests to trigger the concerned library APIs and checks whether they behave consistently across different versions. In our approach, SENSOR does not simply report all detected behavioral inconsistencies as bugs. Instead, it pinpoints the differences in variable states of a project under analysis and provides such fine-grained information to help developers further diagnose SC issues.

We evaluated SENSOR using 92 open-source projects on GitHub and 10 industrial projects from Neusoft Co. Ltd (SSE: 600718) [11]. SENSOR achieved a *Precision* of 0.803 and a *Recall* of 0.760 on open-source projects, and a precision of 0.821 on industrial projects. SENSOR detected 150 real SC issues from 29 open-source projects. We reported these issues to the developers of the corresponding projects and detailed the issues' impact on program behaviors. So far, 81.8% of our reported issues have been confirmed by the developers as real bugs, and 85.2% of the confirmed issues have been fixed quickly. Most of the confirmed issues are from popular projects such as Rest-assured [12] and Java-design-patterns [13]. From the feedback on our reported issues (see Section V-C), we observed that developers acknowledged the pervasiveness of SC issues and the necessity of a testing technique to diagnose such issues. They also expressed great interests in using SENSOR. These results demonstrate the effectiveness and usefulness of SENSOR. In summary, we make four major contributions in this paper:

- An empirical study of 75 real SC issues for exploring the characteristics of test cases that can expose SC issues.
- A fully automated technique, SENSOR, for detecting SC issues.
- A benchmark dataset for assessing SENSOR and similar approaches for detecting the issues induced by semantic inconsistencies of library APIs across different versions.
- A systematic analysis and discussions of SC issues' impacts on program behaviors.

Our tool and dataset are available at: <https://sensordc.github.io/>.

## II. PRELIMINARIES

### A. Motivation

To roughly estimate the scale of SC issues, we statically detected the semantic inconsistency of the conflicting API pairs by comparing their *code structures* in terms of call graphs and control flow graphs. We first collected 1,654 Java

```

// Project: Hmily 2.0.0
// Loaded version in library org.jboss.netty 3.2.5
45: public String getPrefixFromTerm(String term, TimerContext context) {
46:     if (StringUtils.isNotBlank(term)) {
47:         String[] parts = term.split(":");
48:         return parts[0];
49:     }
50:     if (enableTimers && context != null) {
51:         context.stop();
52:     }
53:     return null;
54: }

// Shadowed version in library io.netty 3.10.5
45: public String getPrefixFromTerm(String term, TimerContext context) {
46:     if (StringUtils.isNotBlank(term)) {
47:         String[] parts = term.split(":");
48:         return parts[0];
49:     }
50:     stopTimer(context);
51:     return null;
52: }
53: private void stopTimer(TimerContext context) {
54:     if (enableTimers && context != null) {
55:         context.stop();
56:     }
57: }

```

**Fig. 2:** A motivating example

projects from GitHub based on two criteria: (1) it has achieved over 50 stars or forks (popularity); and (2) it is built on the Maven platform. Then, we compared the code structures of the conflicting API pairs in these projects and labeled them as potential SC issues if their code structures are different. The results showed that 73.1% of the projects contain at least one potential SC issue. Each of them contains on average 20 conflicting library API pairs that potentially cause SC issues.

The static analysis of SC issues based on different code structures can be highly imprecise for some projects. Figure 2 illustrates a false positive SC issue found by the static approach. There are two versions of the class `Netty.bootstrap.ServerBootstrap` on the classpath of the project `Hmily-2.0.0` [14], which are included by the libraries `org.jboss.netty 3.2.5` and `io.netty 3.10.5`, respectively. Due to Maven's *first declaration wins* strategy, only the method `getPrefixFromTerm()` declared in the library `org.jboss.netty 3.2.5` is loaded and invoked by the client project. Although the call graphs of these two versions differ, the method in `io.netty 3.10.5` was simply a code refactoring in `org.jboss.netty 3.2.5` that did not affect the program semantics.

Validation of SC issues is non-trivial. It requires domain knowledge to understand the implementations of the client project and its libraries. This motivates us to validate SC issues using automatically generated tests.

### B. Problem formulation

To formulate our research problem, we introduce the following concepts. In particular, we let  $C'_i$  be a shadowed class version and  $C_i$  be the actually-loaded class version, and use  $C :: m$  to denote an API  $m$  of class  $C$ .

**Definition 1. (Conflicting API pair):** Let  $C'_i :: m_k$  be an API included in the shadowed class version and referenced by the client project  $\mathcal{H}$ , and  $C_i :: m_k$  be an API belonging to the actually-loaded class version, where  $m_k$  represents the method signature. If  $C'_i :: m_k$  and  $C_i :: m_k$  share the same signature, we consider  $C'_i :: m_k$  and  $C_i :: m_k$  as a pair of conflicting APIs, which is denoted as  $RP\langle C'_i :: m_k, C_i :: m_k \rangle$ . Conflicting class versions  $C_i$  and  $C'_i$  caused by a dependency conflict issue, may introduce a set of conflicting API pairs. We denote the set of conflicting API pairs as  $R\langle C_i, C'_i \rangle$ . Specially, if there are implementation differences between  $RP\langle C'_i :: m_k, C_i :: m_k \rangle$ , we consider it as an *isomericous conflicting API pair*.

**Definition 2. (Original dependency path):** For each API  $C'_i :: m_k$  included in a shadowed class version and referenced by a class  $C_1$  in the client project, we define any path  $Ep =$

$\langle C_1 :: m_1, \dots, C_{i-1} :: m_{k-1}, C'_i :: m_k \rangle$  as its original dependency path, where  $C_1 :: m_1$  represents an entry method in the class  $C_1$  of the client project indirectly referencing the method  $C'_i :: m_k$  along  $Ep$ .

**Definition 3. (Actual dependency path):** Suppose that  $RP\langle C'_i :: m_k, C_i :: m_k \rangle$  is a conflicting API pair. For each original dependency path  $Ep$  with respect to API  $C'_i :: m_k$ , we define  $Fp = \langle C_1 :: m_1, \dots, C_{i-1} :: m_{k-1}, C_i :: m_k \rangle$  as the corresponding actual dependency path, as the build environment enforces the interactions between entry method  $C_1 :: m_1$  in the class  $C_1$  of the client project and API  $C_i :: m_k$  included in the actually-loaded class  $C_i$  along  $Fp$ . Note that,  $Ep$  and  $Fp$  share the subpath from entry method  $C_1 :: m_1$  to  $C_{i-1} :: m_{k-1}$ .

**Problem:** Given a project with a set of conflicting API pairs  $R\langle C_i, C'_i \rangle$ , our research problem is how to design an automated test generation technique to trigger the executions of each isomericous conflicting API pair  $RP\langle C'_i :: m_k, C_i :: m_k \rangle \in R\langle C_i, C'_i \rangle$  along their original and actual dependency paths, respectively, thereby identifying their impacts on the client project's program behaviors.

### C. Challenges

RIDDLE is the state-of-the-art technique that generates tests to detect dependency conflict issues in projects where the loaded library versions fail to cover all the referenced APIs based on their method signatures [7]. However, this technique is not applicable to detecting SC issues, since SC issues arise from referencing the APIs with identical method signatures but inconsistent behaviors across multiple library versions.

RIDDLE generates tests forcing the program execution along the path that invokes the shadowed library APIs. We may adapt the mechanism to detect SC issues. For example, after identifying the conflicting library APIs for SC issues, we can use RIDDLE to generate tests to drive the program to execute along the path from an entry method to the conflicting library APIs. Then, we can execute the generated test using the shadowed version and the loaded version, respectively, and compare their test outcomes to check the semantic inconsistency. However, based on our trials on 70 Java projects with RIDDLE, we observed that this approach cannot effectively detect SC issues.

Consider the SC issue #214 [9] described in Section I. To trigger the conflicting library API `validate(ClientResponse)` in the intended version `Jackson-core-asl 1.9.13` or the loaded version `Jackson-core-asl 1.9.4`, a test must instantiate an object of the class `connector.RESTEasyConnector` and call the entry method `createClientExecutor()` provided

```

// connector.RESTEasyConnector.java
61: public RESTEasyConnector(String Path){
62:     OpenStackClient openStackClient = new OpenStackClient(Path);
63:     request = new OpenStackRequest(openStackClient, HttpMethod.GET);
64:     request.addHeader("Accept", "application/json");
65: }
(a) Constructor of the class connector.RESTEeasyConnector
// openstack.keystone.api.Create.java
48: public Create(Endpoint endpoint) {
49:     RestEasyConnector = new RESTEasyConnector("/endpoints");
50:     this.endpoint = endpoint;
51: }
(b) A caller method of connector.RESTEeasyConnector's constructor
// @Test
1: RESTEasyConnector RestEasyConnector = new RESTEasyConnector("a*#");
2: RestEasyConnector.createClientExecutor(); //Throw Exception
(c) A failing test case generated by Riddle

```

**Fig. 3:** Tests generated by Riddle for the example issue #214 [9]

by the object. Figure 3(c) shows a test generated by RIDDLE to trigger the invocation of the entry method. As the randomly generated parameter "a\*#" is an invalid argument of path to instantiate class `OpenStackClient`, the test will trigger a `NullPointerException` at Line 62 of the constructor `RESTEasyConnector(String)` (as shown in Figure 3(a)), when it attempts to construct the instance of class `RESTEasyConnector`. As a result, the SC issue is missed.

As observed from the above example, the SC issue requires a specific argument to be triggered, i.e., a valid path string "/endpoint", which can be found in a caller method of `RESTEasyConnector`'s constructor (as shown in Figure 3(b)). RIDDLE, which generates test inputs randomly, is ineffective in detecting such issues. The example motivates us to conduct an empirical study to understand the characteristics of specific arguments that expose SC issues.

### III. EMPIRICAL INVESTIGATION

In this section, we present an empirical study on a collection of test cases that have ever successfully captured real SC issues, with the aim of answering the following two research questions.

**RQ1:** Are randomly generated arguments in test cases likely to capture inconsistent program behaviors of SC issues? What are the characteristics of divergence arguments?

Generating the desirable objects is a significant challenge for automated test generation techniques [15]. Specifically, to increase the likelihood of triggering the conflicting API pairs and revealing SC issues, we should be able to generate divergence arguments for object constructors. To ease our presentation, we refer to an object constructor taking no arguments as a *no-args constructor* and an object constructor that takes arguments as a *parameterized constructor*. In this paper, we focus on the characteristics of the divergence arguments required by parameterized constructors in the test cases, i.e., the concrete values held by arguments, including strings, primitive types, and object references.

**RQ2:** Can divergence arguments in test cases be found in the source code of the client project?

The above investigation can provide empirical evidences and guidance to help construct divergence arguments for parameterized constructors in test generation.

#### A. Collection for benchmark dataset

Identifying existing tests written by developers or generated by tools that can detect SC issues is difficult. To achieve such a goal, we first simulate a series of dependency conflicts for a given project by altering the actually-loaded versions of its referenced libraries. We then execute the project's associated tests to see if it can capture the inconsistent behaviors introduced by the version substitution. The steps and criteria for constructing such a dataset are described as follows in detail:

**Step 1: Selecting subjects.** We randomly selected Java projects from GitHub satisfying three conditions: (1) including more than 50 test class files designed by the original developers with their domain knowledge; (2) passing all the associated test cases without errors (ensuring no SC issues in the selected version); (3) depending on more than 30 libraries (having more upgraded/downgraded candidate libraries). As such, we obtained 523 open-source projects.

**Step 2: Altering the actually-loaded library version.** For each library on the dependency tree of a subject, we first collected a set of its version numbers released on the Maven central repository, which is denoted as  $V = \{v_1, v_2, \dots, v_n\}$ . We iteratively used each library version  $v_i \in V$  to replace its original version on the dependency tree. Then, we checked whether the associated tests thrown `AssertionErrors` when running on the subject after replacements. The rationale is that the `AssertionError` in JUnit tests is used to indicate whether the actual variable values are equal to their expected values. If a test passes for the selected version of a subject and fails for the revised version with an `AssertionError`, we consider that the failing test captures an SC issue caused by the substitution of library version  $v_i$ . By manually debugging two versions of the program triggered by this failing test, on the execution traces, we can identify a pair of isomeric conflicting APIs defined in both the original and altered library versions.

Eventually, from 48 Java projects, we obtained 75 SC-revealing test cases, which correspond to 75 conflicting API pairs. Table 1 shows the statistics of the subjects. They are large (up to 509.1 kLOC), popular (up to 7,231 stars) and well-maintained (up to 1,108 associated test cases). Moreover, they have large-scale dependency trees (up to 61 referenced libraries), and on average, each library has 17 versions released on the Maven central repository. The statistics indicate that the collected test cases are representative.

#### B. Empirical findings of RQ1

To answer RQ1, we manually checked 325 class instances used in 75 collected SC-revealing test cases, to analyze the characteristics of the divergence arguments required by their corresponding constructors. By investigation, we found that in the above test cases, 320 out of 325 class instances (98.5%) need to be created using parameterized constructors, and only five class instances (1.5%) are constructed without arguments. Among 320 parameterized constructors in the test cases, the number of their required arguments ranges from 1 to 7 (i.e.,  $3 \pm 0.34$ ). Specially, 314 out of 320 parameterized constructors (98.1%) require more than 2 arguments for creating valid class instances. For 1,036 divergence arguments needed by the 320

**TABLE I:** The statistics of the subjects collected in our study

# Project	# Star			Size(# kLOC)			# Test case			# Library			Avg <sub>v</sub>		
	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.
48	73	7,231	692	0.7	509.1	78.4	62	1,108	201	34	61	40	6	109	17

```
(a) // Project: struts-master/core 2.5.22
// Test case: DateInterceptorTest.testDateField
77: DateFormat date = new SimpleDateFormat("yyyy-mm-dd");

(b) // Project: index-providers 5.4.1
// Test case: EsIndexTest.setUpClass
85: EsClient client = new EsClient("localhost", 9200);
86: EsIndex index = new EsIndex(client);

(c) // Project: kitodo-production 3.0.0
// Test case: MockEntity.prepareNodeSettings
59: ConfigMain cm = new ConfigMain("NodeSettings");
60: String host = cm.getParameter("elasticsearch.host", "9205");
61: HttpHost httpHost = new HttpHost(host);
```

**Fig. 4:** The illustrative examples for explaining the constructors' arguments (denoted by arguments ) in test cases

parameterized constructors, we investigated the corresponding source files to understand their characteristics of assignments. Based on our observations, we divide the arguments into the following three types:

**Type 1.** *The arguments are strings or primitive type values with specific semantic meanings or in specific formats (25.3%).* 262 out of 1,036 arguments are strings or primitive type values (i.e., numeric and enumeration variables) in the test cases. By manually checking their values, we found that 168 out of 173 string arguments (97.1%) are constrained in specific formats or have specific semantic meanings that reflect developers' domain knowledge, such as, the protocol or date related strings. For instance, in the test file as shown in Figure 4(a), a string "yyyy-mm-dd" is assigned to the parameter of constructor `SimpleDateFormat(String)`, which can rarely be generated randomly.

Besides, most of the primitive type arguments are used to specify boundary or specific values, e.g., -1, 9200, etc. For 89 primitive type arguments, we used a random value to replace each argument and then run the corresponding revised test cases on both the original and the altered library versions. The above process is repeated ten times for each argument. Over the ten runs, if the revised test cases can capture SC issues once a time, we consider the corresponding argument can be replaced with random values. Unfortunately, 79 out of 89 arguments (88.8%) failed to trigger the inconsistent behaviors after the random replacements.

**Type 2.** *The arguments are the instances of other classes created by the constructors with specific inputs (32.4%).* 336 out of 1,036 arguments are the instances of other classes. 328 out of the above 336 object constructors (97.6%) require arguments. In such cases, developers should recursively construct class instances, and the combination of the involved arguments determines the outermost instance's state. Therefore, it requires rich domain knowledge to create such combination of specific arguments. For example, as shown in Figure 4(b), the argument of constructor `EsIndex(EsClient)` in the test

case is created by constructor `EsClient(String, int)` with specific arguments. The arguments "localhost" and 9200 determine the state of constructed object `client`.

**Type 3.** *The arguments are returned by the other method calls with specific inputs (438/1,036 = 42.3%).* In such cases, the states of the constructed class instances are determined by the method calls with valid arguments and the states of class instances providing the above method calls. Figure 4(c) shows an example of this type of assignment in test file `MockEntity.prepareNodeSettings`, the argument of constructor `HttpHost(String)` is returned by method `ConfigMain.getParameter(String, String)` with valid inputs "elasticsearch.host" and "9205". In such scenario, developers should also instantiate the classes that provide the required method calls, using specific argument "NodeSettings". Similar to **Type 2** cases, the combination of the involved arguments required by method calls and the recursively constructed class instances, makes the parameter assignments more complicated.

For most of the parameterized constructors, they require mixed types of arguments. In many cases, when instantiating a class instance, to construct one type of arguments, we need to recursively create the other types of arguments. An effective technique for the generation of valid class instances is essential to capture real program behaviors.

**Finding 1:** 320 out of 325 constructors (98.5%) require arguments to produce valid class instances, in the test cases that successfully capture the inconsistent behaviors.

**Finding 2:** 1,001 out of 1,036 arguments (97.8%) required by 320 parameterized constructors are subject to semantic constraints, which can hardly be replaced by random values.

### C. Empirical findings of RQ2

To answer RQ2, for 320 parameterized constructors in the collected test cases, we first located their corresponding caller methods in the client projects' source code. Furthermore, we manually checked whether the valid arguments of constructors could be found in the code snippets of these caller methods. The rational is that the caller methods mostly contain the invocation contexts of a constructor. Note that an object constructor may have more than one caller methods in the source code.

For each collected parameterized constructor, suppose  $Argu_t$  is the total number of the required arguments, and  $Argu_s$  is the number of arguments that can be identified in the source code of client project. We found that 182 out of 320 parameterized constructors (56.9%) whose corresponding  $Argu_s/Argu_t$  values are greater than 0. This means that for 182 parameterized constructors, at least one of their required arguments can be found in the source code. Specially, among the above 182 parameterized constructors, 119 constructors' corresponding  $Argu_s/Argu_t$  values are equal to 1. For the

rest 63 cases whose corresponding  $\text{Argu}_s/\text{Argu}_t$  values are between 0 and 1, we observed that 62.2% of their arguments are passed by the input parameters of the constructors' caller methods. However, the above caller methods are the APIs provided for invocation by the third-party projects. As a result, their valid arguments could not be found in the source code of client projects.

For the remaining 138 parameterized constructors (43.1%) whose corresponding  $\text{Argu}_s/\text{Argu}_t$  values are equal to 0, we performed the following tasks: (1) We manually extracted the valid arguments from the code snippets of their caller methods. If an argument could not be found in the source code, we randomly assigned a value to it. (2) Let  $\text{Argu}'_s$  be the number of arguments that are manually extracted from source code. For each class instances created by the original developers in the test cases, we replaced it with our constructed ones whose corresponding  $\text{Argu}'_s/\text{Argu}_t$  values are greater than 0 (i.e., at least one argument could be found in the source code). (3) After the replacement, we executed the revised test cases and checked whether they could still capture the inconsistent behaviors with `AssertionErrors`. Finally, 101 out of 138 parameterized constructors (73.2%) were replaced in 58 test cases. For the above 58 revised test scripts, 37 of them (63.8%) successfully detected the SC issues when running on the project versions with upgraded/downgraded libraries. The average value of  $\text{Argu}'_s/\text{Argu}_t$  of the 76 replaced constructors in the 37 test cases that can capture SC issues is 0.25 higher than that of the 25 replaced constructors in the 21 test cases that fail to detect SC issues.

From the above results, we can draw the conclusion that combining the constructors with their valid arguments extracted from the source code to generate tests can help to expose the inconsistent behaviors. The more valid arguments injected to the constructors, the higher success rate of capturing SC issues.

**Finding 3:** In the collected test cases that can capture the SC issues, 182 out of the 320 parameterized constructors (56.9%) of which parts of their arguments can be found in the source code.

**Finding 4:** When we substituted our injected arguments for the constructor arguments that cannot be found in the source code, 37 out of 58 test cases (63.8%) captured SC issues.

The empirical findings of RQ1 and RQ2 shed lights on understanding the characteristics of the object constructors' arguments and provide valuable guidance to design an automated test generation technique for detecting SC issues.

## IV. METHODOLOGY

### A. Sensor in a nutshell

Figure 5 shows an overview of our approach, which involves three steps: *identification for isomeric conflicting API pairs*, *test generation* and *outcome comparison*. First, SENSOR finds a set of conflicting API pairs introduced by a dependency conflict and identifies which of these conflicting API pairs are isomeric. Second, it generates tests to capture the inconsistent variable states of a given client project affected by isomeric conflicting API pairs. Third, by comparing their test outcomes obtained on two conflicting class versions, SENSOR

identifies the SC issues and points out their impacts on the client project's program behaviors at a fine-grained level to help diagnose SC issues.

### B. Identifying isomeric conflicting API pairs

SENSOR identifies the isomeric conflicting API pairs at a fine-grained level based on code differences detected iteratively using GUMTREE [16]. It considers that those conflicting API pairs with different implementations will potentially cause semantic conflicts.

**Identifying conflicting API pairs.** By analyzing the dependency tree of a client project, SENSOR identifies multiple versions of a class  $C_i$  or a library  $Lib_i$ . For each API  $C'_i :: m_k$  defined in the shadowed class version  $C'_i$  and referenced by the client project, SENSOR considers the API  $C_i :: m_k$  that satisfies one of the following two conditions as its replaceable method:

- API  $C_i :: m_k$  has the same signature (i.e., method name, parameter types and return types) as  $C'_i :: m_k$  and is defined in the actually-loaded class version.
- Suppose class  $CS_i$  is the superclass of  $C_i$ . If the actually-loaded class  $C_i$  does not include the API with the same signature as  $C'_i :: m_k$ , then SENSOR regards the API defined in its superclass  $CS_i :: m_k$  that can be overridden by  $C'_i :: m_k$  as its replaceable method. In this case, the API compatibility will not be broken due to dynamic binding mechanisms.

Finally,  $C'_i :: m_k$  and  $C_i :: m_k$  are identified as a pair of conflicting APIs, which is denoted as  $RP\langle C'_i :: m_k, C_i :: m_k \rangle$ .

**Analyzing isomeric conflicting API pairs.** We adopt GUMTREE [16] to check if implementation differences exist in a conflicting API pair. GUMTREE detects code differences based on abstract syntax trees (ASTs). When applying GUMTREE, SENSOR needs to consider the cases where a pair of conflicting APIs exhibit no difference in terms of their ASTs but are semantically different due to the changes in their depended methods. Typically, an API could invoke a series of methods which constitute a call graph. Any changes in the methods invoked by an API could possibly affect the states of its referenced variables, thereby changing the API's semantics. To perform a comprehensive analysis for a conflicting API pair, we construct the corresponding call graphs of the two APIs, and then iteratively compare each method pair with the same signature on the call graphs in a top-down manner. In the process of iterative analysis, we consider the above conflicting API pair as an *isomeric conflicting API pair*, if there are AST differences identified by GUMTREE between one comparable method pair on the call graphs.

Although our iterative analysis can capture all code differences between a pair of method invocation paths on the call graphs, not all the differences are useful in practice. According to an empirical study conducted by Schröter et. al [17], nearly 90% of the issues are fixed within the **top-10** methods along the invocation paths. Other deeper methods barely affect the program semantics of the client project. To reduce such false positives, SENSOR only analyzes the methods whose call depth is less than **ten**, along the original and actual dependency paths of a conflicting API pair.

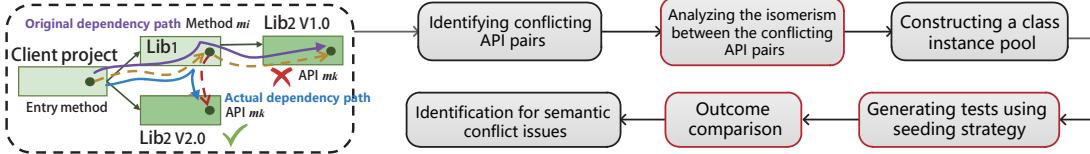


Fig. 5: The overall architecture of SENSOR

### C. Test generation

SENSOR is built on top of EVOSUITE, which adopts a genetic algorithm (GA) to derive a test suite for a given target class. A target class is the one that contains an entry method that directly or indirectly references the identified isomeric conflicting APIs along their original and actual dependency paths, respectively. SENSOR adopts the fitness function defined by RIDDLE, which aims to maximize the possibility of covering the identified isomeric conflicting APIs [7].

To precisely capture the program behaviors, SENSOR adopts a seeding strategy of class instances inspired by our empirical findings summarized in Section III. SENSOR injects the invocation context information extracted from the source code into class instances with the aim of generating divergence arguments. Specifically, SENSOR first constructs a pool of instances with the injected invocation contexts for each class included in a client project, which is denoted as *CIP*. When EVOSUITE needs to instantiate a class in a test, SENSOR tries to select an instance from *CIP* and provide it to EVOSUITE.

**CIP Construction.** SENSOR constructs a class instance involving two steps: identifying its possible object constructors and extracting the constructors' invocation contexts from the source code of the client project. Let  $MOI_i = \{mo_1, mo_2, \dots, mo_n\}$  be a set of possible object constructors of a given class  $C_i$  collected by the static analysis approach [18], where  $mo_k \in MOI_i$  represents an object constructor of this class. For most of the cases, a constructor requires parameters for substantiation. To inject valid arguments into  $\forall mo_k \in MOI_i$ , SENSOR performs the following tasks: (1) it identifies a set of caller methods  $M_k = \{m_1, m_2, \dots, m_n\}$  in the client project, which reference the constructor  $mo_k \in MOI_k$ ; and (2) for each caller method  $m_t \in M_k$ , SENSOR locates the source file where it is defined, and considers the code snippets within the source file as a search scope of the invocation contexts of  $mo_k$ . Specially, SENSOR takes into account the following three cases to extract the invocation contexts of  $mo_k \in MOI_i$ , based on three types of constructors' arguments observed in our empirical study:

**Case 1:** If the required arguments of  $mo_k$  are strings or primitive types that can be identified in source code by exactly matching assignment statements with variable names, SENSOR extracts their corresponding assigned values directly from the source code.

**Case 2:** In the case where the arguments of  $mo_k$  are the input arguments of the caller method  $m_t \in M_k$ , SENSOR recursively searches the corresponding invocation context for caller method  $m_t$ .

**Case 3:** For the arguments whose assigned values are

returned by the other method calls, SENSOR recursively finds invocation contexts of the required method calls and the class instances that provide the method calls, following the above steps.

For the required arguments that are the instances of other classes, SENSOR recursively constructs such class instances following the above steps. Moreover, if the required arguments are strings or primitive types whose valid values cannot be exactly extracted from source code, SENSOR randomly assigns values to them. In the cases of recursively constructing class instances or searching for valid arguments from the intermediate invocation contexts (e.g., Cases 2 and 3), the searching process is terminated if the recursion depth is greater than  $DN$ , or it cannot generate an instance of the current class (e.g., the class is not instantiable for accessibility reasons). In this manner, for each class  $C_i$ , we can obtain a set of possible object constructors  $MOI_i$ , and each constructor  $mo_k \in MOI_i$  corresponds to a set of invocation contexts extracted from the source code.

**Seeding strategy.** During the insertion of new statements into a test case, EVOSUITE tries to resolve dependencies either by re-using objects declared in earlier statements of the same test, or by recursively inserting new calls to generate new instances of the required dependency objects [19], [20]. Whenever EVOSUITE attempts to generate a class instance, SENSOR selects one of the instances of that class from *CIP* with probability  $POC$  to replace it. Note that if *CIP* does not contain the required class instance, such replacement operations are not performed.

In the presence of different object constructors and their various corresponding invocation contexts for a given class, setting the probability  $POC$  to choose one of them is challenging. In our approach, we set a probability  $POC$  based on the complexity of a constructor's invocation contexts. The selection strategy favors the object constructor with a low complexity. Also, it dynamically adjusts the probability  $POC$  according to the number of times that a class instance has been selected during the test generation process, to diversify candidate class instances. Thus, we have

$$POC = 1/Depth_{arg} \times 1/T_s \quad (1)$$

where  $Depth_{arg}$  is an indicator of a constructor's complexity, which represents the recursion depth for constructing the involved class instances or searching valid arguments from the intermediate invocation contexts; and  $T_s$  is the number of times that a class instance has been seeded into the tests.

### D. Test outcome comparison

For each isomeric conflicting API pair, SENSOR generates tests to trigger their executions on the actually-loaded and the

shadowed class versions where they are defined, respectively. It repeats the above test generation process for  $RN$  times and then compares the test outcomes. SENSOR takes the following two types of semantic inconsistencies into account, in the comparison process:

- *Variable states.* It considers three types of affected variables, including: (a) each input parameter of entry method whose type is an object; (b) each variable used by entry method but not defined in it; and (c) return variable of the entry method. As a result, if the state of any affected variable is different across the executions on the above two versions of code, SENSOR regards the behaviors as inconsistent.
- *Test outcomes.* If a test succeeds to run on one version of code but fails on the other, a semantic inconsistency is noted.

SENSOR considers the isomericous conflicting APIs that induce inconsistent behaviors when executing more than one generated tests, as the cases that could cause SC issues. Finally, it illustrates their impacts on the program semantics of the client project, to help developers further diagnose the SC issues.

## V. EVALUATION

This section presents our experimental results through answering the following research questions:

- **RQ1 (Effectiveness):** How effective is SENSOR in detecting SC issues?
- **RQ2 (Usefulness):** Can SENSOR detect unknown SC issues and provide useful diagnosis information?

### A. Experimental design

1) *RQ1:* To study RQ1, we first collected a high quality ground truth dataset and then applied SENSOR to this dataset to assess its effectiveness in detecting SC issues.

**Collection for the ground truth dataset.** We consider 75 isomericous conflicting API pairs that can cause `AssertionErrors` when executing 75 tests collected in our empirical study (in Section III), as the ones introducing SC issues into their client projects. We labeled the isomericous conflicting API pairs that will not cause SC issues in client projects, based on the following steps:

(1) We mined the historical commits of open-source projects on Github and identified the commits that only upgraded/downgraded a library version in the projects' dependency management scripts (e.g., `pom.xml`). We consider that the above change of a library version does not affect the client project's program behaviors, if its corresponding commit satisfies all the following conditions:

- All the tests triggered by a continuous integration build tool (e.g., `TravisCI`) can pass the revised project version, after this commit is submitted.
- In the case that the client project is still active, there are no version changes for this library, in the next 24 months after the above commit being merged. Besides, during the above period, there are no issues or commits whose descriptions and logs mentioning the semantic issues caused by this revised library version. The rational is that a recent study [21] found that bugs are usually repaired within 2

years across different projects since they were introduced into the project.

(2) For an identified semantic-preserving library version change, we labeled the isomericous conflicting API pairs in the original and revised versions of this library, which can be covered by the client project's tests (triggered by the continuous integration build tool) without errors, as the ones that will not introduce SC issues.

Eventually, we collected 150 isomericous conflicting API pairs that will not introduce SC issues into their client projects and 75 isomericous conflicting API pairs definitely causing SC issues.

**Metrics.** The outcomes of SENSOR can be categorized as follows: (1) *True Positive* (TP): The inconsistent behavior identified by SENSOR between a conflicting API pair is a real SC issue. (2) *False Positive* (FP): The inconsistent behavior identified by SENSOR between a conflicting API pair is not a real SC issue. (3) *True Negative* (TN): No inconsistent behavior is identified by SENSOR between a conflicting API pair, and it is not a real SC issue. (4) *False Negative* (FN): No inconsistent behavior is identified by SENSOR between a conflicting API pair, but it is a real SC issue. Based on the outcomes, we use *Recall*, *Precision*, and *F-measure* to evaluate the performance of SENSOR, which are defined as follows.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

*Precision* evaluates whether SENSOR can detect SC issues precisely. *Recall* evaluates the capability of SENSOR in detecting all the SC issues. *F-measure* takes the *Precision* and *Recall* into consideration, and weights these two metrics equally [22].

**Comparison.** We compared SENSOR with the latest dependency conflict detection tool RIDDLE, in terms of their effectiveness in covering the target branches. RIDDLE is chosen as the baseline because it is designed for generating tests to trigger the program execution from an entry method of client project to reach an identified conflicting API that is missing in the actually-loaded library, thereby causing a program crash. We consider that RIDDLE detects a SC issue, if its generated tests can trigger the isomericous conflicting API pairs in the ground truth dataset and capture the variable state or test outcome inconsistencies.

**Experimental setting.** For both SENSOR and RIDDLE, we set the time budget for the evolutionary search to 800 seconds and repeated the test process for  $RN = 10$  times on each code version with different random seeds. The final results were averaged over the ten runs to avoid the biased results.

2) *RQ2:* To answer RQ2, we conducted experiments on 92 open-source Java projects randomly sampled from GitHub using three criteria: (1) it has received more than 50 stars or forks (i.e., popularity); (2) it references multiple versions of libraries or classes detected by static analysis and contains at least one commit after December 2019 (i.e., actively-maintained); (3) it is not included in the subject set of our empirical study in Section III (i.e., new validation). We leveraged SENSOR

**TABLE II:** The experimental results on the ground truth dataset

	<i>TP</i>	<i>FP</i>	<i>TN</i>	<i>FN</i>	<i>Precision</i>	<i>Recall</i>	<i>F-measure</i>
SENSOR	57	14	136	18	0.803	0.760	0.781
RIDDLE	6	2	148	69	0.750	0.080	0.145

to generate issue reports that include: (1) the root causes of SC issues; (2) the isomeric conflicting API pairs that induce semantic inconsistencies and their corresponding original and actual dependency paths; (3) the generated test cases that can trigger the executions of the isomeric conflicting API pairs; (4) the differences in the test outcomes or variable states of the client project after executing the generated test cases. We submitted the issue reports to the corresponding developers via the projects' issue tracking systems and evaluated the usefulness of SENSOR based on developers' feedback.

### B. RQ1: Effectiveness of Sensor

1) *Overall effectiveness.*: Table 2 shows the experimental results on the ground truth dataset. RIDDLE identified 8 SC issues with 2 (25.0%) false positives. Besides, it did not capture any inconsistent behaviors between 217 isomeric conflicting API pairs, with 69 (31.8%) false negatives. Our approach, SENSOR, identified 71 SC issues with 14 (19.7%) false positives, which achieves a *Precision* of 0.803. For the isomeric conflicting API pairs that will not cause SC issues in client projects, SENSOR successfully identified 154 of them, with 18 (11.7%) false negatives, leading to a *Recall* of 0.760. In terms of the *F-measure*, SENSOR also significantly outperformed RIDDLE (0.781 vs. 0.145).

By manually checking the six true positive SC issues detected by RIDDLE, we found that in these cases, all the invocation depths from the entry methods of client projects to the conflicting APIs are less than 3. All the object constructors in the test cases generated by RIDDLE that successfully captured the SC issues, do not need arguments. However, the initialization of the required variables could be found in the method bodies of the above specific simple constructors. Therefore, they could easily trigger the target branches with valid program semantics. Note that the above six true positive SC issues were also detected by our technique. For 57 true positive cases detected by SENSOR, the average invocation depth from the entry methods to the conflicting APIs is 6.9. It is largely ascribed to the effectiveness of SENSOR's seeding strategy of class instances. The seeded objects greatly increase the possibility of reaching the target branches, compared with RIDDLE. We further investigated the reasons why SENSOR generated false positive and negative cases of SC issues and summarized them below.

**False positive examples.** The main cause of false positives generated by SENSOR is that the inconsistent behaviors are caused by the non-deterministic or random variable states, which are benign for the client projects. For example, as shown in Figure 6, a test case generated by SENSOR captured the inconsistent return values of the entry method in project cdap 6.0.0, on conflicting versions of library com.google.code.gson. The return value is a Json string,

```
// Project: cdap-6.0.0
// Conflicting library versions: com.google.code.gson 2.2.4 vs. 2.0
1: public void test01(){           // A test case generated by Sensor
2: NamespaceId ns = new NamespaceId("ns1");
3: String Write = AccessType.WRITE;
4: AccessPayload accessPayload = new AccessPayload(Write, ns);
5: String json = accessPayload.JsontoString();
6: }

// The return value of entry method: accessPayload.JsontoString();
Expected: {namespace=ns1, app=app1, version=v1, type=Worker,
program=worker1, run=run1, entity=RUN} // On gson 2.2.4
Actual: {app=app1, version=v1, type=Worker, program=worker1,
run=run1, namespace=ns1,b entity=RUN} // On gson 2.0
```

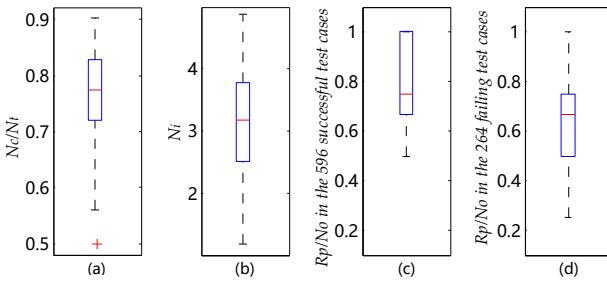
**Fig. 6:** A false positive example

which has the same attributes but different declaration orders on these two library versions. However, the attributes are stored in an unordered collection and the sequence of traversing the attributes is non-deterministic in the program. Such differences do not affect the semantics of client project, and therefore it did not catch developers' attention. The other false positives detected by both SENSOR and RIDDLE are similar cases, in which the inconsistent behaviors affected by conflicting API pairs (e.g., non-deterministic text formats, random values, etc.), are benign for the program semantics.

**False negative examples.** We manually investigated the 18 false negative cases detected by SENSOR, and divided them into the following two categories:

- *The inconsistent branches within conflicting API pairs cannot be reached* ( $10/18 = 55.6\%$ ). In these cases, the required object constructors have multiple caller methods with different invocation contexts in the source code, and the ones seeded by Sensor could not trigger the inconsistent branches within conflicting API pairs, even over the ten runs.
- *The program crashed before reaching the conflicting API pairs* ( $8/18 = 44.4\%$ ). The eight false negative cases were caused by the inaccurate arguments extracted by SENSOR, which led to program crashes before triggering the conflicting API pairs. By further investigation, we found that the inaccurate arguments are manifested into two patterns: (1) *the required arguments are affected by a series of method calls that cannot be extracted from the source code exactly* ( $5/8 = 62.5\%$ ), and (2) *part of a constructor's arguments cannot be found from the source code, to which random values are assigned* ( $3/8 = 37.5\%$ ).

2) *Effectiveness on producing valid class instances.*: Let  $N_c$  be the number of classes in a project for which SENSOR could construct instances;  $N_t$  be the total number of classes in this project; and  $N_i$  be the average number of instances with different divergence arguments constructed for each class in a project. Then,  $N_c/N_t$  describes SENSOR's capability on constructing class instances with extracted invocation contexts. In our ground truth dataset, the 225 isomeric conflicting API pairs are selected from 123 Java projects. As shown in Figures 7(a) and 7(b), the box plots show the distribution of indicators  $N_c/N_t$  and  $N_i$  in these projects. On average, SENSOR could construct instances with extracted arguments for 76.8% of classes in the projects. We looked into the code

**Fig. 7:** Effectiveness on producing valid class instances

and found that SENSOR could not instantiate the remaining classes mainly because their required arguments are not provided in the source code, or for the accessibility reason. On average, SENSOR constructed 3.13 instances for each class with divergence arguments, in these projects. Specially, in project FluentLenium-3.9.0, there are 89 classes having more than ten invocation contexts for their corresponding constructors. The diverse constructed class instances significantly increase the probability of capturing inconsistent behaviors with different invocation contexts.

Let  $R_p$  represent the number of class instances that are successfully seeded by SENSOR in a test case, and  $N_o$  be the total number of instantiated classes in this test case. Then,  $R_p/N_o$  is the substitution rate of class instances by our approach in a test case. Figures 7(c) and 7(d) show the values of  $R_p/N_o$  in the 596 generated test cases that successfully captured the inconsistent behaviors and the 264 test cases that caused crashes before triggering the conflicting APIs, over the ten runs, respectively. The average value of  $R_p/N_o$  in Figure 7(c) is 0.22 higher than that in Figure 7(d). The results demonstrated the validity of our seeded class instances. Furthermore, suppose that  $Argu_s$  is the number of arguments required by a constructor that can be extracted from the source code and  $Argu_t$  is the total number of required arguments in this constructor. For the seeded class instances in the 596 test cases that successfully identified SC issues, the average value of  $Argu_s/Argu_t$  is 0.24 higher than that of class instances in the 264 test cases causing program crashes. This validates the correctness of the arguments extracted by SENSOR for constructing class instances.

*3) Effectiveness on industrial projects.*: To further assess SENSOR's effectiveness, we applied it to the industrial projects in the Neusoft Co. Ltd (SSE: 600718) and received an assessment report [23]. Neusoft [11] is the largest IT solutions & services provider in China, which has considerable large-scale Java projects with hundreds of third party libraries. Diagnosing SC issues is one of the key challenges for their developers. Table 3 reports the results of applying SENSOR to ten industrial subjects that comprise over 0.58 million lines of code. We invited nine developers who participated in development of the selected projects to verify the detected SC issues. We did not evaluate the *Recall* in this experiment since it is difficult to obtain the complete set of SC issues in the projects.

In Table 3, columns “CA” and “ICA” represent the number of conflicting API pairs and isomeric conflicting API pairs caused by two conflicting library versions in the project,

**TABLE III:** The results on anonymized industry projects

Projects	SLOC	CA	ICA	TP	FP
P1	>100K	103	44	7	2
P2	>100K	78	9	3	0
P3	>100K	213	36	10	2
P4	>100K	52	20	5	2
P5	>50K	89	25	6	1
P6	>50K	31	7	2	1
P7	>50K	64	12	3	0
P8	>20K	45	11	3	1
P9	>20K	23	7	2	0
P10	>20K	42	8	3	0
<i>Precision</i>				46 / 56 = 0.821	

**TABLE IV:** The SC issues reported by SENSOR

Htm.java, #550, 1; EasyTransaction, #144, 1; Hydra, #364▲, 1; Motan, #800★, 15; Restx, #297★, 3; Netty-rest, #8★, 12; Netty-rest, #9★, 9; Aws-sdk-java, #1897◊, 1; Ff4j, #336★, 10; Retrofit, #3018★, 12; Guagua, #103★, 19; Wechat-springmvc, #17★, 3 Jss7, #309, 1; Motan, #809★, 15; Product-iot, #1911, 1; Nutzboot, #199★, 2; Atom-hopper, #301★, 1; Quick-media, #41★, 1; Ontop, #287★, 12; Ontop, #288★, 17; Odo, #173★, 1; Openstack-java-sdk, #214★, 1; Java-design-patterns, #868★, 1; Hmily, #86★, 1; Ninja, #654★, 1; JavaCPP, #295◊, 1; FastjsonExploit, #6★, 1; MiA, #11▲, 1; Vertx-examples, #335▲, 1; Vertx-examples, #336▲, 1; Yawp, #121★, 1; Apache/Hive, #21374★, 1; Rest-assured, #1143★, 1
--

( Project name, issue report ID, the number of SC issues in the issue report )

★: The issues have already been fixed. ▲: The issues were confirmed and in the process of being fixed. ◊: False positive cases.

The detailed information is available at: <https://sensordc.github.io/>

respectively. Among the 56 detected SC issues, 46 were confirmed by developers as true positives (TP) and 10 were labeled as false positives, leading to a *Precision* of 0.821. By communicating with the projects' developers, we found that the main cause of the above false positives (FP) is the same as that in the open-source projects. In these cases, the inconsistent variable states affected by the conflicting API pairs are benign for the semantics of industrial subjects. In particular, we received positive feedback from the Neusoft's testing team, on the high precision of SENSOR. Such results indicate that SENSOR not only achieves significant effectiveness on open-source projects, but also performs great on industrial subjects.

### C. RQ2: Usefulness of Sensor

SENSOR successfully detected 150 SC issues from 29 projects among all the 92 projects. Note that the SC issues caused by a pair of conflicting library versions were merged into one issue report. Altogether, we submitted 33 issue reports. As shown in Table 4, 27 out of 33 issue reports (81.8%) were confirmed by developers as real bugs; 23 out of 27 confirmed reports (85.2%) were quickly fixed; and 4 of them (14.8%) were in the process of being fixed. Among the seven unconfirmed issue reports, two were labeled as false positives and the others are not confirmed mainly due to inactive maintenance of the corresponding release versions.

We have received developers' positive feedbacks on the reported SC issues and our tool. Developers in Issue #11 [24] agreed that the provided test case indeed triggered realistic program behaviors, which has facilitated their diagnosis for semantic conflicts. In particular, a developer confirmed the usefulness of our approach :

*"I encountered the same problem when using MiA. I just noticed something strange happened in the [min, max] range of operation progresses. Thanks for your test case. It helped me reproduce this issue. By amazing coincidence, I got the similar outputs as your test."*

Besides, developers have expressed great interests in our detection technique for SC issues. For instance, in Issue #288 [25], an experienced developer [26] in the Ontop community has been looking for a technique of such kind to detect semantic conflicts:

*"I am very interested in the detection method so that in the future we will have a more systematic approach to avoid the issue related to such conflict, which is in general quite subtle and difficult to debug."*

SENSOR was highly recognized in the assessment report [23] provided by Neusoft:

*"On average, it took about 20.5 hours to obtain the diagnosis report for a large-scale Neusoft project, and the run time depends on the number of conflicting APIs in the project. Although the testing task is time-consuming, SENSOR did a great job in automatically detecting SC issues. The generated diagnosis reports indeed helped us identify many issues that could hardly be found using our existing test suites."*

The above results and developers' feedback demonstrate that the information (e.g., test cases) provided by SENSOR is useful for developers to diagnose the SC issues in practice.

#### D. Discussions

We further analyzed the root causes and distributions of the behavioral inconsistencies induced by the identified isomeric conflicting API pairs of all the subjects as shown in Table 4. Statistically, for the 150 isomeric conflicting API pairs that affect the client projects' program behaviors, we further categorized the exposed behavioral inconsistencies into three types: (a) 135 of them (90.0%) only cause *variable state inconsistencies*; (b) 5 of them (3.3%) only lead to *test outcome inconsistencies*; (c) 10 of them (6.7%) result in both *variable state* and *test outcome* inconsistencies.

By manually examining the source code of isomeric conflicting API pairs, we found that *variable state inconsistencies* are mainly caused by adding or deleting control branches in one version of the conflicting API (e.g., issue #550 [27]) or inconsistent function implementations between conflicting API pairs (e.g., issue #1143 [28]). In addition, strengthening the precondition or weakening postcondition of a referenced method will lead to *test outcome inconsistencies*. A method's precondition is the condition that a caller must satisfy before calling the method, and a method's postcondition is the condition that a callee must satisfy before returning from the method [2]. For instance, in issue #9 [29], replacing the shadowed version of a method with the loaded version resulted in strengthening the precondition of this conflicting method. Therefore, the caller in the host project will trigger a `NullPointerException` when the caller of method `setNamesSize()` in the client project passes an empty `HashSet` to it. Similarly, in issue #809 [30], referencing the actually loaded version of a method to substitute the shadowed version will weaken its postcondition, which can trigger a

crash in the caller. The above cases will break the compatibility of libraries in the client projects.

#### VI. THREATS TO VALIDITY

**Ground truth dataset collection.** Collecting the ground truth dataset of SC issues is challenging and can be a threat to the evaluation results. To avoid introducing noises in our dataset, we upgraded/ downgraded the actually-loaded library versions in a series of Java projects. After altering the library versions, we selected the conflicting API pairs that could trigger the `AssertionErrors` when executing the projects' associated tests, as the cases that definitely caused SC issues.

**Validity of developers' feedback.** In this paper, we rely on developers' feedback to validate the effectiveness and usefulness of SENSOR on both industrial and open source projects. However, there might be different opinions towards the validity of the issue reports for different developers. To mitigate such threat, for the industrial subjects, we invited nine original developers with the domain knowledge of the selected projects for verification. For the open source projects, we did not encounter the controversies for all the evaluated subjects. Therefore, the received feedback demonstrate the effectiveness and usefulness of our approach.

#### VII. RELATED WORK

**Dependency conflict.** Library conflicts are challenging to detect for a program analysis and difficult to avoid for library developers. Determining whether two or more libraries cannot be built together is an important issue in the quality assurance process of software projects. Blincoe et al. [31] conducted an in-depth study of millions of dependencies across multiple software ecosystems. They found that using a range of versions to declare dependencies could facilitate the automated repairing for dependency conflict issues, when adopting semantic versioning strategies. Yet, since the vast majority of Java projects declare the fixed versions of their referenced third party libraries, semantic versioning does not play a major role in repairing dependency conflict issues in the Java ecosystem.

Ghorbani et al. [32] formally defined eight inconsistent modular dependencies that may arise in Java-9 applications, and proposed a technique Darcy to detect and repair such specified inconsistent dependencies. So far, there are a significant fraction of Java projects that do not adopt Java-9 mechanism. Therefore, an effective approach to diagnosing the SC issues is still urgently needed in the Java ecosystem.

Suzaki et al.' approach [33] mainly focused on conflicts on resource access, conflicts on configuration data, and interactions between uncommon combinations of packages and categorized them to provide useful suggestions on how to prevent and detect such problem. Patra et al. [34] were the first researchers studying the detection strategy for conflicts among JavaScript libraries. They tackled the huge search space of possible conflicts in two phases, i.e., identifying potentially conflicting library pairs and synthesizing library clients to validate conflicts. Soto-Valero et al. [35] presented quantitative empirical evidence about how the immutability of artifacts in Maven Central supports the emergence of natural software diversity. Wang et al. [5] conducted a study

to characterize the manifestations of dependency conflicts in Java projects, and presented an automated technique to diagnose dependency conflict issues. Afterwards, they developed RIDDLE to generate tests to collect crashing stack traces to facilitate dependency conflict diagnosis [7]. However, there is no previous work analyzing the impacts on semantic behaviors of programs when dependency conflicts happen.

**Differential testing and analysis.** Differential testing and analysis techniques have been used to find bugs across many types of programs [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47]. Zhang et al. [48] implemented an isomorphic regression testing approach named ISON, which compares the behaviors of modified programs to check whether abnormal behaviors are induced in the new code versions. Xie et al. [49] presented a differential unit testing technique, DIFFUT, which compared the methods between different program revisions. Petsios et al. [50] proposed an effective technique NEZHA to trigger semantic bugs, using gray-box and black-box mechanisms to generate inputs for differential testing. NEZHA is applicable to detecting the semantic differences among different libraries providing similar functionalities. Compared with the above differential testing techniques, SENSOR has a different goal: detecting semantic conflicts combining host projects' invocation contexts.

**Test input generation.** Existing automated testing generation approaches use many techniques to create inputs for exercising a software under test with minimal human efforts, including feedback-directed random test generation [51], [52], [53], search-based techniques [54], [55], seeding strategies [56], [57], [58], [18], [59], [60], and symbolic reasoning-based test generators [44], [61], [62], [63]. Xu et al. [64] presented a mining approach to building a decision tree model according to the test inputs generated from Java bytecode. It converts Java bytecode into the Jimple representation, extracts predicates from the control flow graph of the Jimple code, and uses these predicates as attributes for organizing training data to build a decision tree. Dallmeier et al. [65] proposed an improved dynamic specification mining technique, TAUTOKO, to generate test cases. Since previous specification mining technique entirely depends on the observed executions, the resulting specification may be too incomplete to be useful if not enough tests are available. To address this problem, TAUTOKO explores previously unobserved aspects of the execution space. Their evaluation results show that the enriched specifications cover more general behaviors and much more exceptional behaviors. Toffola et al. [66] proposed an approach to extract literals from thousands of tests and to adapt information retrieval techniques to find values suitable for a particular domain.

Despite all successes, test generation still suffers from non-trivial limitations in exposing SC issues. First, approach [64] are more effective to deal with the code snippets with simple data types that can be easily convert into Jimple representations. While our empirical study results show that to trigger the real SC issues, effective test generation techniques should have the ability to construct divergence arguments for parameterized complex constructors. Second, the effectiveness of approaches [65], [66] entirely depends on the detection ability

of existing test suites of projects under test. Our empirical study provides evidences that combining invocation contexts of constructors in source code can effectively improve the possibility of capturing inconsistent program behaviors caused by SC issues. In addition, existing techniques [56], [57], [58], [18], [59], [60] have proposed different seeding strategies for test input generation, especially for strings and primitive types. Since these strategies generate constructor arguments without considering their invocation contexts, they are not effective in constructing valid class instances to expose conflicting API pairs. SENSOR adopts a new seeding strategy of class instances inspired by the our empirical findings summarized in Section III. SENSOR injects the invocation context information extracted from the source code into class instances with the aim of generating divergence arguments.

## VIII. CONCLUSION AND FUTURE WORK

In this paper, we presented an effective and automated test generation technique SENSOR, which are capable of producing valid inputs to trigger the SC issues. The evaluation results show that SENSOR can achieve a *Precision* of 0.803 and a *Recall* of 0.760 on open source projects and a *Precision* of 0.821 on industrial subjects. SENSOR has detected 150 SC issues from 29 open source projects and submitted 33 issue reports to them. Encouragingly, 27 issue reports (81.8%) have been confirmed by developers as real SC issues. Although SENSOR is designed for detecting SC issues, it can be adapted to other problems arising from library evolution (e.g., safeguarding the reliability upgrading libraries). In future, we plan to combine symbolic execution or fuzzing techniques with our technique to improve its test input exploration capability.

## REFERENCES

- [1] L. Bao, Z. Xing, X. Xia, D. Lo, and A. E. Hassan, "Inference of development activities from interaction with uninstrumented applications," *Empirical Software Engineering*, vol. 23, no. 3, pp. 1313–1351, 2018.
- [2] K. Jezek, J. Dietrich, and P. Brada, "How java apis break—an empirical study," *Information and Software Technology*, vol. 65, pp. 129–146, 2015.
- [3] C. Teyton, J.-R. Falleri, M. Palyart, and X. Blanc, "A study of library migrations in java," *Journal of Software: Evolution and Process*, vol. 26, no. 11, pp. 1030–1052, 2014.
- [4] C. Macho, S. McIntosh, and M. Pinzger, "Automatically repairing dependency-related build breakage," in *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2018, pp. 106–117.
- [5] Y. Wang, M. Wen, Z. Liu, R. Wu, R. Wang, B. Yang, H. Yu, Z. Zhu, and S.-C. Cheung, "Do the dependency conflicts in my project matter?" in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, 2018, pp. 319–330.
- [6] S. Liang and G. Bracha, "Dynamic class loading in the java [tm] virtual machine," *Acm sigplan notices*, vol. 33, pp. 36–44, 1998.
- [7] Y. Wang, M. Wen, R. Wu, Z. Liu, S. H. Tan, Z. Zhu, H. Yu, and S.-C. Cheung, "Could I Have a Stack Trace to Examine the Dependency Conflict Issue?" in *Proceedings of the 41th International Conference on Software Engineering*, ser. ICSE. ACM/IEEE, 2019.
- [8] J. Wang, G. Dong, J. Sun, X. Wang, and P. Zhang, "Adversarial sample detection for deep neural network through model mutation testing," *arXiv preprint arXiv:1812.05793*, 2018.
- [9] "Issue #214 of project openstack-java-sdk," <https://github.com/woorea/openstack-java-sdk/issues/214>, 2020, accessed: 2020-01-31.
- [10] "A pr of issue #214 in project openstack-java-sdk," <https://github.com/woorea/openstack-java-sdk/pull/215>, 2020, accessed: 2020-01-31.
- [11] "Neusoft," <https://www.neusoft.com/>, 2020, accessed: 2020-01-31.

- [12] "Rest-assured," <https://github.com/rest-assured/rest-assured>, 2020, accessed: 2020-01-31.
- [13] "Java-design-patterns," <https://github.com/iluwatar/java-design-patterns>, 2020, accessed: 2020-01-31.
- [14] "Hmily," <https://github.com/yu199195/hmily>, 2020, accessed: 2020-01-31.
- [15] S. Artzi, S. Kim, and M. D. Ernst, "Recrash: Making software failures reproducible by preserving object states," in *European conference on object-oriented programming*. Springer, 2008, pp. 542–565.
- [16] J.-R. Falleri, F. Morandat, X. Blanc, M. Martinez, and M. Monperrus, "Fine-grained and accurate source code differencing," in *Proceedings of the 29th ACM/IEEE international conference on Automated software engineering*. ACM, 2014, pp. 313–324.
- [17] A. Schroter, A. Schröter, N. Bettenburg, and R. Premraj, "Do stack traces help developers fix bugs?" in *2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010)*. IEEE, 2010, pp. 118–121.
- [18] A. Sakti, G. Pesant, and Y.-G. Guéhéneuc, "Instance generator and problem representation to improve object oriented code coverage," *IEEE Transactions on Software Engineering*, vol. 41, no. 3, pp. 294–313, 2015.
- [19] G. Fraser and A. Arcuri, "Evosuite at the sbst 2016 tool competition," in *Proceedings of the 9th International Workshop on Search-Based Software Testing*, 2016, pp. 33–36.
- [20] A. Arcuri, G. Fraser, and J. P. Galeotti, "Automated unit test generation for classes with environment dependencies," in *Proceedings of the 29th ACM/IEEE international conference on Automated software engineering*, 2014, pp. 79–90.
- [21] S. Kim and E. J. Whitehead Jr, "How long did it take to fix bugs?" in *Proceedings of the 2006 international workshop on Mining software repositories*. ACM, 2006, pp. 173–174.
- [22] R. Wu, H. Zhang, S. Kim, and S.-C. Cheung, "Relink: recovering links between bugs and changes," in *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*. ACM, 2011, pp. 15–25.
- [23] "Assessment report," <https://sensordc.github.io/>, 2020, accessed: 2020-01-31.
- [24] "Issue #11 of project mia," <https://github.com/tdunning/MiA/issues/11>, 2020, accessed: 2020-01-31.
- [25] "Issue #288 of project ontop," <https://github.com/ontop/ontop/issues/288>, 2020, accessed: 2020-01-31.
- [26] "An experienced developer of project ontop," <https://github.com/ghxiao>, 2020, accessed: 2020-01-31.
- [27] "Issue #550 of project htm.java," <https://github.com/numamenta/htm.java/issues/550>, 2020, accessed: 2020-01-31.
- [28] "Issue #1143 of project rest-assured," <https://github.com/rest-assured/rest-assured/issues/1143>, 2020, accessed: 2020-01-31.
- [29] "Issue #9 of project netty-rest," <https://github.com/buremba/netty-rest/issues/9>, 2020, accessed: 2020-01-31.
- [30] "Issue #809 of motan," <https://github.com/weibocom/motan/issues/809>, 2020, accessed: 2020-01-31.
- [31] J. Dietrich, D. Pearce, J. Stringer, A. Tahir, and K. Blincoe, "Dependency versioning in the wild," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 349–359.
- [32] N. Ghorbani, J. Garcia, and S. Malek, "Detection and repair of architectural inconsistencies in java," in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 2019, pp. 560–571.
- [33] C. Artho, K. Suzuki, R. Di Cosmo, R. Treinen, and S. Zacchiroli, "Why do software packages conflict?" in *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories*. IEEE Press, 2012, pp. 141–150.
- [34] J. Patra, P. N. Dixit, and M. Pradel, "Conflictjs: Finding and understanding conflicts between javascript libraries," in *Proceedings of the 40th International Conference on Software Engineering*. ACM, 2018, pp. 741–751.
- [35] C. Soto-Valero, A. Benelallam, N. Harrand, O. Barais, and B. Baudry, "The emergence of software diversity in maven central," in *MSR 2019-16th International Conference on Mining Software Repositories*. ACM, 2019, pp. 1–11.
- [36] T. Zhang and M. Kim, "Automated transplantation and differential testing for clones," in *Proceedings of the 39th International Conference on Software Engineering*. IEEE, 2017, pp. 665–676.
- [37] P. Chapman and D. Evans, "Automated black-box detection of side-channel vulnerabilities in web applications," in *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011, pp. 263–274.
- [38] Y. Chen, T. Su, C. Sun, Z. Su, and J. Zhao, "Coverage-directed differential testing of jvm implementations," in *ACM SIGPLAN Notices*, vol. 51, no. 6. ACM, 2016, pp. 85–99.
- [39] X. Yang, Y. Chen, E. Eide, and J. Regehr, "Finding and understanding bugs in c compilers," in *ACM SIGPLAN Notices*, vol. 46, no. 6. ACM, 2011, pp. 283–294.
- [40] S. A. Chowdhury, T. T. Johnson, and C. Csallner, "Cyfuzz: A differential testing framework for cyber-physical systems development environments," in *International Workshop on Design, Modeling, and Evaluation of Cyber Physical Systems*. Springer, 2016, pp. 46–60.
- [41] B. Daniel, D. Dig, K. Garcia, and D. Marinov, "Automated testing of refactoring engines," in *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*. ACM, 2007, pp. 185–194.
- [42] A. Groce, G. Holzmann, and R. Joshi, "Randomized differential testing as a prelude to formal verification," in *29th International Conference on Software Engineering (ICSE'07)*. IEEE, 2007, pp. 621–631.
- [43] R. Lämmel and W. Schulte, "Controllable combinatorial coverage in grammar-based testing," in *IFIP International Conference on Testing of Communicating Systems*. Springer, 2006, pp. 19–38.
- [44] C. Cadar, D. Dunbar, D. R. Engler *et al.*, "Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs." in *OSDI*, vol. 8, 2008, pp. 209–224.
- [45] B. Ray, M. Kim, S. Person, and N. Rungta, "Detecting and characterizing semantic inconsistencies in ported code," in *2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2013, pp. 367–377.
- [46] H. Zhong, S. Thummalapenta, and T. Xie, "Exposing behavioral differences in cross-language api mapping relations," in *International Conference on Fundamental Approaches to Software Engineering*. Springer, 2013, pp. 130–145.
- [47] Y. Lin, Z. Xing, Y. Xue, Y. Liu, X. Peng, J. Sun, and W. Zhao, "Detecting differences across multiple instances of code clones," in *Proceedings of the 36th International Conference on Software Engineering*, 2014, pp. 164–174.
- [48] J. Zhang, Y. Lou, L. Zhang, D. Hao, L. Zhang, and H. Mei, "Isomorphic regression testing: Executing uncovered branches without test augmentation," in *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 2016, pp. 883–894.
- [49] T. Xie, K. Taneja, S. Kale, and D. Marinov, "Towards a framework for differential unit testing of object-oriented programs," in *Second International Workshop on Automation of Software Test (AST'07)*. IEEE, 2007, pp. 5–5.
- [50] T. Petsios, A. Tang, S. Stolfo, A. D. Keromytis, and S. Jana, "Nezha: Efficient domain-independent differential testing," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 615–632.
- [51] S. Artzi, J. Dolby, S. H. Jensen, A. Möller, and F. Tip, "A framework for automated testing of javascript web applications," in *Proceedings of the 33rd International Conference on Software Engineering*, 2011, pp. 571–580.
- [52] C. Pacheco, S. K. Lahiri, M. D. Ernst, and T. Ball, "Feedback-directed random test generation," in *29th International Conference on Software Engineering (ICSE'07)*. IEEE, 2007, pp. 75–84.
- [53] M. Pradel and T. R. Gross, "Fully automatic and precise detection of thread safety violations," in *Proceedings of the 33rd ACM SIGPLAN conference on Programming Language Design and Implementation*, 2012, pp. 521–530.
- [54] W.-F. Chiang, G. Gopalakrishnan, Z. Rakamaric, and A. Solovyev, "Efficient search for inputs causing high floating-point errors," in *Proceedings of the 19th ACM SIGPLAN symposium on Principles and practice of parallel programming*, 2014, pp. 43–52.
- [55] G. Fraser and A. Arcuri, "Evosuite: automatic test suite generation for object-oriented software," in *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*. ACM, 2011, pp. 416–419.
- [56] N. Alshahwan and M. Harman, "Automated web application testing using search based software engineering," in *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*. IEEE Computer Society, 2011, pp. 3–12.
- [57] J. M. Rojas, G. Fraser, and A. Arcuri, "Seeding strategies in search-based unit test generation," *Software Testing, Verification and Reliability*, vol. 26, no. 5, pp. 366–401, 2016.
- [58] G. Fraser and A. Arcuri, "The seed is strong: Seeding strategies in search-based software testing," in *2012 IEEE Fifth International*

- Conference on Software Testing, Verification and Validation.* IEEE, 2012, pp. 121–130.
- [59] P. McMinn, M. Shahbaz, and M. Stevenson, “Search-based test input generation for string data types using the results of web queries,” in *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*. IEEE, 2012, pp. 141–150.
- [60] M. Alshraideh and L. Bottaci, “Search-based software test data generation for string data using program-specific search operators,” *Software Testing, Verification and Reliability*, vol. 16, no. 3, pp. 175–203, 2006.
- [61] P. Godefroid, N. Klarlund, and K. Sen, “Dart: directed automated random testing,” in *Proceedings of the 2005 ACM SIGPLAN conference on Programming language design and implementation*, 2005, pp. 213–223.
- [62] K. Sen, D. Marinov, and G. Agha, “Cute: a concolic unit testing engine for c,” *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 5, pp. 263–272, 2005.
- [63] S. Thummalapenta, T. Xie, N. Tillmann, J. De Halleux, and Z. Su, “Synthesizing method sequences for high-coverage testing,” *ACM SIGPLAN Notices*, vol. 46, no. 10, pp. 189–206, 2011.
- [64] W. Xu, T. Ding, H. Wang, and D. Xu, “Mining test oracles for test inputs generated from java bytecode,” in *2013 IEEE 37th Annual Computer Software and Applications Conference*. IEEE, 2013, pp. 27–32.
- [65] V. Dallmeier, N. Knopp, C. Mallon, S. Hack, and A. Zeller, “Generating test cases for specification mining,” in *Proceedings of the 19th international symposium on Software testing and analysis*, 2010, pp. 85–96.
- [66] L. Della Toffola, C.-A. Staicu, and M. Pradel, “Saying ‘hi’ is not enough: Mining inputs for effective test generation,” in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, pp. 44–49.