JSEFT: Automated JavaScript Unit Test Generation

Shabnam Mirshokraie  Ali Mesbah  Karthik Pattabiraman
University of British Columbia
Vancouver, BC, Canada
{shabnamm, amesbah, karthikp}@ece.ubc.ca

Abstract—The event-driven and highly dynamic nature of JavaScript, as well as its runtime interaction with the Document Object Model (DOM) make it challenging to test JavaScript-based applications. Current web test automation techniques target the generation of event sequences, but they ignore testing the JavaScript code at the unit level. Further they either ignore the oracle problem completely or simplify it through generic soft oracles such as HTML validation and runtime exceptions. We present a framework to automatically generate test cases for JavaScript applications at two complementary levels, namely events and individual JavaScript functions. Our approach employs a combination of function coverage maximization and function state abstraction algorithms to efficiently generate test cases. In addition, these test cases are strengthened by automatically generated mutation-based oracles. We empirically evaluate the implementation of our approach, called JSEFT, to assess its efficacy. The results, on 13 JavaScript-based applications, show that the generated test cases achieve a coverage of 68% and that JSEFT can detect injected JavaScript and DOM faults with a high accuracy (100% precision, 70% recall). We also find that JSEFT outperforms an existing JavaScript test automation framework both in terms of coverage and detected faults.

Keywords—Test generation; oracles; JavaScript; DOM

I. INTRODUCTION

JavaScript plays a prominent role in modern web applications. To test their JavaScript applications, developers often write test cases using web testing frameworks such as SELENIUM (GUI tests) and QUNIT (JavaScript unit tests). Although such frameworks help to automate test execution, the test cases still need to be written manually, which is tedious and time-consuming.

Further, the event-driven and highly dynamic nature of JavaScript, as well as its runtime interaction with the Document Object Model (DOM) make JavaScript applications error-prone [1] and difficult to test.

Researchers have recently developed automated test generation techniques for JavaScript-based applications [2], [3], [4], [5], [6]. However, current web test generation techniques suffer from two main shortcomings, namely, they:

1) Target the generation of event sequences, which operate at the event-level or DOM-level to cover the state space of the application. These techniques fail to capture faults that do not propagate to an observable DOM state. As such, they potentially miss this portion of code-level JavaScript faults. In order to capture such faults, effective test generation techniques need to target the code at the JavaScript unit-level, in addition to the event-level.

2) Either ignore the oracle problem altogether or simplify it through generic soft oracles, such as W3C HTML validation [2], [5], or JavaScript runtime exceptions [2]. A generated test case without assertions is not useful since coverage alone is not the goal of software testing. For such generated test cases, the tester still needs to manually write many assertions, which is time and effort intensive. On the other hand, soft oracles target generic fault types and are limited in their fault finding capabilities. However, to be practically useful, unit testing requires strong oracles to determine whether the application under test executes correctly.

To address these two shortcomings, we propose an automated test case generation technique for JavaScript applications.

Our approach, called JSEFT (JavaScript Event and Function Testing) operates through a three step process. First, it dynamically explores the event-space of the application using a function coverage maximization method, to infer a test model. Then, it generates test cases at two complementary levels, namely, DOM event and JavaScript functions. Our technique employs a novel function state abstraction algorithm to minimize the number of function-level states needed for test generation. Finally, it automatically generates test oracles, through a mutation-based algorithm.

A preliminary version of this work appeared in a short New Ideas paper [7]. In this current paper, we present the complete technique with conceptually significant improvements, including detailed new algorithms (Algorithms 1–2), a fully-functional tool implementation, and a thorough empirical analysis on 13 JavaScript applications, providing evidence of the efficacy of the approach.

This work makes the following main contributions:

- An automatic technique to generate test cases for JavaScript functions and events.
- A combination of function coverage maximization and function state abstraction algorithms to efficiently generate unit test cases;
- A mutation-based algorithm to effectively generate test oracles, capable of detecting regression JavaScript and DOM-level faults;
- The implementation of our technique in a tool called JSEFT, which is publicly available [8];
- An empirical evaluation to assess the efficacy of JSEFT using 13 JavaScript applications.

The results of our evaluation show that on average (1) the generated test suite by JSEFT achieves a 68% JavaScript code coverage, (2) compared to ARTEMIS, a feedback-directed JavaScript testing framework [2], JSEFT achieves 53% better coverage, and (3) the test oracles generated by JSEFT are able to detect injected faults with 100% precision and 70% recall.

II. RELATED WORK

Web application testing. Marchetto and Tonella [3] propose a search-based algorithm for generating event-based sequences
to test Ajax applications. Mesbah et al. [9] apply dynamic analysis to construct a model of the application’s state space, from which event-based test cases are automatically generated. In subsequent work [5], they propose generic and application-specific invariants as a form of automated soft oracles for testing Ajax applications. Our earlier work, JSART [10], automatically infers program invariants from JavaScript execution traces and uses them as regression assertions in the code. Sen et al. [11] recently proposed a record and replay framework called Jalangi. It incorporates selective record-replay as well as shadow values and shadow execution to enable writing of heavy-weight dynamic analyses. The framework is able to track generic faults such as null and undefined values as well as type inconsistencies in JavaScript. Jensen et al. [12] propose a technique to test the correctness of communication patterns between client and server in Ajax applications by incorporating server interface descriptions. They construct server interface descriptions through an inference technique that can learn communication patterns from sample data. Saxena et al. [6] combine random test generation with the use of symbolic execution for systematically exploring a JavaScript application’s event space as well as its value space, for security testing. Our work is different in two main aspects from these: (1) they all target the generation of event sequences at the DOM level, while we also generate unit tests at the JavaScript code level, which enables us to cover more and find more faults, and (2) they do not address the problem of test oracle generation and only check against soft oracles (e.g., invalid HTML). In contrast, we generate strong oracles that capture application behaviours, and can detect a much wider range of faults.

Perhaps the most closely related work to ours is ARTEMIS [2], which supports automated testing of JavaScript applications. ARTEMIS considers the event-driven execution model of a JavaScript application for feedback-directed testing. In this paper, we quantitatively compare our approach with that of ARTEMIS (Section V).

Oracle generation. There has been limited work on oracle generation for testing. Fraser et al. [13] propose \( \mu \)TEST, which employs a mutant-based oracle generation technique. It automatically generates unit tests for Java object-oriented classes by using a genetic algorithm to target mutations with high impact on the application’s behaviour. They further identify [14] relevant pre-conditions on the test inputs and post-conditions on the outputs to ease human comprehension. Differential test case generation approaches [15], [16] are similar to mutation-based techniques in that they aim to generate test cases that show the difference between two versions of a program. However, mutation-based techniques such as ours, do not require two different versions of the application. Rather, the generated differences are in the form of controllable mutations that can be used to generate test cases capable of detecting regression faults in future versions of the program. Staats et al. [17] address the problem of selecting oracle data, which is formed as a subset of internal state variables as well as outputs for which the expected values are determined. They apply mutation testing to produce oracles and rank the inferred oracles in terms of their fault finding capability. This work is different from ours in that they merely focus on supporting the creation of test oracles by the programmer, rather than fully automating the process of test case generation. Further, (1) they do not target JavaScript; (2) in addition to the code-level mutation analysis, we propose DOM-related mutations to capture error-prone \( \mu \) dynamic interactions of JavaScript

III. Challenges and Motivation

In this section, we illustrate some of the challenges associated with test generation for JavaScript applications.

Figure 1 presents a snippet of a JavaScript game application that we use as a running example throughout the paper. This simple example uses the popular jQuery library [18] and contains four main JavaScript functions:

1) `cellClicked()` is bound to the event-handlers of DOM elements with IDs `cello` and `cell1` (Lines 34–35). These two DOM elements become available when the DOM is fully loaded (Line 32). Depending on the element clicked, `cellClicked()` inserts a `div` element with ID `divElem` (Line 3) after the clicked element and makes it clickable by attaching either `setup()` or `setDim()` as its event-handler function (Lines 5–6, 9–10).

2) `setDim()` calls `setDim(dimension)` (Line 15) to change the value of the global variable `currentDim`. It further makes an element with ID `startCell` clickable by setting its event-handler to `startCell` (Line 16).

3) `setDim()` receives an input variable. It performs some computations to set the `height` value of the `divElem` property of a DOM element with ID `endCell` and the value of `currentDim` (Lines 20–22). It also returns the computed dimension.

4) `start()` is called at runtime when the element with ID `startCell` is clicked (Line 16), which either updates the width dimension of the element on which it was called, or removes the element (Lines 27–29).

There are four main challenges in testing JavaScript applications.
The first challenge is that a fault may not immediately propagate into a DOM-level observable failure. For example, if the ‘+’ sign in Line 21 is mistakenly replaced by ‘−’, the affected result does not immediately propagate to the observable DOM state after the function exits. While this mistakenly changes the value of a global variable, currentDim, which is later used in start (Line 27), it neither affects the returned value of the setDim function nor the css value of element endCell. Therefore, a GUI-level event-based testing approach may not help to detect the fault in this case.

The second challenge is related to fault localization: even if the fault propagates to a future DOM state and a DOM-level test case detects it, finding the actual location of the fault is challenging for the tester as the DOM-level test case is agnostic of the JavaScript code. However, a unit test case that targets individual functions, e.g., setDim in this running example, helps a tester to spot the fault, and thus easily resolve it.

The third challenge pertains to the event-driven dynamic nature of JavaScript, and its extensive interaction with the DOM resulting in many state permutations and execution paths. In the initial state of the example, clicking on cell0 or cell1 takes the browser to two different states as a result of the if-else statement in Lines 4 and 8 of the function cell1Clicked. Even in this simple example, expanding either of the resulting states has different consequences due to different functions that can be potentially triggered. Executing either of the resulting states has different consequences due to different functions that can be potentially triggered. Executing either setup or setDim in Lines 6 and 10 results in different execution paths, DOM states, and code coverage. It is this dynamic interaction of the JavaScript code with the DOM (and indirectly CSS) at runtime that makes it challenging to generate test cases for JavaScript applications.

The fourth important challenge in unit testing JavaScript functions that have DOM interactions, such as setDim, is that the DOM tree in the state expected by the function, has to be present during unit test execution. Otherwise the test will fail due to a null or undefined exception. This situation arises often in modern web applications that have many DOM interactions.

IV. APPROACH

Our main goal in this work is to generate client-side test cases coupled with effective test oracles, capable of detecting regression JavaScript and DOM-level faults. Further, we aim to achieve this goal as efficiently as possible. Hence, we make two design decisions. First, we assume that there is a finite amount of time available to generate test cases. Consequently we guide the test generation to maximize coverage under a given time constraint. The second decision is to minimize the number of test cases and oracles generated to only include those that are essential in detecting potential faults. Consequently, to examine the correctness of the test suite generated, the tester would only need to examine a small set of assertions, which minimizes their effort.

Our approach generates test cases and oracles at two complementary levels:

DOM-level event-based tests consist of DOM-level event sequences and assertions to check the application’s behaviour from an end-user’s perspective.

Function-level unit tests consist of unit tests with assertions that verify the functionality of JavaScript code at the function level.

![Fig. 2. Overview of our test generation approach.](image)

An overview of the technique is depicted in Figure 2. At a high level, our approach is composed of three main steps:

1) In the first step (Section IV-A), we dynamically explore various states of a given web application, in such a way as to maximize the number of functions that are covered throughout the program execution. The output of this initial step is a state-flow graph (SFG) [5], capturing the explored dynamic DOM states and event-based transitions between them.

2) In the second step (Section IV-B), we use the inferred SFG to generate event-based test cases. We run the generated tests against an instrumented version of the application. From the execution trace obtained, we extract DOM element states as well as JavaScript function states at the entry and exit points, from which we generate function-level unit tests. To reduce the number of generated test cases to only those that are constructive, we devise a state abstraction algorithm that minimizes the number of states by selecting representative function states.

3) To create effective test oracles for the two test case levels, we automatically generate mutated versions of the application (Section IV-C). Assuming that the original version of the application is fault-free, the test oracles are then generated at the DOM and JavaScript code levels by comparing the states traced from the original and the mutated versions.

A. Maximizing Function Coverage

In this step, our goal is to maximize the number of functions that can be covered, while exercising the program’s event space. To that end, our approach combines static and dynamic analysis to decide which state and event(s) should be selected for expansion to maximize the probability of covering uncovered JavaScript functions. While exploring the web application under test, our function coverage maximization algorithm selects a next state for exploration, which has the maximum value of the sum of the following two metrics:
1. Potential Uncovered Functions. This pertains to the total number of unexecuted functions that can potentially be visited through the execution of DOM events in a given DOM state $s_i$. When a given function $f_i$ is set as the event-handler of a DOM element $d \in s_i$, it makes the element a potential clickable element in $s_i$. This can be achieved through various patterns in web applications depending on which DOM event model level is adopted. To calculate this metric, our algorithm identifies all JavaScript functions that are directly or indirectly attached to DOM elements as event handlers, in $s_i$ through code instrumentation and execution trace monitoring.

2. Potential Clickable Elements. The second metric, used to select a state for expansion, pertains to the number of DOM elements that can potentially become clickable elements. If the event-handlers bound to those clickables are triggered, new (uncovered) functions will be executed. To obtain this number, we statically analyze the previously obtained potential uncovered functions within a given state in search of such elements.

While exploring the application, the next state for expansion is selected by adding the two metrics and choosing the state with the highest sum. The procedure repeats the aforementioned steps until the designated time limit, or state space size is reached.

In the running example of Figure 1, in the initial state, clicking on elements with IDs cell0 and cell1 results in two different states due to an if-else statement in Lines 4 and 8 of cellClicked. Let’s call the state in which a DIV element is located after the element with ID cell0 as $s_0$, and the state in which a DIV element is placed after the element with ID cell1 as $s_1$. If state $s_0$, with the clickable cell10, is chosen for expansion, function setup is called. As shown in Line 15, setup calls setDim, and thus, by expanding $s_0$ both of the aforementioned functions get called by a single click. Moreover, a potential clickable element is also created in Line 16, with start as the event-handler. Therefore, expanding $s_1$ results only in the execution of setDim, while expanding $s_0$ results in the execution of functions setup, setDim, and a potential execution of start in future states. At the end of this step, we obtain a state-flow graph of the application that can be used in the next test generation step.

B. Generating Test Cases

In the second step, our technique first extracts sequences of events from the inferred state-flow graph. These sequences of events are used in our test case generation process. We generate test cases at two complementary levels, as described below.

DOM-level event-based testing. To verify the behavior of the application at the user interface level, each event path, taken from the initial state (Index) to a leaf node in the state-flow graph, is used to generate DOM event-based test cases. Each extracted path is converted into a JUNIT SELENIUM-based test case, which executes the sequence of events, starting from the initial DOM state. Going back to our running example, one possible event sequence to generate is: `$('cell0').click -> $('div #divElem').click -> $('startCell').click`.

To collect the required trace data, we capture all DOM elements and their attributes after each event in the test path is fired. This trace is later used in our DOM oracle comparison, as explained in Section IV-C.

JavaScript function-level unit testing. To generate unit tests that target JavaScript functions directly (as opposed to event-triggered function executions), we log the state of each function at their entry and exit point, during execution. To that end, we instrument the code to trace various entities. At the entry point of a given JavaScript function we collect (1) function parameters including passed variables, objects, functions, and DOM elements, (2) global variables used in the function, and (3) the current DOM structure just before the function is executed. At the exit point of the JavaScript function and before every return statement, we log the state of the (1) return value of the function, (2) global variables that have been accessed in that function, and (3) DOM elements accessed (read/written) in the function. At each of the above points, our instrumentation records the name, runtime type, and actual values. The dynamic type is stored because JavaScript is a dynamically typed language, meaning that the variable types cannot be determined statically. Note that complex JavaScript objects can contain circular or multiple references (e.g., in JSON format). To handle such cases, we perform a de-serialization process in which we replace such references by an object in the form of $ref: Path$, where Path denotes a JSONPath string\(^1\) that indicates the target path of the reference.

In addition to function entry and exit points, we log information required for calling the function from the generated test cases. JavaScript functions that are accessible in the public scope are mainly defined in (1) the global scope directly (e.g., function $f()$), (2) variable assignments in the global scope (e.g., var $f = function()$), (3) constructor functions (e.g., function constructor() {this.member= function(){}}, and (4) prototypes (e.g., Constructor.prototype.$f = function(){}). Functions in the first and second case are easy to call from test cases. For the third case, the constructor function is called via the new operator to create an object type, which can be used to access the object’s properties (e.g., container=new Constructor(); container.member()); This allows us to access the inner function, which is a member of the constructor function in the above example. For the prototype case, the function can be invoked through container.$f() from a test case.

Going back to our running example in Figure 1, at the entry point of setDim, we log the value and type of both the input parameter dimension and global variable currentDim, which is accessed in the function. Similarly, at the exit point, we log the values and types of the returned variable dim and currentDim.

In addition to the values logged above, we need to capture the DOM state for functions that interact with the DOM. This is to address the fourth challenge outlined in Section III. To mitigate this problem, we capture the state of the DOM just before the function starts its execution, and include that as a test fixture [19] in the generated unit test case.

In the running example, at the entry point of setDim, we log the innerHTML of the current DOM as the function contains several calls to the DOM, e.g., retrieving the element with ID endCell in Line 22. We further include in our execution trace the way DOM elements and their attributes are modified by the JavaScript function at runtime. The information that we

\(^1\) http://goessner.net/articles/JsonPath/
Algorithm 1: Function State Abstraction

\[
\begin{align*}
\text{Algorithm 1: Function State Abstraction} \\
\text{input}: & \text{ The set of function states } st_i \in ST_i \text{ for a given function } f \\
\text{output}: & \text{ The obtained abstracted states set } \text{AbsStates} \\
\text{begin} & \\
\text{for } st_i \in ST_f \text{ do} & \\
L = 1; & \text{StSet}_L \leftarrow \emptyset \\
\text{if } \text{BrnCovLNS}[st_i] \neq \text{BrnCovLNS}[\text{StSet}_L^{\text{dim}}] \text{ then} & \\
\text{StSet}_{L+1} \leftarrow st_i & \\
L + 1 & \\
\text{else} & \\
\text{StSet}_L \leftarrow \text{StSet}_L \cup \text{StSet}_i & \\
K = L + 1; & \text{StSet}_K \leftarrow \emptyset \\
\text{if } \text{DOMPROPS}[st_i] \neq \text{DOMPROPS}[\text{StSet}_K^{\text{dim}}] \text{ then} & \\
\text{StSet}_{K+1} \leftarrow st_i & \\
K + 1 & \\
\text{else} & \\
\text{StSet}_k \leftarrow \text{StSet}_k \cup \text{StSet}_i & \\
\text{while } \text{StSet}_{K+L} \neq \emptyset \text{ do} & \\
\text{SelectedSt} \leftarrow \text{SELECTMAXST}(st_i; \text{StSet}_{j=1}^{K+L}) & \\
\text{AbsStates.ADD}[\text{SelectedSt}] & \\
\text{StSet}_{k+L} \leftarrow \text{StSet}_{K+L} \setminus \text{SelectedSt} & \\
\text{return } \text{AbsStates} & \\
\end{align*}
\]

log for accessed DOM elements includes the ID attribute, the XPath position of the element on the DOM tree, and all the modified attributes. Collecting this information is essential for oracle generation in the next step. We use a set to keep the information about DOM modifications, so that we can record the latest changes to a DOM element without any duplication within the function. For instance, we record ID as well as both the XPath position of the element on the DOM tree, and all the properties accessed in a function can be seen as entry states. Changes in such DOM entry states can affect the behaviour of the function. For example, in line 29 this keyword refers to the clicked DOM element of which function start is an event-handler. Assuming that currentDim ≤ 40, depending on which DOM element is clicked, by removing the element in line 29 the resulting state of the function start differs. Therefore, we take into consideration the DOM elements accessed by the function as well as the type of accessed DOM properties.

Algorithm 1 shows our function state abstraction algorithm. The algorithm first collects covered branches of individual functions per entry state (BrnCovLNS[st_i] in Line 3). Each function’s states exhibiting same covered branches are categorized under the same set of states (Lines 4 and 7). StSet_l corresponds to the set of function states, which are classified according to their covered branches, where l = 1,...,L and L is the number of current classified sets in that category. Similarly, function states with the same accessed DOM characteristics as well as return value type, are put into the same set of states (Lines 10 and 13). StSet_k corresponds to the set of function states, which are classified according to their DOM/return value type, where k = 1,...,K and K is the number of current classified sets in that category. After classifying each function’s states into several sets, we cover each set by selecting one of its common states. The state selection step is a set cover problem [20], i.e., given a universe U and a family S of subsets of U, a cover is a subfamily \( C \subseteq S \) of sets whose union is U. Sets to be covered in our algorithm are StSet_k+L, where st_i ∈ StSet_k+L. We use a common greedy algorithm for obtaining the minimum number of states that can cover all the possible sets (Lines 15-17). Finally, the abstracted list of states is returned in Line 18.

C. Generating Test Oracles

In the third step, our approach automatically generates test oracles for the two levels of test cases generated in the previous step, as depicted in the third step of Figure 2. Instead of randomly generating assertions, our oracle generation uses a mutation-based process.

Mutation testing is typically used to evaluate the quality of a test suite [21], or to generate test cases that kill mutants [13]. In our approach, we adopt mutation testing to (1) reduce the number of assertions automatically generated, (2) target critical and error-prone portions of the application. Hence, the tester would only need to examine a small set of effective assertions to verify the correctness of the generated oracles. Algorithm 2 shows our algorithm for generating test oracles. At a high level, the technique iteratively executes the following steps:
Algorithm 2: Oracle Generation

\[
\text{input} : \text{A web application (App), list of event sequences obtained from SFG (EvSeq), maximum number of mutations (n)} \\
\text{output: Assertions for function-level (FcAsserts) and DOM event-level tests (DomAsserts)} \\
\]

1. \text{App = INSTRUMENT(App)}
2. \text{while GenMuts < n do}
3. \hspace{1em} \text{foreach EvSeq in SFG do}
4. \hspace{2em} OnEvDomSt = Trace.GETONEVDOMST(Ev ∈ EvSeq)
5. \hspace{2em} AfterEvDomSt = Trace.GETAFTEREVDOMST(Ev ∈ EvSeq)
6. \hspace{2em} AccDomProps = GETACCDDOMNDS(OnEvDomSt)
7. \hspace{2em} EquivalentDomMut = true
8. \hspace{2em} while EquivalentCodeMut do
9. \hspace{3em} MutDom = MUTATEDOM(AccDomProps, OnEvDomSt)
10. \hspace{3em} ChangedSt = EvSeq.EXECEVENT(MutDom)
11. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
12. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
13. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
14. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
15. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
16. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
17. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
18. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
19. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
20. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
21. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
22. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
23. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
24. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
25. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
26. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
27. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
28. \hspace{3em} Diff(ChangedSt, AfterEvDomSt) = false
29. \text{return \{FcAsserts, DomAsserts\}}

1. A mutant is created by injecting a single fault into the original version of the web application (Line 9 and 19 in Algorithm 2 for DOM mutation and code-level mutation, respectively).
2. Related entry/exit program states at the DOM and JavaScript function levels of the mutant and the original version are captured. \text{OnEvDomSt} in Line 4 is the original DOM state on which the event \text{Ev} is triggered. \text{AfterEvDomSt} in line 5 is the observed DOM state after the event is triggered, \text{MutDom} in line 9 is the mutated DOM, and \text{ChangedSt} in line 10 is the corresponding affected state for DOM mutations. \text{FcExit} in Line 22 is the exit state of the function in the original application and \text{MutFcExit} in line 23 is the corresponding exit state for that function after the application is mutated for function-level mutations.
3. Relevant observed state differences at each level are detected and abstracted into test oracles (Diff in Line 11 and 24 for DOM and function-level oracles, respectively).
4. The generated assertions (Lines 15 and 28) are injected into the corresponding test cases.

DOM-level event-based test oracles. After an event is triggered in the generated Selenium test case, the resulting DOM state needs to be compared against the expected structure. One naïve approach would be to compare the DOM tree in its entirety, after the event execution. Not only is this approach inefficient, it results in brittle test-cases, i.e., the smallest update on the user interface can break the test case. We propose an alternative approach that utilizes DOM mutation testing to detect and selectively compare only those DOM elements and attributes that are affected by an injected fault at the DOM-level of the application. Our DOM mutations target only the elements that have been accessed (read/written) during execution, and thus have a larger impact on the application’s behaviour. To select proper DOM elements for mutation, we instrument JavaScript functions that interact with the DOM, i.e., code that either accesses or modifies DOM elements.

We execute the instrumented application by running the generated Selenium test cases and record each accessed DOM element, its attributes, the triggered event on the DOM state, and the DOM state after the event is triggered (\text{GETONEVDOMST} in line 4, \text{GETAFTEREVDOMST} in line 5, and \text{GETACCDDOMNDS} in line 6 to retrieve the original DOM state, DOM state after event \text{Ev} is triggered, and the accessed DOM properties as event \text{Ev} is triggered, respectively; in Algorithm 2). To perform the actual mutation, as the application is re-executed using the same sequence of events, we mutate the recorded DOM elements, one at a time, before the corresponding event is fired. \text{MUTATEDOM} in line 9 mutated the DOM elements, and \text{EvSeq.EXECEVENT} in line 10 executes the event sequence on the mutated DOM. The mutation operators include (1) deleting a DOM element, and (2) changing the attribute, accessed during the original execution. As we mutate the DOM, we collect the current state of DOM elements and attributes.

Figure 3 shows part of a DOM-level test case generated for the running example. Going back to our running example, as a result of clicking on \$\langle\text{div} \#\text{divElem}\rangle$ in our previously obtained event sequence \$\langle\text{#cell0}.\text{click}ightarrow\text{#cell1}.\text{click}\rightarrow\text{#startCell}\rangle\$, the height and width properties of DOM element with ID \text{endCell}, and the DOM element with ID \text{startCell} are accessed. One possible DOM mutation is altering the width value of the \text{endCell} element before click on \$\langle\text{div} \#\text{divElem}\rangle$ happens. We log the consequences of this modification after the click event on \$\langle\text{div} \#\text{divElem}\rangle$ as well as the remaining events. This mutation affects the height property of DOM element with ID \text{endCell} in the resulting DOM state from clicking on \$\langle\text{div} \#\text{divElem}\rangle$. Line 6 in Figure 3 shows the corresponding assertion. Furthermore, Assuming that the DOM mutation makes \text{currentDim} ≤ 40 in line 27, after click on element \text{#startCell} happens, the element is removed and no longer exists in the resulting DOM state. The generated assertion is shown in line 10 of Figure 3.

Hence, we obtain two sets of execution traces that contain information about the state of DOM elements for each fired event in the original and mutated application. By comparing these two traces (\text{DIFF} in line 11 in Algorithm 2), we identify all changed DOM elements and generate assertions for these elements. Note that any changes detected by the \text{DIFF} operator (line 12 in Algorithm 2) is an indication that the corresponding DOM mutation is not equivalent (line 13); if no change is detected, another DOM mutation is generated. We automatically place the generated assertion immediately after the corresponding line of code that executed
1 assertEquals(int 10)
2 assertEquals(startCellHeight, 50);
3 assertEquals((currentDim, 50);
4 assertEquals(startCell1Height, 30);
5 assertEquals(startCell2Height, 30);
6 assertEquals(result, 30);
7 assertEquals(result, 30);
8 assertEquals(startCell1Height, 50);
9 assertEquals(startCell2Height, 50);
10 assertEquals(true);
11 assertEquals(true);
12 assertEquals(true);

Fig. 3. Generated SELENIUM test case.

the event, in the generated event-based (SELENIUM) test case. 
DomAssertsEV.AfterEvDOMSt in line 15 contains all DOM assertions for the state AfterEvDOMSt and the triggered event Ev.

Function-level test oracles. To seed code level faults, we use our recently developed JavaScript mutation testing tool, MUTANDIS [22]. Mutations generated by MUTANDIS are selected through a function rank metric, which ranks functions in terms of their relative importance from the application’s behaviour point of view. The mutation operators are chosen from a list of common operators, such as changing the value of a variable or modifying a conditional statement. Once a mutant is produced (MUTATEJSCode in line 19), it is automatically instrumented. We collect a new execution trace from the mutated program by executing the same sequence of events that was used on the original version of the application. This way, the state of each JavaScript function is extracted at its entry and exit points. AbsFcSts.GETFCENTRIES in line 21 retrieves the function’s entries from the abstracted function’s states. GETFCEXIT in line 22, and GETMUTFCEXIT in line 23 retrieve the corresponding function’s exit state in the original and mutated application. This process is similar to the function state extraction algorithm explained in Section IV-B.

After the execution traces are collected for all the generated mutants, we generate function-level test oracles by comparing the execution trace of the original application with the traces we obtained from the modified versions (DIFF in line 24 in Algorithm 2). If the DIFF operator detects no changes (line 25 of the algorithm), an equivalent mutant is detected, and thus another mutant will be generated.

Our function-level oracle generation targets postcondition assertions. Such postcondition assertions can be used to examine the expected behaviour of a given function after it is executed in a unit test case. Our technique generates postcondition assertions for all functions that exhibit a different exit-point state but the same entry-point state, in the mutated execution traces. FcAssertEv in line 27 contains all such post condition assertions. Due to the dynamic and asynchronous behaviour of JavaScript applications, a function with the same entry state can exhibit different outputs when called multiple times. In this case, we need to combine assertions to make sure that the generated test cases do not mistakenly fail. FcAssertEvEntry in line 28 contains the union of function assertions generated for the same entry but different outputs during multiple executions. Let’s consider a function f with an entry state entry in the original version of the application (A), with two different exit states exit1 and exit2. If in the mutated version of the application (Am), f exhibits an exit state exit that is different from both exit1 and exit2, then we combine the resulting assertions as follows: assert1(exit1,expRes1)‖assert2(exit2,expRes2), where the expected values expRes1 and expRes2 are obtained from the execution trace of A.

D. Tool Implementation

We have implemented our JavaScript test and oracle generation approach in an automated tool called JSEFT. The tool is written in Java and is publicly available for download [8]. Our implementation requires no browser modifications, and is hence portable. For JavaScript code interception, we use a web proxy, which enables us to automatically instrument JavaScript code before it reaches the browser. The crawler for JSEFT extends and builds on top of the event-based crawler, CRAWLJAX [9], with random input generation enabled for form inputs. As mentioned before, to mutate JavaScript code, we use our recently developed mutation testing tool, MUTANDIS [22]. The upper-bound for the number of mutations can be specified by the user. However, the default is 50 for code-level and 20 for DOM-level mutations. We observed that these default numbers provide a balanced trade-off between oracle generation time, and the fault finding capability of the tool. DOM-level test cases are generated in a JUNIT format that uses SELENIUM (WebDriver) APIs to fire events on the application’s DOM inside the browser. JavaScript function-level tests are generated in the QUNIT unit testing framework [19], capable of testing any generic JavaScript code.

V. EMPIRICAL EVALUATION

To quantitatively assess the efficacy of our test generation approach, we have conducted an empirical study, in which we address the following research questions:

Fig. 4. Generated QUNIT test case.

of the application (Am), f exhibits an exit state exit that is different from both exit1 and exit2, then we combine the resulting assertions as follows: assert1(exit1,expRes1)‖assert2(exit2,expRes2), where the expected values expRes1 and expRes2 are obtained from the execution trace of A.

Each assertion for a function contains (1) the function’s returned value, (2) the used global variables in that function, and/or (3) the accessed DOM element in that function. Each assertion is coupled with the expected value obtained from the execution trace of the original version.

The generated assertions that target variables, compare the value as well as the runtime type against the expected ones. An oracle that targets a DOM element, first checks the existence of that DOM element. If the element exists, it checks the attributes of the element by comparing them against the observed values in the original execution trace. Assuming that width and height are 100 and 200 accordingly in Figure 1, and ‘+’ sign is mutated to ‘-’ in line 20 of the running example in Figure 1, the mutation affects the global variable currentDim, height property of element with ID endCell, and the returned value of the function setDim. Figure 4 shows a QUNIT test case for setDim function according to this mutation with the generated assertions.
RQ1 How effective is JSEFT in generating test cases with high coverage?

RQ2 How capable is JSEFT of generating test oracles that detect regression faults?

RQ3 How does JSEFT compare to existing JavaScript testing frameworks?

JSEFT and all our experimental data in this paper are available for download [8].

### A. Objects

Our study includes thirteen JavaScript-based applications in total. Table I presents each application's ID, name, lines of custom JavaScript code (LOC, excluding JavaScript libraries) and resource. The first five are web-based games, AjaxTabs is a JQUERY plugin for creating tabs, NarrowDesign and JointLondon are websites. FractalViewer is a fractal tree zoom application. SimpleCart is a shopping cart library, WymEditor is a web-based HTML editor, TuduList is a web-based task application. SimpleCart is a shopping cart library, WymEditor is a web-based HTML editor, TuduList is a web-based task application. TinyMCE is a Java-based WYSIWYG editor control. The applications range from 206 to 27K lines of JavaScript code.

The experimental objects are open-source and cover different application types. All the applications are interactive in nature and extensively use JavaScript on the client-side.

### B. Setup

To address our research questions, we provide the URL of each experimental object to JSEFT. Test cases are then automatically generated by JSEFT. We give JSEFT 10 minutes in total for each application. 5 minutes of the total time is designated for the dynamic exploration step.

**Test Case Generation (RQ1).** To measure client-side code coverage, we use JSCover [23], an open-source tool for measuring JavaScript code coverage. We report the average results over five runs to account for the non-determinism behaviour that stems from crawling the application. In addition, we assess each step in our approach separately as follows: (1) compare the statement coverage achieved by our function coverage maximization with a method that chooses the next state/event for the expansion uniformly at random, (2) assess the efficacy of our function state abstraction method (Algorithm 1), and (3) evaluate the effectiveness of applying mutation techniques (Algorithm 2) to reduce the number of assertions generated.

**Test Oracles (RQ2).** To evaluate the fault finding capability of JSEFT (RQ2), we simulate web application faults by automatically seeding each application with 50 random faults. We automatically pick a random program point and seed a fault at that point according to our fault category. While mutations used for oracle generation have been selectively generated (as discussed in Section IV-C), mutations used for the purpose of evaluation are randomly generated from the entire application. Note that if the mutation used for the purpose of evaluation and the mutation used for generating oracles happen to be the same, we remove the mutant from the evaluation set. Next we run the whole generated test suite (including both function-level and event-based test cases) on the faulty version of the application. The fault is considered detected if an assertion generated by JSEFT fails and our manual examination confirms that the failed assertion is detecting the seeded fault. We measure the precision and recall as follows:

**Precision** is the rate of injected faults found by the tool that are actual faults: $\frac{TP}{TP + FP}$

**Recall** is the rate of actual injected faults that the tool finds: $\frac{TP}{TP + FN}$

where $TP$ (true positives), $FP$ (false positives), and $FN$ (false negatives) respectively represent the number of faults that are correctly detected, falsely reported, and missed.

**Comparison (RQ3).** To assess how JSEFT performs with respect to existing JavaScript test automation tools, we compare its coverage and fault finding capability to that of ARTEMIS [2]. Similar to JSEFT, we give ARTEMIS 10 minutes in total for each application; we observed no improvements in the results obtained from running ARTEMIS for longer periods of time. We run ARTEMIS from the command line by setting the iteration option to 100 and enabling the coverage priority strategy, as described in [2]. Similarly, JSCover is used to measure the coverage of ARTEMIS (over 5 runs). We use the output provided by ARTEMIS to determine if the seeded mutations are detected by the tool, by following the same procedure as described above for JSEFT.

### C. Results

**Test Case Generation (RQ1).** Figure 5 depicts the statement coverage achieved by JSEFT for each application. The results show that the test cases generated by JSEFT achieve a coverage of 68.4% on average, ranging from 41% (ID 12) up to 99% (ID

---

TABLE I. CHARACTERISTICS OF THE EXPERIMENTAL OBJECTS.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>LOC</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SameGame</td>
<td>206</td>
<td>crawljax.com/same-game/</td>
</tr>
<tr>
<td>2</td>
<td>Tunnel</td>
<td>334</td>
<td>arcade.christianmontoya.com/tunnel/</td>
</tr>
<tr>
<td>3</td>
<td>GhostBusters</td>
<td>282</td>
<td>tdk.anerentapart.com/2/tplaside/63/</td>
</tr>
<tr>
<td>4</td>
<td>Peg</td>
<td>509</td>
<td><a href="http://www.ccomthnewhunreg/pregoame.htm">www.ccomthnewhunreg/pregoame.htm</a></td>
</tr>
<tr>
<td>5</td>
<td>HennyHunt</td>
<td>580</td>
<td>themannblinde.com/experiment/HennyHunt/</td>
</tr>
<tr>
<td>6</td>
<td>AjaxTabs</td>
<td>592</td>
<td>github.com/amazingSurge/jquery-tabs/</td>
</tr>
<tr>
<td>7</td>
<td>NarrowDesign</td>
<td>1.245</td>
<td><a href="http://www.narrowdesign.com">http://www.narrowdesign.com</a></td>
</tr>
<tr>
<td>8</td>
<td>JointLondon</td>
<td>1.231</td>
<td><a href="http://www.jointlondon.com">http://www.jointlondon.com</a></td>
</tr>
<tr>
<td>9</td>
<td>FractalViewer</td>
<td>1.235</td>
<td>onecm.com/projects/canopy/</td>
</tr>
<tr>
<td>10</td>
<td>SimpleCart</td>
<td>1.900</td>
<td>github.com/wjosdesign/simplecart-jp/</td>
</tr>
<tr>
<td>11</td>
<td>WymEditor</td>
<td>3.035</td>
<td><a href="http://www.wymeditor.org">http://www.wymeditor.org</a></td>
</tr>
<tr>
<td>12</td>
<td>TuduList</td>
<td>1.963</td>
<td><a href="http://tudu.ess.ch/tudu">http://tudu.ess.ch/tudu</a></td>
</tr>
<tr>
<td>13</td>
<td>TinyMCE</td>
<td>26.908</td>
<td><a href="http://www.tinymce.com">http://www.tinymce.com</a></td>
</tr>
</tbody>
</table>

---

TABLE II. RESULTS SHOWING THE EFFECTS OF OUR FUNCTION COVERAGE MAXIMIZATION, FUNCTION STATE ABSTRACTION, AND MUTATION-BASED ORACLE GENERATION ALGORITHMS.

<table>
<thead>
<tr>
<th>App ID</th>
<th>St. Coverage</th>
<th>State Abstraction</th>
<th>Oracles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fun. cov. maximize (%)</td>
<td>Random exploration (%)</td>
<td>FP</td>
</tr>
<tr>
<td>1</td>
<td>95</td>
<td>80</td>
<td>447</td>
</tr>
<tr>
<td>2</td>
<td>78</td>
<td>78</td>
<td>828</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>66</td>
<td>422</td>
</tr>
<tr>
<td>4</td>
<td>75</td>
<td>75</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>49</td>
<td>45</td>
<td>534</td>
</tr>
<tr>
<td>6</td>
<td>78</td>
<td>75</td>
<td>797</td>
</tr>
<tr>
<td>7</td>
<td>63</td>
<td>58</td>
<td>1653</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>50</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>82</td>
<td>82</td>
<td>1509</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>69</td>
<td>71</td>
</tr>
<tr>
<td>11</td>
<td>56</td>
<td>54</td>
<td>1383</td>
</tr>
<tr>
<td>12</td>
<td>41</td>
<td>38</td>
<td>1530</td>
</tr>
<tr>
<td>13</td>
<td>51</td>
<td>47</td>
<td>1401</td>
</tr>
</tbody>
</table>

**AVG** 68.4  42.8  -  -  85.5  -
We investigated why JSEFT has low coverage for some of the applications. For instance, we observed that in JointLondon (ID 7), the application contains JavaScript functions that are browser/device specific, i.e., they are exclusively executed in Internet Explorer, or iDevices. As a result, we are unable to cover them using JSEFT. We also noticed that some applications required more time to achieve higher statement coverage (e.g., in NarrowDesign ID 8), or they have a large DOM state space (e.g., BunnyHunt ID 5) and hence JSEFT is only able to cover a portion of these applications in the limited time it had available.

Table II columns under “St. Coverage” present JavaScript statement coverage achieved by our function coverage maximization algorithm versus a random strategy. The results show a 9% improvement on average, for our algorithm, across all the applications. We observed that our technique achieves the highest improvement when there are many dynamically generated clickable DOM elements in the application, for example, GhostBusters (ID 3).

The columns under “State Abstract” in Table II present the number of function states before and after applying our function state abstraction algorithm. The results show that the abstraction strategy reduces function states by 85.5% on average. NarrowDesign (ID 7) and FractalViewer (ID 9) benefit the most by a 97% state reduction rate. Note that despite this huge reduction, our state abstraction does not adversely influence the coverage as we include at least one function state from each of the covered branch sets as described in Section IV-B.

The last two columns of Table II, under “Oracles”, present the number of assertions obtained by capturing the whole application’s state, without any mutations, and with our mutation-based oracle generation algorithm respectively. The results show that the number of assertions is decreased by 86.5% on average due to our algorithm. We observe the most significant reduction of assertions for JointLondon (ID 7) from more than 198000 to 342.

Fault finding capability (RQ2). Table III presents the results on the fault finding capabilities of JSEFT. The table shows the total number of injected faults, the number of false negatives, false positives, true positives, and the precision and recall of JSEFT.

JSEFT achieves 100% precision, meaning that all the detected faults reported by JSEFT are real faults. In other words, there are no false-positives. This is because the assertions generated by JSEFT are all stable i.e., they do not change from one run to another. However, the recall of JSEFT is 70% on average, and ranges from 48 to 100%. This is due to false negatives, i.e., missed faults by JSEFT, which occur when the injected fault falls is either in the uncovered region of the application, or is not properly captured by the generated oracles.

The table also shows that on average 32% percent of the injected faults (ranges from 15–73%) are detected by function-level test cases, but not by our DOM event-based test cases. This shows that a considerable number of faults do not propagate to observable DOM states, and thus cannot be captured by DOM-level event-based tests. For example in the SimpleCart application (ID 10), if we mutate the mathematical operation that is responsible for computing the total amount of purchased items, the resulting error is not captured by event-based tests as the fault involves internal computations only. However, the fault is detected by a function-level test that directly checks the returned value of the function. This points to the importance of incorporating function-level tests in addition to event-based tests for JavaScript web applications. We also observed that even when an event-based test case detects a JavaScript fault, localizing the error to the corresponding JavaScript code can be quite challenging. However, function-level tests pinpoint the corresponding function when an assertion fails, making it easier to localize the fault.

Comparison (RQ3). Figure 5 shows the code coverage achieved by both JSEFT and ARTEMIS on the experimental objects running for the same amount of time, i.e., 10 minutes. The test cases generated by JSEFT achieve 68.4% coverage on average (ranging from 41–99%), while those generated by ARTEMIS achieve only 44.8% coverage on average (ranging from 0–92%). Overall, the test cases generated by JSEFT achieve 53% more coverage than ARTEMIS, which points to the effectiveness of JSEFT in generating high coverage test cases. Further, as can be seen in the bar plot of Figure 5,
for all the applications, the test cases generated by JSEFT achieve higher coverage than those generated by ARTEMIS. This increase was more than 226% in the case of Bunnyhunt (ID 5). For two of the applications, NarrowDesign (ID 7) and JointLondon (id 8), ARTEMIS was not able to complete the testing task within the allocated time of ten minutes. Thus we let ARTEMIS run for an additional 10 minutes for these applications (i.e., 20 minutes in total). Even then, neither application completes under ARTEMIS.

Table III shows the precision and recall achieved by JSEFT and ARTEMIS. With respect to fault finding capability, unlike ARTEMIS that detects only generic faults such as runtime exceptions and W3C HTML validation errors, JSEFT is able to accurately distinguish faults at the code-level and DOM-level through the test oracles it generates. Both tools achieve 100% precision; however, JSEFT achieves five-fold higher recall (70% on average) compared with ARTEMIS (12.8% recall on average).

D. Threats to Validity

An external threat to the validity of our results is the limited number of web applications that we use to evaluate our approach. We mitigated this threat by using JavaScript applications that cover various application types. Another threat is that we validate the failed assertions through manual inspection, which can be error-prone. To mitigate this threat, we carefully inspected the code in which the assertion failed to make sure that the injected fault was indeed responsible for the assertion failure. Regarding the reproducibility of our results, JSEFT and all the applications used in this study are publicly available, thus making the study replicable.

VI. CONCLUSIONS

In this paper, we presented a technique to automatically generate test cases for JavaScript applications at two complementary levels: (1) individual JavaScript functions, (2) event sequences. Our technique is based on algorithms to maximize function coverage and minimize function states needed for efficient test generation. We also proposed a method for effectively generating test oracles along with the test cases, for detecting faults in JavaScript code as well as on the DOM tree. We implemented our approach in an open-source tool called JSEFT. We empirically evaluated JSEFT on 13 web applications. The results show that the generated tests by JSEFT achieve high coverage (68.4% on average), and that the injected faults can be detected with a high accuracy rate (recall 70%, precision 100%).

ACKNOWLEDGMENT

This work was supported in part by a Strategic Project Grant (SPG) from the Natural Science and Engineering Research Council of Canada (NSERC), and a research gift from Intel Corporation. We thank the anonymous reviewers of ICST’15 for their insightful comments.

REFERENCES


