Flycatcher

Automatic unit test generation for JavaScript

Interim Report

Author: Jerome de Lafargue
Supervisor: Susan Eisenbach
Technical Supervisor: Tristan Allwood

Department of Computing
Imperial College London

Contents

1 Introduction **2**

2 Background **3**

2.1 Dynamic software testing 3

2.1.1 Overview 3

2.1.2 Structural testing 4

2.1.3 Unit testing 4

2.1.4 Regression testing 5

2.2 Automatic test data generation 5

2.2.1 Overview 5

2.2.2 Static test data generation 6

2.2.3 Hybrid test data generation 7

2.2.4 Dynamic test data generation 8

2.2.5 Genetic algorithms 9

2.2.6 Challenges of dynamic languages 10

2.3 Object-oriented test case generation 12

2.4 JavaScript 12

2.4.1 Prominence 12

2.4.2 Idiosyncratic features 12

2.4.3 Object-oriented programming 12

3 Project plan **13**

4 Evaluation strategy **14**

Chapter 1

Introduction

Notes on [19]

- testing major part of the development process, crucial for quality and accounts for 50% or more of the development effort
- manual process is tedious, costly and difficult and the testing is often biased
- exhaustive enumeration of inputs is infeasible for a reasonable program, coverage gives a good indication of test quality but full coverage is often infeasible too therefore ideally ATDG should achieve the best possible coverage
- random methods are easy to implement but unreliable and unlikely to discover deep errors
- size and complexity of real software means a lot of the research deals with toy examples/language subsets. ATDG 'undecidable' problem.
- metaheuristic search techniques use an *objective function* that estimates the value of a solution
- as more and more people rely on software it will be required to be of greater quality

Chapter 2

Background

In this chapter we will start by giving an overview of software testing, with particular emphasis on aspects of it that are relevant to this project. We will then take a look at the state of the art in automatic test data generation (ATDG) in order to understand the approach that will be used for Flycatcher. Because the tests will be object-oriented, method call sequences also need to be generated for the test cases and we will look at the state of the art for doing that too. Finally we will explain why we chose JavaScript as our target language and describe features of it that are of interest to us for this project.

2.1 Dynamic software testing

2.1.1 Overview

We can define the activity of dynamic testing as testing that requires execution of the software with test data as input [17] and characterise it with respect to three parameters namely the amount of knowledge assumed by the tester, the target of the tests and the stage of development at which they are executed. The amount of knowledge of the software under test can be divided into three categories, structural (white-box testing) testing, functional (black-box testing) and a hybrid of the two (grey-box testing). The target of the tests refers to their granularity, from testing specific units of code (unit testing) to an entire integrated system (system testing). The stage at which the tests are undertaken determines whether they are regression tests, alpha-tests, beta-tests, acceptance tests *etc.* With Flycatcher we hope to generate suites of structural tests, focused at the unit level of object-oriented classes, most likely to perform incremental regression testing. Hence, structural testing, unit testing and regression testing will be described in more detail in this section.

2.1.2 Structural testing

The goal of structural testing is to test the internal workings [19] of an application in order to verify that it does not contain errors. While functional testing determines whether the software provides the required functionality, structural testing tries to ensure that it does not crash under any circumstances, regardless of how it is called. It concerns *how* well the software operates, its structure, rather than *what* it can do, its function. As a result, the measure to determine good structural testing is the code covered during the testing process — code coverage. It gives us an idea of the amount of code that should be bug free. However, there are various types of code coverage criteria and the confidence that our code is bug free varies depending on which one is chosen.

Test coverage

Edvardsson lists the most cited criteria [7], from weakest to strongest:

- **Statement Coverage** Each statement must be executed at least once.
- **Branch[1]/Decision Coverage** Each branch condition must evaluate to true and false.
- **Condition/Predicate Coverage** Each clause within each branch condition must evaluate to true and false.
- **Multiple-condition Coverage** Each combination of truth values of each clause of each condition must be executed.
- **Path[1] Coverage** Each path in the control flow graph[1] must be traversed.

  The stronger criteria of condition, multiple-condition and path coverage are often infeasible to achieve for programs of more than moderate complexity, and thus branch coverage has been recognised as a basic measure for testing [7].

2.1.3 Unit testing

Unit testing consists in testing individual and independently testable units of source code [23]. Therefore, unit testing is made easier if the code is designed in a modular way. The nature of the units depends on the programming language and environment but is often a class or a function. As opposed to system tests which can be aimed at the client, unit tests are usually white-box tests. Although they do not guarantee that the overall software works

[1] For an explanation of program analysis terminology such as branch, path and control flow graph see Appendix A.

as required, they give confidence in specific units of code and narrow down errors, helping the development process. In Flycatcher, the target unit will be a JavaScript prototype, which will be introduced later in this chapter.

2.1.4 Regression testing

Automatically generating structural unit tests can be of great use for regression testing. Regression testing aims to ensure that enhancing a piece of software does not introduce new faults [23]. The difficulty in testing this is that programmers do not always appreciate the extent of their changes. Hence, having a suite of unit tests with good structural coverage can reduce this problem by verifying the software in a systematic, unbiased way.

2.2 Automatic test data generation

2.2.1 Overview

As can be seen in Mahmood's systematic review of ATDG techniques [17], many classifications exist for ATDG techniques. For our purposes, the first distinction that we need to make is between white-box, black-box and grey-box ATDG techniques, as for Flycatcher we are only interested in white-box testing. In the literature we found that white-box ATDG techniques are usually classified in two ways [17, 7, 30].

The first concerns the target selection stage of ATDG techniques: where either paths or individual nodes that contribute to the overall coverage criterion are successively selected from the control flow graph, so that test data that respectively traverses the path or reaches the node can be generated. When specific paths are targeted, the ATDG technique is known as *path-oriented* [7] whereas if a node is targeted then it is *goal-oriented.* When data is generated purely randomly *i.e.* there is no specific target, then as part of this classification the ATDG technique is simply *random.*

The other classification of white-box ATDG concerns the type of implementation: *static*, *dynamic* or a *hybrid* of the two [12, 19]. We will focus on the latter classification of structural testing as it governs our choice of implementation for Flycatcher. Moreover, the former concerns the path selection stage of ATDG and this step will be ignored in Flycatcher as it is in many recent ATDG techniques [30]. Figure 2.1 summarises what we believe is an intuitive characterisation of ATDG techniques with respect to this project and the one which will guide our choice of implementation. Many techniques can be found under each of the static, dynamic and hybrid implementation categories and we will only list the most noteworthy to us.

The choice of implementation for Flycatcher is dynamic random ATDG for our benchmark and dynamic search-based ATDG using genetic algorithms for our solution. The rest of this section will present in further detail

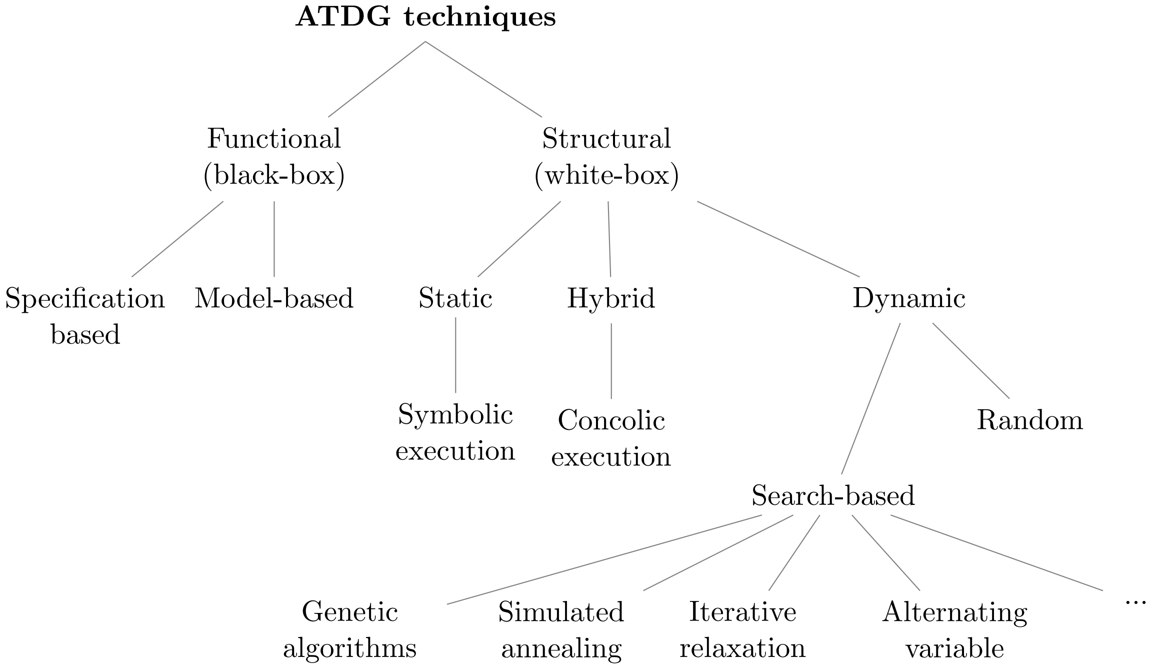

Figure 2.1: Overview of ATDG techniques

the structural ATDG implementation categories we have chosen and the difficulties of ATDG for a dynamic language, so that we can understand the rationale behind this choice.

2.2.2 Static test data generation

Static structural test data generation is based on information available from the static analysis of the program, without requiring that the program is actually executed [19]. Static program analysis produces control flow information that can be used to select execution paths in order to try and achieve good coverage. The goal of ATDG is then to generate data that executes these paths.

Every time control flow branches, *e.g.* at `if` statements, there is a corresponding predicate or branch condition. These predicates can be collected along a path and conjoined to form the path predicate. By solving the path predicate in terms of the input variables, we can obtain test data that executes that path. However, in order to rewrite the path predicate in terms of the input variables we need to take into account the execution of the program. Hence, for generation of test data statically a technique called symbolic execution [14] is used.

Symbolic execution gathers constraints along a simulated execution of a program path, where symbolic variables are used instead of actual values, such that the final path predicate can be rewritten in terms of the input variables. Solving the resulting system of constraints then yields the data necessary for the traversal of the given path [13, 14]. There are a lot of technical difficulties associated with symbolic execution [7, 20, 19]:

- the presence of input variable dependent loops can lead to infinite execution trees[2] as the loops can be executed any number of times
- array references become problematic if the indexes are not constants but variables, as is typically the case
- features such as pointers and dynamically-allocated objects that rely on execution are hard to analyse statically
- static analysis is not possible for function calls to precompiled modules or libraries
- if the path constraint is non-linear, solving it is an undecidable problem
- even if the path constraint is linear, solving it can lead to very high complexity

Although various static solutions have been proposed for these issues [27, 10, 24], they often dramatically increase the complexity of the ATDG process. As a result, tools purely based on symbolic execution can typically handle only subsets of programming languages and are not applicable in industry. A better trend that has developed in the past decade, is the combination of concrete and symbolic execution, which tackles most of the aforementioned issues [25] — we will cover this type of ATDG implementation in the subsection on hybrid ATDG. Due to the numerous problems posed by purely static ATDG, its weakness with dynamic types and constructs [7, 30] and the complexity of building a fully-fledged symbolic executor for a language [7, 12], we chose not to use static ATDG for the implementation of Flycatcher.

2.2.3 Hybrid test data generation

The hybrid approach to ATDG consists in combining symbolic and concrete execution, which is known as *concolic execution* [25]. In other words, hybrid analysis tools run programs on actual inputs, while collecting symbolic constraints in order to direct the search for new inputs. In doing so, they avoid the main weaknesses of the static approach, such as solving non-linear constraints or dealing with dynamic structures. This type of technique has been popular in recent years, mainly because it overcame the limitations

[2] the execution paths followed during the symbolic execution of a procedure [14]

that prevented static ATDG techniques from being applied to industry software. Notable tools that implement it are DART [9], CUTE [29], JPF-SE [2], PEX [31], EXE [5] and KLEE [4].

Yet, although it deals with some limitations of static ATDG, hybrid ATDG still requires static analysis of the source code under test, which is unfeasible for Flycatcher, given the dynamically typed, object-oriented nature of JavaScript.

2.2.4 Dynamic test data generation

Dynamic structural test data generation is purely based on actual execution of the software. The program under test is run with, possibly randomly, selected input and feedback is collected at runtime regarding the chosen coverage objective [7]. The feedback is usually obtained through some form of instrumentation of the program that monitors the program flow. Inputs can be continually generated randomly, relying on probability to achieve the coverage objective — this is known as *random* test data generation and does not perform well in general [7]. On the other hand, inputs can be incrementally tuned based on the feedback (using different kinds of search methods) in order to satisfy the coverage objective — this is known as *search-based* test data generation [19], where the search-space is the control flow graph of a program. The main drawback of dynamic ATDG is that it is reliant on the speed of execution of a program and as the number of required executions to achieve satisfactory coverage may be high, this leads to an overall expensive process. Below we will present the random and search-based approaches in more detail, in order to understand which would be more suitable for Flycatcher.

Random approach

Random test data generation consists in producing inputs at random in the hope of achieving the chosen coverage criterion through probability. Although random test data generation is relatively simple to implement, it does not perform well in terms of coverage, as the chances of finding faults that are revealed by only a small percentage of program inputs are low [7]. In other words, it is difficult for it to exercise ‘deep’ features of a program that are exercised only through specific and unlikely paths. As a result, random ATDG only works well for straightforward programs. However, because it is the simplest ATDG technique and is considered to have the lowest acceptance rate [7], it is often used as a benchmark and is a suitable candidate for us to benchmark our Flycatcher application.

Search-based approach

Search-based test data generation uses heuristics to guide the generation of input data so that the inputs execute paths that contribute to the overall test coverage objective. This involves modelling the test coverage objective as a heuristic function or *objective function*, that evaluates the fitness of a chosen set of inputs with respect to a coverage objective. Based on those fitness values, many search techniques exist to find optimal inputs in order to achieve the desired coverage. Various objective functions exist and they are dependent on the ATDG method used. Some of the well-known search-based ATDG techniques are *alternating variable* (local search optimisation) [15, 8], *simulated annealing* [33, 32], *iterative relaxation* [11] and *genetic algorithms* [21, 22].

In 2008, Han and Kwon [12] conducted a robust empirical evaluation of search-based test data generation techniques, comparing iterative-relaxation, local search optimisation and genetic algorithms. Genetic algorithms came out on top of the other techniques regarding the rate of coverage and generality, both essential characteristics when considering ATDG techniques. However, genetic algorithms proved expensive both in time and resources, but this can be improved upon [12]. Although simulated annealing is not part of this empirical evaluation, we can see from [21] that although it performs as well as genetic algorithms, it appears to be less efficient.

Additionally, genetic algorithms offer another advantage: as opposed to most other ATDG techniques, they do not require static analysis of the program under test [12]. This advantage is especially significant in our case, as to the best of our knowledge, tools that produce control and data flow analysis information for JavaScript source code are not available. However, this is not surprising since the dynamic, object-oriented nature of JavaScript makes this task very laborious.

Consequently, due to their effectiveness in terms of coverage and the advantages they possess with regard to dynamic languages, genetic algorithms appear to be the most suitable for our implementation of ATDG for Flycatcher. We will therefore go on to explain what genetic algorithms are and how they can be applied to test data generation.

2.2.5 Genetic algorithms

Overview

Genetic algorithms attempt to model the robustness and flexibility of natural selection in order to guide a search[22]. They start with a randomly initialised *population* of potential candidate solutions, called *individuals* or *chromosomes*. The population is iteratively recombined and mutated to evolve successive populations, known as *generations*. The recombination takes the ‘fittest’ parent solutions and ‘breeds’ them to produce new offspring. Their fitness is determined by a *fitness function* that evaluates how good a candidate solution is for a particular problem.

This process favours evolution towards fitter individuals, mimicking natural selection. The 'breeding' of two individuals usually involves a crossover operation which swaps their information at a randomly selected position. In the interest of maintaining diversification, a *mutation* phase usually occurs after the crossover, to introduce new genetic material into the search and avoid premature convergence on one area of the search space. The mutation operation randomly modifies some information of a selected individual, The process of iterative recombination and mutation, illustrated in Figure 123, is repeated until a specific termination criterion is fulfilled.

Application to test data generation

For the purposes of ATDG, the population is one of test cases. The test cases are evolved according to the genetic algorithm in order to satisfy the chosen coverage criteria [22]. In other words, the *fitness function* is based on the *objective function*, the function that evaluates test data according to a certain coverage criterion — the more a candidate test case contributes to coverage, the fitter it is. Instead of using search heuristics to minimise our objective function, we use it to guide our natural selection process towards the test cases that will achieve the best coverage.

Regarding the objective function used in the context of genetic algorithms, different variants exist [19]. The *coverage-oriented approach* rewards individuals on the basis of *all* covered program structures. This type of objective function rewards coverage with respect to the overall goal *i.e.* the system is not required to pick a specific path or node and attempt to execute it. The other approaches are *structure-oriented* where, similarly to most other search-based ATDG methods, a path or node is selected from the data flow analysis and a separate search is undertaken for each selected structure. Once more, seeing that the former does not require static analysis of the program under test, it is the most attractive to us for the implementation of Flycatcher.

Finally, the population, individual representation, fitness evaluation and methods of crossover, mutation and selection are all highly dependent on the test data generation problem at hand, hence we will discuss them with regard to Flycatcher when we come to the implementation.

2.2.6 Challenges of dynamic languages

Most of the research on ATDG concerns static programming languages [17] and it is only in the past few years that dynamic programming languages have sparked some interest in that field. A possible reason for this is that dynamic programming languages make ATDG harder by enabling features that allow programs to significantly change at runtime. These features can include modifying the type system, extending objects or adding new code during program execution. The challenges this type of behaviour introduces, and that Flycatcher will try and overcome, are listed below [6].

Generating test data of the required type

Given that method parameters do not have static types in dynamically typed languages, we do not know what arguments to pass to them. A potential solution to this is to use a method called *type inference* [26], which tries to infer the type of arguments from the way they are used inside the program. Although this method does not guarantee 100% precision, it is a good starting point for generating accurately typed test data. Mairhofer uses this technique for RuTeG [18], his search-based ATDG tool for Ruby, where the search for test data refines the initially inferred type, by discarding poor candidates.

Generating object instances

Sometimes the type of an object constructor or method parameters will be a complex type and this complicates the test data generation task even further. Generating well-formed object instances to use as arguments inside tests for a dynamically typed object-oriented language is problematic because there isn't a blueprint to construct them from. There is previous work on input data generation for dynamic data structures [15, 34, 28, 35], but all these approaches focus on statically typed languages (C/C++), require static program analysis and mostly lack generality.

Another approach uses something called needed-narrowing [3] or lazy instantiation [16] — where instances are only created when they are actually put to use by the program. This enables test data generators to adjust object instances during execution, when attempts are made to use them, so that they always have the required type. This technique is used by Irulan [1] for generating tests in Haskell, which has lazy evaluation by default. For the purpose of complex type test data generation in Flycatcher, we will investigate whether we can mimic lazy evaluation for JavaScript objects in order to build test data with the appropriate complex type.

Identifying bugs

In dynamic languages such as JavaScript, the function signatures bear no type information and this makes it difficult to know whether an exception is raised due to a wrongly typed test argument or a true program bug. In the case where the exception is not a bug it could be due to two things: manipulating a badly initialised object or breaking a program precondition. The bad initialisation problem might be prevented through the use of lazy evaluation to initialise objects. The breaking of preconditions could be avoided by giving the tester the ability to impose restrictions on the test data generator, so that preconditions for the program are respected. Being able to identify real program bugs during test data generation is an issue that will need further attention during the implementation of Flycatcher.

Dealing with dynamically generated code

2.3 Object-oriented test case generation

2.4 JavaScript

2.4.1 Prominence

2.4.2 Idiosyncratic features

2.4.3 Object-oriented programming

Chapter 3

Project plan

Chapter 4

Evaluation strategy

Bibliography

[1] Allwood, T., Cadar, C., and Eisenbach, S. High coverage testing of haskell programs. In *Proceedings of the 2011 International Symposium on Software Testing and Analysis* (2011), ACM, pp. 375–385.

[2] Anand, S., Păsăreanu, C., and Visser, W. Jpf–se: A symbolic execution extension to java pathfinder. *Tools and Algorithms for the Construction and Analysis of Systems* (2007), 134–138.

[3] Antoy, S., Echahed, R., and Hanus, M. A needed narrowing strategy. In *Proceedings of the 21st ACM SIGPLAN-SIGACT symposium on Principles of programming languages* (1994), ACM, pp. 268–279.

[4] Cadar, C., Dunbar, D., and Engler, D. Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs. In *Proceedings of the 8th USENIX conference on Operating systems design and implementation* (2008), USENIX Association, pp. 209–224.

[5] Cadar, C., Ganesh, V., Pawlowski, P., Dill, D., and Engler, D. Exe: automatically generating inputs of death. *ACM Transactions on Information and System Security (TISSEC) 12*, 2 (2008), 10.

[6] Ducasse, S., Oriol, M., Bergel, A., et al. Challenges to support automated random testing for dynamically typed languages.

[7] Edvardsson, J. A survey on automatic test data generation. In *Proceedings of the 2nd Conference on Computer Science and Engineering* (1999), no. x, pp. 21–28.

[8] Gallagher, M., and Lakshmi Narasimhan, V. Adtest: A test data generation suite for ada software systems. *Software Engineering, IEEE Transactions on 23*, 8 (1997), 473–484.

[9] Godefroid, P., Klarlund, N., and Sen, K. Dart: directed automated random testing. In *ACM Sigplan Notices* (2005), vol. 40, ACM, pp. 213–223.

[10] Goldberg, A., Wang, T., and Zimmerman, D. Applications of feasible path analysis to program testing. In *Proceedings of the 1994 ACM SIGSOFT international symposium on Software testing and analysis* (1994), ACM, pp. 80–94.

[11] Gupta, N., Mathur, A., and Soffa, M. Automated test data generation using an iterative relaxation method. In *ACM SIGSOFT Software Engineering Notes* (1998), vol. 23, ACM, pp. 231–244.

[12] Han, S., and Kwon, Y. An empirical evaluation of test data generation techniques. *Journal of Computing Science and Engineering 2*, 3 (2008), 274–300.

[13] King, J. A new approach to program testing. *Programming Methodology* (1975), 278–290.

[14] King, J. Symbolic execution and program testing. *Communications of the ACM 19*, 7 (1976), 385–394.

[15] Korel, B. Automated software test data generation. *Software Engineering, IEEE Transactions on 16*, 8 (1990), 870–879.

[16] Lindblad, F. Property directed generation of first-order test data. *TFP 7* (2007), 105–123.

[17] Mahmood, S. A systematic review of automated test data generation techniques. *School of Engineering, Blekinge Institute of Technology Box 520* (2007).

[18] Mairhofer, S. Search-based software testing and complex test data generation in a dynamic programming language. *Master's thesis, Blekinge Institute of Technology* (2008).

[19] McMinn, P. Search-based software test data generation: a survey. *Software Testing, Verification and Reliability 14*, 2 (2004), 105–156.

[20] Meudec, C. Atgen: automatic test data generation using constraint logic programming and symbolic execution. *Software Testing, Verification and Reliability 11*, 2 (2001), 81–96.

[21] Michael, C., and McGraw, G. Automated software test data generation for complex programs. In *Automated Software Engineering, 1998. Proceedings. 13th IEEE International Conference on* (1998), IEEE, pp. 136–146.

[22] Michael, C., McGraw, G., and Schatz, M. Generating software test data by evolution. *Software Engineering, IEEE Transactions on 27*, 12 (2001), 1085–1110.

[23] Myers, G., Sandler, C., and Badgett, T. *The art of software testing*. Wiley, 2011.

[24] Offutt, A., Jin, Z., and Pan, J. The dynamic domain reduction procedure for test data generation. *Software-Practice and Experience 29*, 2 (1999), 167–194.

[25] Păsăreanu, C., and Visser, W. A survey of new trends in symbolic execution for software testing and analysis. *International Journal on Software Tools for Technology Transfer (STTT) 11*, 4 (2009), 339–353.

[26] Pluquet, F., Marot, A., and Wuyts, R. Fast type reconstruction for dynamically typed programming languages. In *ACM SIGPLAN Notices* (2009), vol. 44, ACM, pp. 69–78.

[27] Ramamoorthy, C., Ho, S., and Chen, W. On the automated generation of program test data. *Software Engineering, IEEE Transactions on*, 4 (1976), 293–300.

[28] Sai-ngern, S., Lursinsap, C., and Sophatsathit, P. An address mapping approach for test data generation of dynamic linked structures. *Information and Software Technology 47*, 3 (2005), 199–214.

[29] Sen, K., Marinov, D., and Agha, G. *CUTE: A concolic unit testing engine for C*, vol. 30. ACM, 2005.

[30] Tahbildar, H., and Kalita, B. Automated software test data generation: Direction of research. *International Journal of Computer Science and Engineering 2*.

[31] Tillmann, N., and De Halleux, J. Pex–white box test generation for. net. *Tests and Proofs* (2008), 134–153.

[32] Tracey, N., Clark, J., and Mander, K. The way forward for unifying dynamic test-case generation: The optimisation-based approach. In *International Workshop on Dependable Computing and Its Applications (DCIA)* (1998), pp. 169–180.

[33] Tracey, N., Clark, J., Mander, K., and McDermid, J. An automated framework for structural test-data generation. In *Automated Software Engineering, 1998. Proceedings. 13th IEEE International Conference on* (1998), IEEE, pp. 285–288.

[34] Visvanathan, S., and Gupta, N. Generating test data for functions with pointer inputs. In *Automated Software Engineering, 2002. Proceedings. ASE 2002. 17th IEEE International Conference on* (2002), IEEE, pp. 149–160.

[35] Zhao, R., and Li, Q. Automatic test generation for dynamic data structures. In *Software Engineering Research, Management & Applications, 2007. SERA 2007. 5th ACIS International Conference on* (2007), IEEE, pp. 545–549.