Specify and Measure, Cover and Reveal: A Unified Framework for Automated Test Generation

Sébastien Bardin, Nikolai Kosmatov, Michaël Marcozzi, Mickaël Delahaye

Abstract

Automatic test input generation (ATG) is a major topic in software engineering, analysis and security. In this paper, we bridge the gap between state-of-the-art white-box ATG techniques, especially Dynamic Symbolic Execution, and the diversity of test objectives that they may be used to cover in practice, including many of those defined by common source-code coverage criteria. We define a new coverage specification mechanism, called labels, for specifying test objectives, and prove it to be both expressive and amenable to efficient automation. We present an efficient approach for detecting – revealing – infeasible (i.e. unreachable) test objectives expressed as labels. We demonstrate that measuring the achieved coverage can be efficiently performed for labels. Finally, we propose an innovative extension of DSE resulting in an efficient support for label coverage, while the existing naive approach induces an exponential blow-up of the search space. Experiments show that our ATG technique yields very significant savings and confirm the interest of infeasible label detection, enabling to lift DSE to label coverage with only a slight overhead. Overall, we show that label coverage provides the basis of a rich framework allowing one to express and handle test objectives from various contexts in an efficient and generic manner. To illustrate this framework, we describe LTest, an all-in-one testing toolset based on labels and used in the industry, which offers automatic program annotation, ATG, coverage measurement and detection of infeasible test objectives.

Keywords: software testing, test generation, symbolic execution, static analysis, coverage criteria
1. Introduction

Context. Efficient test input generation is a major issue for software engineering, analysis and security, so that a tremendous amount of work has been carried out to develop Automatic Test Generation (ATG) techniques and apply them in various contexts. These techniques may select tests randomly or craft them to cover specific targets in the code, relying for example on search meta-heuristics or symbolic analysis to produce the relevant tests. In the latter case, advances in constraint solving and dynamic analysis have led to the surge of Dynamic Symbolic Execution (DSE) [46, 69, 74], implemented into many tools (e.g. [30, 31, 37, 38, 47, 71]) leading to impressive case-studies (e.g. [37, 38, 49]).

In so-called white-box testing, ATG tools are used to fulfill various kinds of test objectives defined from the code of the program under test. For instance, they can be employed to search for inputs triggering specific types of failures during code execution (like buffer overflows) or capable of stressing specific zones in the code (like those affected by a software patch). ATG may also be used to generate a general test suite for a piece of software, which is then passed to one or several external oracles, in order to assess for example functional correctness, security or performance. The more different code behaviours are exercised by the test suite, the better. A standard way of measuring this diversity involves coverage criteria [25, 76] (a.k.a. adequacy criteria). We focus in this paper on source code coverage criteria, simply referred to as coverage criteria in the following. Many such criteria have been defined along the years, from basic control-flow or data-flow criteria to mutations [44] and MCDC [43].

Problem. Common white-box ATG techniques may face issues to cope efficiently with the diversity of test objectives that they are confronted with in practice. For example, DSE mostly follows an exhaustive exploration of the path space of the program under test, aiming typically at covering most execution paths up to a given bound. While such a path-oriented exploration proves successful in some contexts, it is well known that the resulting test suite can miss interesting behaviors related to data rather than control. Moreover, standard DSE does not support coverage objectives defined over artifacts not explicitly present in the code, such as multiple-condition coverage [25] or mutations, while it could efficiently guide test generation towards covering them.

Another important issue is that many white-box test objectives are defined in a structural way, i.e. expressed in terms of generic code artifacts (e.g. cover all instructions, all decisions, all conditions, etc.), without taking into account the semantics of the program. This leads to the situation when some of the resulting concrete test objectives (instructions, decisions, conditions, etc.) can be impossible to activate by a test case – they are infeasible, i.e. uncoverable.

---

1 The scope of this paper does not include model-based testing techniques, aiming at producing inputs covering the specification of the code under test. We consider here only deterministic sequential programs, even if extending the presented techniques to concurrent programs can be an interesting work perspective.
Infeasible test objectives waste the test generation effort and prevent testers from measuring the objectives coverage ratio (proxy for test effectiveness [2, 3]) precisely.

Goals. Our first objective is to adapt a state-of-the-art ATG technique, namely DSE, to make it able to cover efficiently a wide class of test objectives derived from the source code of the tested program. Recent works have aimed at lifting DSE to various coverage criteria [51, 65, 66, 67, 68, 77], or improving DSE bug-detection abilities by making explicit run-time error conditions [40, 48, 53]. These approaches are mainly based on an instrumentation of the code under test and allow for black-box reuse of existing DSE tools. However, they come at a high price since they may induce a blow-up of the path space and a significant overhead (previous work [51, Table 2] reports on a 272x average time-overhead, with a worst case of 2,000x). While still relying on instrumentation of the tested program and emphasizing black-box reuse of DSE tools as much as possible, our goal is to provide DSE support for even more kinds of objectives with a minimal and acceptable overhead.

Our second objective is to automate three additional key testing services, which we argue should be typically performed before any actual input generation. First, we offer a simple, generic and formal way to specify what the test objectives are, for many kinds of programs and testing contexts. Second, we enable one to efficiently detect – reveal – those of the test objectives that are actually infeasible (or at least a significant part of them). Third, we provide means to measure the coverage ratio of any existing test suite w.r.t. the feasible specified objectives. One can then leverage our adapted DSE to cover efficiently those of the feasible specified objectives that have not been covered yet.

Approach. We introduce labels, i.e. predicates attached to given program instructions. Labels were named with reference to C labels, which attach an identifier to a given program instruction (ou labels attach predicates instead). We define label coverage, a new source code coverage criterion which appears to be both expressive and amenable to efficient automation. A label is covered if a test execution reaches the instruction and satisfies the predicate. Actually, labels can be thought of as a convenient specification mechanism for test objectives, enabling to simulate notably many common classes of coverage criteria in a unified way. This idea encompasses and extends several existing works [40, 48, 51, 53, 67, 77].

We propose two novel ways of taming the blow-up that may appear while trying to cover labels with DSE. Namely, we introduce a tight instrumentation, where “tight” is made precise in the paper, and a coupling between DSE and label coverage named iterative label deletion. Their combination results in a much more effective support for label coverage in DSE. In addition, both techniques can be implemented using black-box DSE tools. We also show that labels are amenable to an efficient coverage measurement and an efficient detection of infeasible test objectives. Experiments reveal that DSE with tight instrumentation, iterative label deletion and infeasible label detection reaches better label coverage ratios than vanilla DSE, with only a slight time overhead.
Overall, we demonstrate that label coverage provides the basis of a rich framework allowing one to express and handle various test objectives in an efficient and generic manner.

**Contributions.** The purpose of the paper is to provide a complete panorama of various efforts related to label coverage, including basic definitions, key results and techniques, as well as a summary of most recent extensions and industrial applications. Our main contributions are the following:

- We show that label coverage is expressive enough to faithfully emulate notably many standard white-box coverage criteria, from decision or condition coverage (Theorem 1) to advanced logic criteria (Theorem 2) and a substantial subset of weak mutations (the side-effect free fragment, Theorem 3). Labels can thus be seen as a convenient and powerful specification mechanism for coverage criteria.

- We demonstrate that infeasible labels can be detected by existing assertion checkers after translating labels into assertions (Lemma 4).

- We formally characterize the properties of the naive instrumentation used in previous works lifting DSE to some kinds of the test objectives that labels can encode. This instrumentation provides a sound way to achieve label coverage and leads to very efficient coverage score computation. However, it also yields an exponential increase as well as a complexification of the paths space (Theorem 7).

- We propose DSE*, a variant of DSE with efficient handling of all the kinds of test objectives encodable by labels. This approach relies on tight instrumentation and iterative label deletion to reduce the complexity introduced by labels. Tight instrumentation yields only a linear growth of the paths space without any complexification (Theorem 10). Both techniques are orthogonal and allow for a significant speed-up. Moreover, they can be both implemented either within the DSE algorithm or using existing DSE tools in a black-box manner.

- We have implemented DSE* inside the PathCrawler DSE tool [74]. Experiments show that tight instrumentation and iterative label deletion yield very significant reductions of both the search space and computation time compared to naive instrumentation (several orders of magnitude speed-up in some cases).

- Finally, we describe LTest, an all-in-one ATG toolset based on label coverage. Along with DSE*-based test generation, LTest offers several integrated services: program annotation with labels, label coverage score computation, as well as detection of infeasible labels via static analysis. Our experiments with LTest demonstrate that better ATG for label coverage can be achieved at a very reasonable cost compared to vanilla DSE. For example, considering the test objectives from the advanced MCDC
criterion over our benchmark programs, DSE* with infeasible label detection has a mean 1.85x time overhead compared to vanilla DSE, while the mean reported coverage ratio is increased from 78% to 91%.

Outline. First, we present a motivating example in Section 2. After detailing our basic notation (Section 3), we define labels and explore their expressiveness (Section 4). Next, we focus on automation. Detection of infeasible labels is described in Section 5. The naive instrumentation is studied in Section 6.1 and the optimized DSE* approach is presented in Section 6.2. Thereon, we describe LTest, the automated testing framework based on labels (Section 7). Our experiments are presented in Section 8, followed by the discussion of threats to their validity in Section 9. Recent extensions and applications of labels are summarized in Section 10. Finally, we discuss related work (Section 11) and provide a conclusion (Section 12).

Earlier works. The present paper attempts to offer the first consolidated panorama of more than six years of research and industry transfer efforts around labels. As a consequence, this paper provides an integrated, as well as carefully revised, enhanced and extended version of several previously published works. More precisely, Sections 3.1, 3.3, 4, 6, 8.1, 11 are based on previous work presented at ICST 2014 [32]. The principles described in Section 5 were first introduced at ICST 2015 [27]. Section 7 was originally presented at TAP 2014 [26]. These earlier works have been extended in several ways.

Firstly, we have provided more explanations wherever possible, additional examples of criteria simulation, clearly stated theorems for all theoretical results, a thorough discussion on the limitations of labels, a better description of the experimental protocols, and an extended related work. A new motivating example (Section 2) has also been included to emphasise the integrated vision of the various aspects of testing addressed in the paper. We have also better presented all the coverage criteria considered in the paper (Section 3.2).

Secondly, the experiments detailed in Section 8 have been extended compared to those presented in the original papers. Section 8.1 has been strengthened with a new comparison with random testing and a better comparison with standard DSE (e.g. including coverage information). Section 8.2 was added to provide a novel comparison between vanilla DSE and DSE* with infeasible label detection. Section 8.3 was added to provide results involving an advanced coverage criterion (MCDC). A new section was also added to discuss threats to validity (Section 9).

Finally, all recent extensions [8, 9, 19, 1], applications [35], and industrial adoption efforts [33, 34] are now synthesised in Section 10.

2. Motivating Example

In this paper, we introduce a unified framework for automated test input generation. Conceptually, we argue that the test generation process should be divided into four main activities: (1) specify the test objectives, (2) reveal the
infeasible objectives, (3) measure the objective coverage rate of the existing tests and (4) generate new tests to cover the uncovered test objectives. We illustrate now how our framework handles these four activities in two of the most common use cases of test generation tools, namely crafting coverage-adequate test suites and detecting runtime failures. To do so, we consider the problem of automated test generation for the simple C function `numPos` in Figure 1.

### 2.1. Specifying test objectives

To be efficient, automated test generation should be driven by precise and specific objectives to fulfill. Our framework introduces a generic test objective specification language called labels, which can encode many of the diverse kinds of test objectives occurring in the diverse use cases of test generation tools.

When a test generation tool is used to craft a test suite for a given program, the testers often aim at satisfying one of the many existing code coverage criteria, which define a set of syntactic elements in the program as the test objectives to be covered. Higher coverage test suites have indeed better chances to find bugs [2, 3]. For a significant proportion of these criteria, labels can be used to easily specify the syntactic elements which should be covered in the program under test to satisfy the criterion. As a simple example, the Decision Coverage criterion requires that running the test suite should cover all the branches in the control-flow graph of the program. For the program in Figure 1, this leads to four test objectives corresponding to the two branches of each of the two conditional statements. Each of these four test objectives can be encoded by one of the four labels $l_1, l_2, l_3$ and $l_4$ on Figure 2. A label is basically a code assertion and covering the test objective encoded by a label means crafting a test that makes the program execution reach and fulfill the assertion. For example, one can check that covering $l_1$ is equivalent to making the condition of the first conditional statement true, i.e. to covering the then branch of this statement.

Test generation tools can also be used to probe a program for different kinds of runtime failures (see [4] for a successful example), like integer overflows. For the program in Figure 1, this means crafting test inputs able to make one of the two `n++` statements overflow. These two test objectives can be encoded with the labels $l_5$ and $l_6$ on Figure 3.

Given a program under test and a coverage criterion (or well-defined kind of runtime failures to probe for), our framework enables annotating automatically the program with the corresponding labels.
2.2. Revealing infeasible test objectives

Many of the test objectives considered in automated test generation are purely syntactic and thus blind to the semantics of the program under test. As a consequence, a significant proportion of them can turn out to be infeasible, i.e. no input can lead to an execution satisfying them. For example, both labels \(l_5\) and \(l_6\) on Figure 3 are infeasible, because \(n\) can only range between 0 and 2 during program execution.

Infeasible objectives are a threat to the efficiency of test generation tools, because a significant part of the test budget might be lost trying to cover them. Our framework proposes to deal with this issue in a generic way, by introducing a sound approach to prune out infeasible objectives encoded as labels. In a nutshell, proving that a label is infeasible is equivalent to proving that its corresponding opposite assertion (i.e. with the negated predicate) can never be violated during any program execution. Proving the latter is a standard feature of many formal verification or model-checking tools, to which this proof can be delegated. While label infeasibility is not decidable, i.e. the tool may not always be able to conclude, leaving the status of some labels unresolved, our experiments show that existing tools are able to flag many infeasible labels in practice. For example, for the program of Figure 4 such a tool could show that none of its two assertions can be violated, hence we can deduce that \(l_5\) and \(l_6\) on Figure 3 are infeasible.
```
int numPos(int a, int b) {
    int n = 0;
    if (a > 0) {
        assert(n != MAX_INT); // Never violated -> 15 infeasible
        n++;
    }
    if (b > 0) {
        assert(n != MAX_INT); // Never violated -> 16 infeasible
        n++;
    }
    return n;
}
```

Figure 4: Assertions to detect infeasible test objectives for the test objectives of Figure 3.

2.3. Measuring the coverage

While building a test suite for a program, it is often useful to evaluate the strength of the produced test suite, to determine if additional tests should be generated to achieve a more acceptable coverage level or, on the contrary, if the test suite can be pruned to reduce the testing overhead. This is typically done by measuring which proportion of the test objectives from a chosen criterion are covered by the current test suite.

Our framework enables transparently instrumenting the program under test to measure the coverage level of an existing test suite, for any coverage criterion whose objectives can be encoded as labels. This can indeed be simply done through textually replacing any label by some code that reports the coverage of the corresponding objective. For example, considering the labels `l1` and `l2` of Figure 2, the instrumented code could be:

```
... int n = 0;
if (a > 0) report_covered_label("l1");
if (a <= 0) report_covered_label("l2");
if (a > 0) n++;
...
```

2.4. Generating test cases covering test objectives

While dozens of kinds of test objectives are used in the use cases of test generation tools, these different flavours of objectives are seen as dissimilar bases for automation, so that most tools only provide a direct support for a very small subset of them. Supporting new flavours of objectives is time-consuming. Our framework bridges the gap between the variety of test objective flavors and their limited support in test generation tools, by tailoring a state-of-the-art test generation approach, namely dynamic symbolic execution, to efficiently and generically cover objectives encoded as labels in the program under test. For example, considering the program of Figure 2 and its four labels, our tailored dynamic symbolic execution would need only 4 trials before covering all the labels, while the common naive approach might necessitate up to 16 trials.
3. Background

3.1. Notation

Given a program $P$ over a vector $V$ of $m$ input variables taking values in a domain $D \triangleq D_1 \times \cdots \times D_m$, a test datum $t$ for $P$ is a valuation of $V$, i.e. $t \in D$. The execution of $P$ over $t$, denoted $P(t)$, is a path (or run) $\sigma \triangleq (loc_1, S_1) \cdots (loc_n, S_n)$, where the $loc_i$ denote control-locations (or simply locations) of $P$ and the $S_i$ denote the successive internal states of $P$ ($\approx$ valuation of all global and local variables as well as memory-allocated structures) before the execution of each $loc_i$. A test datum $t$ reaches a location $loc$ with internal state $S$, denoted $t \xrightarrow{P} (loc, S)$, if $P(t)$ is of the form $\sigma_1 \cdot (loc, S) \cdot \sigma_2$. A test suite $TS$ is a finite set of test data.

Assume that for a given test objective $c$, there is an adequate notion of covering (left unspecified for the moment), and we write $t \xrightarrow{P} c$ if test datum $t$ covers $c$. We extend the notation for a test suite $TS$ and a set of test objectives $C$, writing $TS \xrightarrow{P} C$ when for any $c \in C$, there exists $t \in TS$ such that $t \xrightarrow{P} c$. A (source-code based) coverage criterion $C$ is defined as a systematic way of deriving a set of test objectives $C = C(P)$ for any program under test $P$. A test suite $TS$ satisfies (or achieves) a given coverage criterion $C$ if $TS$ covers $C(P)$. When no confusion is possible, we can identify the coverage criterion $C$ for a given program $P$ with the derived set of test objectives $C = C(P)$.

These definitions are generic and leave the exact definition of “covering” to the considered coverage criterion. For example, test objectives derived from the Decision Coverage criterion are of the form $c \triangleq (loc, \text{cond})$ or $c \triangleq (loc, !\text{cond})$, where $\text{cond}$ is the condition of the branching instruction at location $loc$, and $t \xrightarrow{P} c$ if $t$ reaches some $(loc, S)$ such that $\text{cond}$ evaluates to true (resp. false) in $S$.

Finally, given a test suite $TS$ and a set $C$ of test objectives, the coverage score of $TS$ w.r.t. $C$ is the ratio of the number of test objectives in $C$ covered by $TS$ to its cardinality $|C|$. The coverage score of $TS$ w.r.t. a coverage criterion $C$ is defined as its coverage score w.r.t. the set $C = C(P)$. We also designate it as the $C$ score (resp. $C$ score) of $TS$ to emphasize the underlying set of test objectives (resp. coverage criterion).

3.2. Coverage criteria

This section defines the standard coverage criteria (a.k.a adequacy criteria) used throughout the paper and their associated notions of covering. We follow the classification of Ammann and Offutt [25].

Control-flow graph and call graph coverage criteria. The Statement Coverage (SC) criterion (a.k.a. Instruction Coverage) requires a test suite to cover, that is, to reach, each statement of the program under test, while Decision Coverage (DC, a.k.a. Branch Coverage) requires a test suite to cover each branch of the program, that is, to activate both true and false branches of each program decision. Function Coverage (FC) is a restricted form of SC, requiring only to reach all function entrypoints.
Logic expression coverage criteria. The three simplest criteria of this family are CC, DCC and MCC. Condition coverage (CC) requires to activate both true and false values for each of the atomic conditions in the predicates appearing in any program decision point (e.g. conditional or loop predicates). Decision-Condition Coverage (DCC) requires to satisfy both DC and CC. Multiple-Condition Coverage (MCC) requires to activate all combinations of truth values of atomic conditions in each decision. MCDC is a family of more intricate logic expression coverage criteria [43], well-known for being required for certification of aeronautic software. In a nutshell, the MCDC criteria require to demonstrate that each single atomic condition, alone, can influence the value of the whole decision. In this work, we focus on two MCDC criteria, General Active Clause Coverage (GACC) and General Inactive Clause Coverage (GICC). Notice that while MCDC is much more complex to cover than DCC, it requires a number of tests only linear in the number of atomic conditions, whereas MCC may require an exponential number of tests. Finally, the Implicant Coverage (IC) and Unique True Point Coverage (UTPC) criteria also aim at covering different behaviors of the predicates appearing at the program decision points, but considering that these predicates have been first syntactically normalized into Disjunctive Normal Form (DNF). The reader can find a detailed definition of these criteria in [25]. (It will be explicitly expressed using labels below in the sketch of proof for Theorem 1.)

Mutation criteria. In mutation testing [44], test objectives consist of mutants, i.e. slight syntactic modifications of the program under test \( P \). In the strong mutation setting \( M \), a mutant \( M \) is covered (or killed) by a test datum \( t \) if the output of \( P(t) \) differs from the output of \( M(t) \). In the weak mutation setting \( WM \) [50], a mutant \( M \) is covered by \( t \) if the internal states of \( P(t) \) and \( M(t) \) differ from each other right after the mutated location (see Figure 5). \( M \) is a very powerful coverage criterion [24, 61]. While less powerful in theory, \( WM \) is almost equivalent to \( M \) in a practical setting [59]. Mutation testing is parameterized by a set of mutation operators \( O \). An (atomic) mutation operator \( op \in O \) is a function mapping a program \( P \) into a finite set of well-defined programs (mutants), such that \( P \) differs from each mutant \( M \) in only one location. We denote \( \text{M}_O \) and \( \text{WM}_O \) the strong and weak mutation criteria restricted to mutants created through operators in \( O \).

Black-box criteria. The Input Domain Coverage criterion (IDC) assumes a partition of the input domain \( D \) of \( P \) given as disjoint predicates \( \varphi_1, \ldots, \varphi_k \), and consists of considering one input for each \( \varphi_i \).

3.3. Symbolic execution

We recall here a few basic facts about Symbolic Execution (SE) [23, 54]. Let us consider again the program \( P \) presented in Section 3.1 and its input variables \( V \), defined over domain \( D \). Let us also consider an execution path \( \sigma \) in the control-flow graph of \( P \), i.e. a path in the graph linking the starting node of \( P \) to one of its intermediate or exiting nodes. The goal of SE is then to generate an input valuation \( t_\sigma \in D \) so that \( P(t_\sigma) \) covers (i.e. activates) \( \sigma \).
The key insight of SE is that, if $P$ is deterministic, it is possible to compute a path constraint $\phi_\sigma$ for $\sigma$ such that for any input valuation $t \in D$, we have: $t$ satisfies $\phi_\sigma$ iff $P(t)$ covers $\sigma$. Indeed, such a path constraint $\phi_\sigma$ can be built by virtually executing $P$ over symbolic inputs and aggregating the constraints that arise over these symbolic inputs, as one forces the execution to follow path $\sigma$. For example, executing the program of Figure 1 over the two symbolic inputs $a$ and $b$ and forcing the execution to follow the path along the then branches of the two conditional statements produces the path constraint $a > 0 \land b > 0$.

Once the path constraint is built, it is solved using an off-the-shelf constraint solver, yielding the expected input valuation $t_\sigma$ or proving that the path is infeasible if the constraint is unsatisfiable. In practice, the path constraint must sometimes be under-approximated, as $P$ might contain statements (like calls to external libraries) from which the corresponding constraints cannot be easily extracted. Moreover, SE requires the availability of an (efficient) solving procedure for the conditions of the path constraint. These two issues have nevertheless been strongly alleviated during the last two decades with (a) the rise of Dynamic Symbolic Execution (DSE), which interleaves concrete and symbolic execution and uses the dynamically collected data to suggest better approximations for the path constraints, and (b) the development of fast constraint solvers based on a Satisfiability Modulo Theories (SMT) approach.

Nowadays, DSE is studied and used by a large and dynamic research community and it is the core principle of a wide variety of test input generation tools, successfully applied in the industry. Yet, the large amount of time that
Input: a program $P$ and a finite set of its execution paths $Paths(P)$

Output: a test suite $TS$, i.e. a set of pairs $(t, \sigma)$, such that

$\forall$ feasible $\sigma \in Paths(P)$, $\exists (t, \sigma) \in TS$, $P(t) \rightsquigarrow_{P} \sigma$

$TS := \emptyset$;

$S_{paths} := Paths(P)$;

while $S_{paths} \neq \emptyset$ do

choose $\sigma \in S_{paths}$; $S_{paths} := S_{paths} \setminus \{\sigma\}$;

compute path constraint $\phi_{\sigma}$ for $\sigma$;

switch $\text{Solve}(\phi_{\sigma})$ do

  case $\text{sat}(t)$ do $TS := TS \cup \{(t, \sigma)\}$;

  case $\text{unsat}$ do skip;

end

end

return $TS$;

Algorithm 1: Symbolic Execution algorithm

may be required to attempt solving the constraints (often written in an undecidable logic), together with the explosion in the number of paths to process, remain the two main bottlenecks faced by the technique when used to generate a test suite $TS$ covering a significant set of paths in a real-world application.

A simplified description of the SE process is depicted in Algorithm 1. While highly abstracted, it is sufficient to understand the remainder of the paper. Note that $\text{Solve}$ represents a call to the constraint solver, which can either return unsat (no solution found) or $\text{sat}(t)$ (where $t$ is a solution).

4. Generic Specification of Test Objectives with Labels

4.1. Definitions

Given a program $P$, a label $l$ is a pair $(loc, \varphi)$ where $loc$ is a location of $P$ and $\varphi$ is a predicate such that:

- $\varphi$ contains only valid expressions using variables visible at location $loc$;
- $\varphi$ contains no side-effect expressions

An annotated program is a pair $\langle P, L \rangle$ where $L$ is a set of labels defined over $P$. A test datum $t$ covers a label $l$ if $t \rightsquigarrow_{(P, L)} l$, if $t$ reaches (at least once) the location $loc$ with some internal state $S$ such that the predicate $\varphi$ is satisfied in $S$.

---

We choose to forbid side-effect expressions in label predicates for practical reasons, as it would make the implementation of our testing framework more complex. For example, measuring label coverage in a safe way would require to sandbox the side-effects occurring during label predicate evaluation or to undo them after it.
4.2. Simulating coverage criteria using label coverage

Given an annotated program \( \langle P, L \rangle \), we define the label coverage criterion, denoted LC, as the function returning \( L \) as the set of test objectives. Thus, a test suite satisfies LC if it covers all labels in \( L \), denoted \( TS \sim_{\psi(P)} \text{LC} \).

We seek to characterize the power of the LC coverage criterion to emulate other criteria. A key notion here is that of labelling function. A labelling function \( \psi \) maps a program \( P \) into an annotated program \( \psi(P) \triangleq \langle P, L \rangle \).

**Definition 1.** A coverage criterion \( C \) can be simulated by LC if there exists a labelling function \( \psi \) that annotates any given program \( P \) with labels corresponding to the test objectives derived following \( C \), so that, for any test suite \( TS \), we have \( TS \sim_{\psi(P)} \text{LC} \) iff \( TS \sim_{\psi(P)} \text{C} \).

In order to make the test objectives of some of the criteria discussed below directly encodable by labels, we consider in the rest of the paper only normalized programs, i.e. programs such that no side-effect occurs in any condition of a branching instruction. This is not a fundamental restriction since any (well-defined) program \( P_1 \) can be rewritten into a normalized program \( P_2 \), using intermediate variables to evaluate the side-effect prone conditions outside the branching instruction. For example, \( \text{if} \ (x++ <= y \&\& e == f) \ \{ \ldots \} \) becomes \( \text{tmp} = x++; \ \text{if} \ (\text{tmp} <= y \&\& e == f) \ \{ \ldots \} \). Notice that similar transformations are automatically performed by the Cil library [63] frequently used by DSE tools for C programs [69, 74].

**Basic graph and logic expression coverage criteria.** To simulate SC and FC, we add one label with a true predicate, respectively, before each statement of the program and at the beginning of each function body. To emulate DC, CC, DCC and MCC, we introduce one label per truth value to cover before any decision in \( P \). Some illustrating examples are given in Figure 6. For instance, for CC the corresponding labelling function \( \psi_{\text{CC}} \) inserts labels that enforce coverage of both truth values of the two atomic conditions \( x==y \) and \( a<b \).

**Theorem 1.** The coverage criteria SC, DC, CC, DCC and MCC can be simulated by LC.

**Sketch of proof.** We need to define a suitable labelling function for any of the considered criteria. For SC, we choose the labelling function \( \psi_{\text{SC}} \) adding all labels of the form \((\text{loc, true})\), where \( \text{loc} \) is any location of \( P \). Given a test suite \( TS \), \( TS \sim_{\psi(P)} \text{SC} \) iff \( TS \) can reach any \( \text{loc} \) of \( P \) iff \( TS \) covers any \((\text{loc, true})\) iff \( TS \sim_{\psi_{\text{SC}}(P)} \text{LC} \). We conclude that SC can be simulated by LC. The reasoning is similar for FC. Other criteria are handled similarly. The labelling function \( \psi_{\text{DC}} \) adds the set of all \((\text{loc, } \varphi)\) and \((\text{loc, } \neg \varphi)\), where \( \text{loc} \) contains a conditional statement with condition \( \varphi \). \( \psi_{\text{CC}} \) adds the set of all \((\text{loc, } a_i)\) and \((\text{loc, } \neg a_i)\), where \( \text{loc} \) contains a conditional statement whose atomic conditions are exactly the \( a_i \). \( \psi_{\text{DCC}} \) adds the union of labels added by \( \psi_{\text{DC}} \) and \( \psi_{\text{CC}} \). \( \psi_{\text{MCC}} \) adds the set of all \((\text{loc, } \bigwedge_i \bar{a}_i)\), where \( \text{loc} \) contains a conditional statement whose atomic conditions are the \( a_i \), and \( \bar{a}_i \) denotes either \( a_i \) or \( \neg a_i \).
Advanced logic expression coverage criteria. Pandita et al. [68] show that GACC (and thus GICC) can be simulated through additional branches to cover, which can be directly specified in terms of labels, as we did for DC. For IC and UTPC, we introduce one label per truth value to cover in the DNF of each decision predicate in $P$.

**Theorem 2.** The coverage criteria GACC, GICC, IC and UTPC can be simulated by LC.

**Sketch of proof.** Let us consider a predicate $p$ in $P$ that involves $n$ atomic conditions $c_1, \ldots, c_n$. GACC requires that for each clause $c_i$, the test suite triggers...
two distinct evaluations of \( p \): one execution \( A \) where \( c_i \) is true, one execution \( B \) where \( c_i \) is false, and both such that the truth value of \( c_i \) impacts the truth value of the whole predicate, i.e.:

\[
p(c_1^A, \ldots, c_{i-1}^A, true, c_{i+1}^A, \ldots, c_n^A) \neq p(c_1^A, \ldots, c_{i-1}^A, false, c_{i+1}^A, \ldots, c_n^A)
\]

for execution \( A \) and:

\[
p(c_1^B, \ldots, c_{i-1}^B, false, c_{i+1}^B, \ldots, c_n^B) \neq p(c_1^B, \ldots, c_{i-1}^B, true, c_{i+1}^B, \ldots, c_n^B)
\]

for execution \( B \). For each clause \( c_i \), this requirement can be encoded in two atomic labels: \((loc_p, l_{i,A})\) and \((loc_p, l_{i,B})\) with

\[
loc_p \equiv \text{the location of predicate } p,
\]

\[
l_{i,A} \equiv c_i \land (p(c_1, \ldots, c_{i-1}, true, c_{i+1}, \ldots, c_n))
\]

\[
\neq p(c_1, \ldots, c_{i-1}, false, c_{i+1}, \ldots, c_n)
\]

\[
l_{i,B} \equiv \neg c_i \land (p(c_1, \ldots, c_{i-1}, false, c_{i+1}, \ldots, c_n))
\]

\[
\neq p(c_1, \ldots, c_{i-1}, true, c_{i+1}, \ldots, c_n)
\]

Similarly, to encode GICC, for each clause \( c_i \) of a predicate \( p \) one needs to define four atomic labels: \((loc_p, l_{i,A}), (loc_p, l_{i,B}), (loc_p, l_{i,C})\) and \((loc_p, l_{i,D})\) with

\[
l_{i,A} \equiv c_i \land p(c_1, \ldots, c_n)
\]

\[
l_{i,B} \equiv \neg c_i \land p(c_1, \ldots, c_n)
\]

\[
l_{i,C} \equiv c_i \land \neg p(c_1, \ldots, c_n)
\]

\[
l_{i,D} \equiv \neg c_i \land \neg p(c_1, \ldots, c_n)
\]

For the next criterion, for a predicate \( p \), we consider the disjunctive normal form (DNF) of \( p \) and of its negation \( \neg p \): \( \text{dnf}(p) = \bigvee_i \text{imp}_i^{\text{dnf}(p)} \) and \( \text{dnf}(\neg p) = \bigvee_k \text{imp}_k^{\text{dnf}(\neg p)} \). (The formulas \( \text{imp}_i^{\text{dnf}(p)} \) and \( \text{imp}_k^{\text{dnf}(\neg p)} \) are called implicants of these DNFs.) To encode Implicant Coverage (IC), one needs to define, for each predicate \( p \):

- a label \((loc_p, \text{imp}_i^{\text{dnf}(p)})\) for any implicant \( \text{imp}_i^{\text{dnf}(p)} \) of the DNF of \( p \),

- a label \((loc_p, \text{imp}_k^{\text{dnf}(\neg p)})\) for any implicant \( \text{imp}_k^{\text{dnf}(\neg p)} \) of the DNF of \( \neg p \).

Finally, to encode Unique True Point Coverage (UTPC), one needs to define, for each predicate \( p \):

- a label \((loc_p, \text{imp}_i^{\text{dnfmin}(p)} \land \bigwedge_{j \neq i} \neg \text{imp}_j^{\text{dnfmin}(p)})\) for each implicant \( \text{imp}_i^{\text{dnfmin}(p)} \) of the minimal DNF\footnote{A DNF is minimal if (a) no implicant can be omitted, and (b) no subterm of an implicant can be omitted. We refer the reader to [25] for detailed definitions.} \( \text{dnf}_{\text{min}}(p) \) of \( p \).
a label $(\text{loc}_p, \text{imp}_{\text{dnf}_\text{min}}(\neg p) \land \bigwedge_{i\neq k} \neg \text{imp}_{\text{dnf}_\text{min}}(\neg p))$ for each implicant $\text{imp}_{\text{dnf}_\text{min}}(\neg p)$ of the minimal DNF $\text{dnf}_{\text{min}}(\neg p)$ of $\neg p$.

The reader can easily check that these labels precisely represent the required test objectives according to the definition of these criteria in [25].

**Mutation testing.** We now consider an even more involved coverage criterion, namely weak mutations, and show that a well-defined part of WM can be simulated by LC. We consider only atomic mutations operators that can affect either a left-hand side expression (lhs), an expression or a condition. This is a very generic model of mutations, encompassing all standard operators [25], as well as deletion of assignments, since the replacement of an assignment $x:=\text{exp}$ by $x:=x$ models its deletion. Finally, we restrict ourselves to mutation operators neither affecting nor introducing side-effect expressions (in particular, calls to side-effect prone functions). We refer to such operators as side-effect free mutation operators.

**Theorem 3.** For any finite set $O$ of side-effect free mutation operators, $\text{WM}_O$ can be simulated by LC.
Sketch of proof. We have to define a suitable labelling function. For simplicity, let us consider first a single mutation operator $op \in O$. The main idea is to introduce one label for each mutant $M$ created by $op$, so that covering the label is equivalent to distinguishing $M$ from $P$ once the modified location has been reached. This transformation is depicted in Figure 7. Let us consider a mutant $M$ differing from $P$ only at location $loc$. We consider three cases, depending on the modification introduced by $op$:

1. $lhs := expr$ becomes $lhs := expr'$: we add label $l \triangleq (loc, expr \neq expr')$. We must prove that $t \leadsto_P M$ iff $t \leadsto_{\psi(P)} l$. Note that $t \leadsto_{\psi(P)} l$ iff $t$ reaches $loc$ with an internal state such that $expr$ and $expr'$ evaluate to different values. This is equivalent to say that $P(t)$ and $M(t)$ are in different internal states right after $loc$, which corresponds by definition $t \leadsto_P M$.

2. if $(cond)$ becomes if $(cond')$: we add label $l \triangleq (loc, cond \oplus cond')$, where $\oplus$ is the xor-operator. We follow the same line of reasoning as in the previous case. The $\oplus$ operator ensures that $P(t)$ and $M(t)$ will not follow the same branching condition.

3. $lhs := expr$ becomes $lhs' := expr$: we add label $l \triangleq (loc, \alpha(lhs) \neq \alpha(lhs') \land (lhs \neq expr \lor lhs' \neq expr))$, where $\alpha(x)$ denotes the memory location ($\approx$ address) of $x$, not its value. For example, in C the memory location is given by the & operator. This case requires a little bit more explanation. In order to observe a difference between $P(t)$ and $M(t)$ right after the mutated location, we need first that $lhs'$ and $lhs$ refer to different memory locations (which is not always obvious in the case of aliasing expressions). Moreover, no difference can be observed if both locations had the assigned value, i.e. if the old value of $lhs$ and the old value of $lhs'$ were equal to $expr$ in $P(t)$ before the assignment. To observe the difference, at least one of them should be modified by the assignment. This is exactly what $l$ encodes.

By iterating this technique for all considered mutation operators $op \in O$, we obtain the desired labelling function. 

The subset of mutations we have been considering so far is limited to (1) atomic mutations and (2) side-effect free operators. The first restriction is not a major issue as atomic mutations have been observed to be almost as powerful as high-order mutations [58]. The second restriction has two sides: (2.a) it forbids mutation operators introducing side-effects, for example mapping $x$ to $x++$, and (2.b) it forbids to mutate a side-effect prone expression. Restriction (2.a) is not severe: it encompasses operators ABS, ROR, AOR, COR and UOI [25], which have been shown mostly equivalent to much larger sets of operators [60, 73]. It is left as an open question to quantify more precisely what is lost with restriction (2.b).

Black-box criteria. Assuming a partition of the input domain $D$ of program $P$ given as disjoint predicates $\varphi_1, \ldots, \varphi_k$, the IDC criterion requires one input
for each $\varphi_i$. The corresponding labelling function adds all labels of the form $(\text{loc}_0, \varphi_j)$, where $\text{loc}_0$ is the entry point of $P$. The approach is independent of the way the partition is obtained, covering both interface-based and functionality-based partitions [25, Chap. 4]. Figure 6 illustrates the approach with $k = 3$, $\varphi_1 \equiv (x == 0)$, $\varphi_2 \equiv (x > 0)$ and $\varphi_3 \equiv (x < 0)$.

4.3. Specifying other useful test objectives with labels

While test generation tools are often used to generate coverage adequate test suites, they have also been applied in other use cases, like e.g. runtime failure detection [21] or patch testing [20]. Test objectives corresponding to run-time failures such as those implicitly searched for in active testing or assertion-based testing [40, 48, 53] can be easily captured by labels, including division by zero, out-of-bound array accesses or null-pointer dereference. Typically, any error-prone instruction at location $\text{loc}$ with a precondition $\varphi_{\text{safe}}$ will be tagged by a label $(\text{loc}, \neg \varphi_{\text{safe}})$. Test objectives corresponding to reaching code zones affected by a patch can be encoded by labels with a true predicate at the entrance of each basic block modified by the patch.

4.4. Limitations of label expressiveness

The following classes of test objectives cannot be directly encoded through labels [19]:

- objectives constraining paths rather than program locations (e.g. data-flow or prime paths coverage criteria [25]),
- objectives relating different paths (e.g. $\text{MCDC}$ criteria other than $\text{GACC}$ and $\text{GICC}$, hyperproperties), possibly in slightly different programs (i.e. strong mutations).

While the first class of criteria can be encoded by labels with the help of additional instrumentation (see e.g. [18] for data-flow criteria), for the others no simple encoding has been found yet. As already pointed out, weak mutations with side-effect operators are also outside the direct scope of labels. When no exact simulation is known, labels can still be used for approximations. For example, even the most intricate $\text{MCDC}$ criterion (a.k.a. $\text{RACC}$) can be upper-approximated (with $\text{MCC}$) or lower-approximated (with $\text{GACC}$).

An extension of labels, named hyperlabels, is presented in Section 10.1. Hyperlabels [8, 9] provide combination operators over labels, yielding a very expressive framework for the specification of test objectives. While standard labels are restricted to state-reachability constraints (the test datum must reach a specific state), hyperlabels can express test objectives defined over trace reachability (the test datum must follow a particular sequence of states) or even hyper-reachability (constraints are here expressed over finite sets of traces, typically pairs).
5. Infeasible Label Detection

A significant proportion of the labels in an annotated program can turn out to be infeasible, i.e. no input can satisfy them. Infeasible labels are a threat to the efficiency of label-driven test generation, because a significant part of the test budget might be lost trying to cover them. In this section, we introduce a sound approach to prune out infeasible labels.

5.1. Definitions

We first formally define infeasible labels.

Definition 2. Given a label \( l \triangleq (\text{loc}, \varphi) \) in an annotated program \( ⟨P, L⟩ \), we say that \( l \) is infeasible if there is no input datum \( t \) such that \( t \leadsto_{⟨P, L⟩} l \).

Given a program \( P \), an assertion is a pair \((\text{loc}, \varphi)\) where \( \text{loc} \) is a location of \( P \) and \( \varphi \) is a predicate that contains only valid and side-effect free expressions.

Definition 3. An assertion of \( P \) is valid iff for any test datum \( t \) and for any internal state \( S \) of \( P \) such that \( t \) reaches \( \text{loc} \) with internal state \( S \), we have that \( \varphi \) evaluates to true in \( S \).

Definition 4. Given a label \( l \triangleq (\text{loc}, \varphi) \) in an annotated program \( ⟨P, L⟩ \), the opposite assertion of \( l \) is the assertion \((\text{loc}, \lnot \varphi)\).

5.2. Reducing label infeasibility to assertion validity

Lemma 4. A label is infeasible iff its opposite assertion is valid.

Sketch of proof. If a label \( l \triangleq (\text{loc}, \varphi) \) is infeasible, then for any test datum \( t \) of \( P \) that reaches \( \text{loc} \) with some internal state \( S \), we have that \( \varphi \) evaluates to false in \( S \). As a consequence, \( \lnot \varphi \) evaluates to true in \( S \). Thus, by definition, the opposite assertion of \( l \) is valid. The reverse can be proven by a similar reasoning. □

5.3. A practical approach to detecting infeasible labels

Assertion validity is a very common kind of safety properties and many assertion checker tools are available \cite{18, 27}, relying on various formal techniques such as weakest-precondition or value analysis, as well as model-checking. By Lemma 4, a natural approach to detect if a label is infeasible is to send its opposite assertion to an off-the-shelf assertion checker: if the checker is able to prove that the assertion is valid, then we know that the label is infeasible. As assertion checking is an undecidable and complex problem, the assertion checker may time out or return no answer, so that our infeasible label detection approach is only partial.
6. Label Coverage Measurement and Label-Driven Symbolic Execution

6.1. Naive approach

Given an annotated program \( \langle P, L \rangle \), we seek for automatic methods for: (1) computing the LC score of a given test suite \( TS \), and (2) deriving a test suite achieving high LC score. We propose first a black-box approach, reusing standard automatic testing tools through a direct instrumentation of \( P \). This technique underlies previous works aiming at extending DSE coverage abilities \([40, 48, 51, 53, 67, 68, 77]\). While it allows for cheap LC score computation, it is far from efficient for automated test generation (abbreviated as ATG below), mainly because of an exponential blow-up of the path space of the program.

6.1.1. Direct instrumentation

The direct instrumentation \( P' \) for \( \langle P, L \rangle \) consists in inserting for each label \( l \triangleq (\text{loc}, \varphi) \in L \) a new branching instruction \( I : \text{if } (\varphi) \{\} \); such that all instructions leading to \( \text{loc} \) in \( P \) lead to \( I \) in \( P' \), and \( I \) leads to \( \text{loc} \). The transformation is depicted in Figure 8. When different labels are attached to the same location, the new instructions are chained together in a sequence ultimately leading to \( \text{loc} \).

Let us denote by \( NTD \) the set of test objectives over \( P' \) requiring to cover all \textit{New Then} Decisions introduced by this transformation. Direct instrumentation is obviously sound w.r.t. LC in the following sense.

**Theorem 5** (Soundness). Given an annotated program \( \langle P, L \rangle \), its direct instrumentation \( P' \) and a test suite \( TS \), we have: \( TS \rightsquigarrow_{\langle P, L \rangle} LC \iff TS \rightsquigarrow_{P'} NTD. \)
```c
int numPos(int a, int b) {
    int n = 0;
    if (a > 0) {
        // l1: a == 5
        n++;
    }
    if (b > 0) n++;
    // l2: n > 0
    return n;
}
```

Figure 9: Function `numPos` with two labels

This is interesting for both LC score computation and ATG. Indeed, any ATG tool run on $P'$ will produce a test suite $TS$ covering LC for $\langle P, L \rangle$ as soon as $TS$ covers all branches of interest in $P'$. Besides, a slightly modified version of direct instrumentation, updating coverage information in the new then-branches, allows efficient coverage score computation.

**Theorem 6.** Given an annotated program $\langle P, L \rangle$, its direct instrumentation $P'$ and a test suite $TS$, then the LC score of $TS$ can be computed in time bounded by $|TS| \cdot \text{maxtime}(\{P'(t) | t \in TS\})$.

Note that by computing the maxtime over $P'$ in the above formula, we implicitly include the overhead of evaluating labels within the code. The expectation is that only a small fraction of the labels in the program are evaluated in any path, so that $\text{maxtime}(P')$ is substantially smaller than $|L| \times \text{maxtime}(P)$. It follows from Theorems 3 and 6 that coverage measurement of the side-effect free subset of weak mutation (WM) can then be efficient in practice: rather than re-running program execution for each mutant, we can measure a relatively small number of labels for a given test case.

6.1.2. Discussion

The direct instrumentation, while useful for label coverage measurement, is inefficient for ATG using symbolic execution, because of two main issues that we illustrate on the function `numPos` annotated with two labels as shown in Figure 9.

The paths of this program and those resulting from its direct instrumentation $P'$ are shown in Figure 10 (a) and (b). The first issue is that the initial program has only 4 paths, while the direct instrumentation leads to 12 more intricate paths (some of which being infeasible). In particular, all the feasible paths covering the second label contain a constraint coming from the condition of the first label, that is, the condition $a=5$ or its negation. Not present in the initial program logic, this constraint is irrelevant for covering subsequent branches and uselessly increases the size of the path constraint to be solved during DSE.

The second issue is that running DSE on $P'$ will lead to covering the second label five times since five feasible partial paths lead to the condition $n>0$ in $P'$,
and five test cases will be generated to cover it (e.g. \((a, b) = (5, 1), (5, 0), (1, 1), (1, 0)\) and \((0, 1)\)). In contrast, a minimal test suite covering the two labels is the single case \((a, b) = (5, 1)\).

To synthesize, the two issues of direct instrumentation that make it inefficient for ATG are:

(†) \(P'\) is too complex: it exhibits many more paths than \(P\), most of them being unduly complex for covering the labels we are targeting.

(‡) DSE will naturally produce test suites that cover the same labels several times, requiring substantially more analysis effort than necessary.

Let us formalize the first point (†) hereafter. We consider two dimensions in which \(P'\) is “too complex”: the size of the search space, denoted \(|\text{Paths}(P')|\), and the shape of the paths in \(\text{Paths}(P')\). Let us call label constraints the conditions of all additional branches \(\varphi\) and \(\neg\varphi\) introduced in \(P'\) compared to \(P\).

**Theorem 7** (Non-tightness). Given an annotated program \((P, L)\) and its direct instrumentation \(P'\), let us assume that \(\text{Paths}(P)\) is bounded, that \(k\) represents the maximal length of paths in \(\text{Paths}(P)\) and that \(m\) is the maximal number of labels per location in \(P\). Then:

- \(|\text{Paths}(P')|\) can be exponentially larger than \(|\text{Paths}(P)|\) by a factor \(2^{m \cdot k}\);
- any \(\sigma' \in \text{Paths}(P')\) may carry up to \(m \cdot k\) (positive or negative) label constraints.

*Sketch of proof.* A single path \(\sigma \in P\) may correspond to up to \(2^{m \cdot |\sigma|}\) paths in \(P'\), since each label of \(P\) creates a branching in \(P'\) and at most \(m\) such branchings can be found at each step of \(\sigma\). Note also that the paths \(\sigma' \in P'\) corresponding to \(\sigma \in P\) have length bounded by \(m \cdot |\sigma|\). Therefore they can pass through up to \(m \cdot |\sigma|\) label constraints, while (by definition) \(\sigma\) does not pass through any label constraint. \(\square\)

Both aspects are problematic for symbolic execution: more paths means more requests to a constraint solver, while more constraint-laden paths means more expensive requests.

**6.2. Efficient label-driven DSE**

We describe in this section two main ingredients in order to obtain efficient ATG for \(\text{LC}\): (1) a tight instrumentation avoiding all drawbacks of the direct instrumentation, and (2) a strong coupling of label coverage and DSE through iterative label deletion.
6.2.1. Tight instrumentation

Given a label \( l \equiv (\text{loc}, \varphi) \), the key insights behind the tight instrumentation are the following:

- label constraint \( \varphi \) is useful only for covering \( l \), and should not be propagated beyond that point;
- label constraint \( \neg \varphi \) is pointless w.r.t. covering \( l \), and should not be enforced in any way.

Keeping these lines in mind, the instrumentation works as depicted in Figure 11: for each label \( (\text{loc}, \varphi) \), we introduce a new instruction

\[
\text{if (nondet) }
\{ \text{assert(\varphi); exit} \}
\]

where \( \text{assert(\varphi)} \) requires \( \varphi \) to be verified, \( \text{exit} \) forces the execution to stop and \( \text{nondet} \) is a non-deterministic choice.

In the resulting instrumented program \( P^* \) (Figure 11 right column), when an execution reaches \( \text{loc} \), it gives rise to two execution paths: the first one tries to cover the label by asserting \( \varphi \) and \textit{stops right there}, the second one simply follows its execution as \textit{it would do in P}, neither \( \varphi \) nor \( \neg \varphi \) being enforced.

Note that any DSE engine can simulate non-deterministic choices by an additional input array of (symbolic) Boolean values.
Let us denote by $\text{NA}$ the set of test objectives over $P^*$ requiring to cover all New Assert introduced by the instrumentation with condition evaluating to true (cf. Figure 11). Alternatively, $\text{NA}$ comes down to cover all new instructions coming from the tight instrumentation, or to cover all the newly introduced exit(0). We also extend our definition of coverage for $\text{NA}$ as follows, as $P^*$ contains non-deterministic choices: $TS \rightsquigarrow P$. $\text{NA}$ if for each $na \in \text{NA}$, there is a test datum $t \in TS$ such that one of the possible executions of $P^*$ over $t$ covers $na$. Tight instrumentation is sound w.r.t. $\text{LC}$ in the following sense.

**Theorem 8** (Soundness). Given an annotated program $(P, L)$, its tight instrumentation $P^*$ and a test suite $TS$, we have: $TS \rightsquigarrow (P, L) \text{LC}$ iff $TS \rightsquigarrow P^* \text{NA}$.

**Sketch of proof.** We use the notation of Figure 11. Clearly, if (the execution of) a test datum $t$ can reach a newly introduced exit(0) in $P^*$ (preceded by instruction assert(p) which corresponds to location loc), then $t$ over $P$ reaches location loc with a memory state satisfying $p$, as the newly introduced instructions cannot modify memory states, they can just observe it. Conversely, if a test datum $t$ over $P$ reaches a location loc with a memory state satisfying predicate $p$, there is an execution in $P^*$ where $t$ reaches the corresponding newly introduced exit(0): indeed, it happens when all non-deterministic choices but the one covering the label choose to follow the original program execution, while the remaining one chooses to bifurcate over the assert(p) when the execution
reaches loc with an adequate memory state – which is ensured to happen, as it does over P.

In particular, an ATG tool run on P⋆ and producing a test suite TS covering NA will also cover the labels of the annotated program ⟨P, L⟩. The following result now follows from Theorems 5 and 8.

**Corollary 9.** Let ⟨P, L⟩ be an annotated program with direct instrumentation P′ and tight instrumentation P⋆. Then both DSE(P⋆) and DSE(P′) achieve the same coverage and can be used to cover all coverable labels in ⟨P, L⟩.

Interestingly, tight instrumentation does not show any of the issues reported in Theorem 7. The underlying reasons have been sketched at the beginning of Section 6.2.1 and are depicted in Figure 12. A single execution path in P going through n labels gives birth up to 2^n paths in P′ (left column), while it creates at most n + 1 paths in P⋆ (right column). Moreover, each path in P⋆ can go through at most one single positive label constraint (because of the exit node), while a path σ′ in P′ can carry up to |σ′| (positive or negative) label constraints.

These results are summarized in Theorem 10.

**Theorem 10 (Tightness).** Given an annotated program ⟨P, L⟩ and its instrumented version P⋆, let us assume that Paths(P) is bounded, that k is the maximal length of paths in Paths(P) and that m is the maximal number of labels per location in P. Then P⋆ is tight in the following sense:
\[|Paths(P^*)| \text{ is linear in } |Paths(P)| \text{ and } m \cdot k. \] More precisely, \[|Paths(P^*)| \text{ is bounded by } (m \cdot k + 1) \cdot |Paths(P)|; \]

- any \(\sigma \in Paths(P^*)\) carries at most one label constraint.

**Sketch of proof.** The main reasons behind this result directly follow from the tight instrumentation, and have already been exposed just before Theorem 10.

To illustrate the benefits of tight instrumentation on our running example, consider the paths of \(P\), \(P'\) and \(P^*\) shown in Figure 10 (a), (b) and (c). Instead of 12 paths of \(P'\), the tight instrumentation \(P^*\) contains only 9 paths, that is, one additional path for each partial path reaching a label in \(P\). Unlike in direct instrumentation, neither of these paths uselessly keeps any constraint for the first label: either the label constraint is the last one in the path and is necessary to cover the label, or it is bypassed without complicating the constraint set for covering subsequent branches.

Theorems 7 and 10 imply that any path-based program analysis like DSE conducted over \(P^*\) will have a much easier task than if conducted over \(P'\), since \(P^*\) contains exponentially fewer paths and those paths are simpler. The issue (†) identified in Section 6.1.2 is thus avoided for tight instrumentation.

### 6.2.2. Iterative label deletion

We focus now on issue (‡) pointed out in Section 6.1.2. A DSE procedure launched on \(P^*\) tries to cover all paths from \(P^*\), while we are only interested in covering branches corresponding to labels. Especially, it may try to cover (partial) paths ending in an already-covered \(\text{assert}(\phi)\). Whether they fail or not, these computations are redundant since Theorem 8 only requires each new \(\text{assert}\) to be covered once.

For our running example, DSE on the tight instrumentation \(P^*\) will still try to cover the second label four times since four feasible partial paths lead to the condition \(n>0\) in \(P^*\) (cf. Figure 10 (c)), and will generate three test data to cover it (e.g. \((a,b) = (1,1), (1,0)\) and \((0,1)\)).

Iterative deletion of labels (IDL) aims at taming issue (‡) by (conceptually) erasing label constraints as soon as they are covered, so that that does not affect the rest of the path search. In practice, this can be implemented by making our program instrumentation keep track of the coverage state of each label during the path exploration by the DSE tool. We depict in Figure 13 how our tight instrumentation can be modified in such a way. In a nutshell, each label \(l\) is associated with a boolean flag \(\text{covered}_l\) in an external database called \(\text{label\_state\_db}\), and this flag should be true iff \(l\) has already been covered during path exploration. In addition, the value of \(\text{covered}_l\) in the database can be read or set by using the \(\text{is\_covered(label\_state\_db,l)}\) and \(\text{set\_covered(label\_state\_db,l,value)}\) primitives. Before the path exploration starts, flag \(\text{covered}_l\) is initialized to false. As a consequence, the call to \(\text{is\_covered(label\_state\_db,l)}\) guarding the
access to label \( l \) in the modified tight instrumentation will always return false and the DSE tool will be allowed to explore paths leading to \( l \) being covered. Yet, as such paths end by a call to \( \text{set\_covered(label\_state\_db, l, true)} \), as soon as one of them has been explored by the DSE tool, all the subsequent calls to \( \text{is\_covered(label\_state\_db, l)} \) guarding the access to the label \( l \) will always return true and the DSE tool will thus be barred from exploring any additional path leading to \( l \) being covered. This behaviour corresponds to the label being erased from the program after being covered by the DSE tool.

We denote by \( \text{DSE}^\star(P^\star) \) this combination of tight instrumentation and \( \text{IDL} \).

Considering only deterministic DSE techniques, the following result holds.

**Theorem 11** (Relative completeness). Let \( \langle P, L \rangle \) be an annotated program, and \( P^\star \) its tight instrumentation. Then \( \text{DSE}^\star(P^\star) \) covers as many labels as \( \text{DSE}(P^\star) \) does.

**Sketch of proof.** The main argument here is that discarding a path \( \sigma \) passing through an already covered label \( l \) in \( \text{DSE}^\star(P^\star) \) can never prevent from covering another not-yet-covered label \( l' \), as by construction there always exist a “label free” path \( \sigma' \) – similar to \( \sigma \) but without label constraints – passing through \( l' \).

7. Implementation: the LTest Toolset

Putting Sections 4, 5 and 6 together, we see that labels form the basis of a very powerful framework for automated testing, providing various services and handling many different flavours of test objectives in a complete and uniform fashion. Figure 14 gives an overview of this framework. Starting from a program \( P \) and choosing a way \( C \) to derive test objectives (like a coverage

---

\(^5\)A reader familiar with DSE may point out that the DSE tool will consider the label state database as another input of the program and try to generate values for the label flags as well. However, this undesired behaviour can be easily avoided in most DSE tools by making the state database a purely concrete value.
The labelling function $\psi_C$ creates the $C$-compliant label-annotated program $\langle P, L \rangle$. In particular, predefined labelling functions are available for all common coverage criteria. Assertion checkers can be leveraged on $\langle P, L \rangle$ in order to detect infeasible (or uncoverable) labels. Finally, efficient LC score computation can be performed to measure the coverage of a given test suite, and efficient label-driven DSE can be used to cover feasible labels.

7.1. Overview of the LTEST platform

We have implemented our framework in the context of C programs within the LTEST toolset [26], which offers the following services:

**Program annotation:** the LANNOTATE module annotates the program with labels according to a chosen coverage criteria.

**Uncoverable label detection:** the LUNCOV module leverages off-the-shelf assertion checkers to detect infeasible labels. The information is primarily used by other modules, but it can also be exported for external use.

**Coverage estimation:** the LREPLAY module replays a given test suite and reports its label coverage. Coverage is given as a whole (all test objectives taken into account) and per criterion. Moreover, infeasible and uncovered labels are reported.

**ATG:** the LGENTEST module uses DSE*(P*) to build a test suite covering the labels. In case a test suite has already been replayed, LGENTEST will try to complete the achieved coverage rather than starting from scratch.
The platform currently provides automatic label annotation for the following coverage criteria: decision coverage (DC), function coverage (FC), condition coverage (CC), multiple-condition coverage (MCC), weak mutation (WM, operators AOR, ROR, COR, ABS), interface-based input domain partition (IDC) and general active clause coverage (GACC). Moreover, coverage criteria can be combined together, test objectives can be restricted to certain procedures of the program under test and it is possible to add hand-written test objectives.

7.2. Technical details

The four modules LAnnotate, LReplay, LUncov and LGenTest interact through shared information comprising the annotated program and a database mapping each label to its current status, namely: covered, uncoverable, unknown (i.e. neither covered nor proven uncoverable). LAnnotate acts as a front-end and comes with predefined labelling functions for all the criteria supported by the platform. It annotates the program with the corresponding labels according to the criteria chosen by the user and creates the status database. An API is provided for developers to easily write labelling functions for additional criteria. Labels are inserted in the C code as external function calls and can be suppressed on-demand by the C compiler via a simple C macro.

Manual edition of labels in the source code and status database is made easy if needed. The three other modules provide user-level services. They can update label statuses and take advantage of them.

The LTest toolkit is built on top of the Frama-C verification platform for C programs [42] whose main analyzers are open source (LGPL). We took advantage of the plugin-based architecture of Frama-C, reusing existing analyses of interest for our needs. In particular, LUncov takes advantage of the assertion checking capabilities of Frama-C to detect infeasible labels. These capabilities are based on the weakest-precondition and value analysis approaches. The Frama-C kernel, LAnnotate, LGendTest and LUncov modules are written in OCaml. The ATG engine of LGendTest relies on an implementation of DSE* within the DSE tool PathCrawler [74], written in ECLiPSe/Prolog.

The LTest code (except for the LGenTest module) is open-source (LGPL).

8. Experimental Evaluation

8.1. Evaluating label-driven DSE for basic coverage criteria

Objectives. This first batch of experiments have been conducted in order to investigate the following properties of our label-driven DSE approach over basic common coverage criteria:

- the relative gain of our optimizations (cf. Sections 6.2.1 and 6.2.2) w.r.t. direct instrumentation,

---

6LTest source code is available online at: https://sites.google.com/view/michaelmarcozzi/software/frama-cltest
<table>
<thead>
<tr>
<th>trityp</th>
<th>50 loc</th>
<th>Random (witness)</th>
<th>DSE((P)) (witness)</th>
<th>DSE((P'))</th>
<th>DSE((P''))</th>
<th>DSE(*) ((P''))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>time</td>
<td>#paths</td>
<td>time</td>
<td>#paths</td>
<td>time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cover</td>
<td></td>
<td>cover</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.500</td>
<td></td>
<td>1.95</td>
<td>183</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.6s</td>
<td>1.3s</td>
<td>1.6s</td>
<td>2.0s</td>
<td>1.6s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>24/24</td>
<td>24/24</td>
<td>24/24</td>
<td>24/24</td>
</tr>
<tr>
<td></td>
<td>MCC</td>
<td>1.619</td>
<td>35</td>
<td>337</td>
<td>110</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.1s</td>
<td>1.3s</td>
<td>1.9s</td>
<td>3.0s</td>
<td>2.1s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16/28</td>
<td>17/28</td>
<td>28/28</td>
<td>28/28</td>
<td>28/28</td>
</tr>
<tr>
<td></td>
<td>WM</td>
<td>4.301</td>
<td>35</td>
<td>x</td>
<td>506</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.1s</td>
<td>1.3s</td>
<td>x</td>
<td>12s</td>
<td>5.1s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>61/129</td>
<td>x</td>
<td>125/129</td>
<td>125/129</td>
<td></td>
</tr>
<tr>
<td>4balls</td>
<td>WM</td>
<td>1.903</td>
<td>7</td>
<td>195</td>
<td>75</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.1s</td>
<td>1.2s</td>
<td>1.9s</td>
<td>1.1s</td>
<td>2.1s</td>
</tr>
<tr>
<td>utf8-3</td>
<td>WM</td>
<td>2.551</td>
<td>134</td>
<td>1,111</td>
<td>3,239</td>
<td>743</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.8s</td>
<td>1.4s</td>
<td>4.2s</td>
<td>4.3s</td>
<td>3.8s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>52/84</td>
<td>53/84</td>
<td>55/84</td>
<td>55/84</td>
<td></td>
</tr>
<tr>
<td>utf8-5</td>
<td>WM</td>
<td>5.396</td>
<td>680</td>
<td>81,133</td>
<td>14,676</td>
<td>3,265</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.1s</td>
<td>2s</td>
<td>576s</td>
<td>110s</td>
<td>35s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>53/84</td>
<td>80/84</td>
<td>82/84</td>
<td>82/84</td>
<td>82/84</td>
</tr>
<tr>
<td>utf8-7</td>
<td>WM</td>
<td>23,053</td>
<td>3,069</td>
<td>81,133</td>
<td>14,676</td>
<td>3,265</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35s</td>
<td>5.8s</td>
<td>576s</td>
<td>110s</td>
<td>35s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>53/84</td>
<td>80/84</td>
<td>82/84</td>
<td>82/84</td>
<td>82/84</td>
</tr>
<tr>
<td>tcas</td>
<td>CC</td>
<td>3.096</td>
<td>2,787</td>
<td>3,508</td>
<td>3,508</td>
<td>2,815</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.4s</td>
<td>2.9s</td>
<td>3.6s</td>
<td>4.9s</td>
<td>3.4s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>MCC</td>
<td></td>
<td>3.182</td>
<td>2,787</td>
<td>3,988</td>
<td>3,988</td>
<td>3,059</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.9s</td>
<td>2.9s</td>
<td>4.2s</td>
<td>5.2s</td>
<td>3.9s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11/12</td>
<td>10/12</td>
<td>11/12</td>
<td>11/12</td>
<td>11/12</td>
</tr>
<tr>
<td>tcas'</td>
<td>WM</td>
<td>20,347</td>
<td>4,420</td>
<td>300,213</td>
<td>20,312</td>
<td>6,014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27s</td>
<td>5.6s</td>
<td>602s</td>
<td>101/111</td>
<td>27s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90/111</td>
<td>88/111</td>
<td>101/111</td>
<td>101/111</td>
<td></td>
</tr>
<tr>
<td>replace</td>
<td>WM</td>
<td>11,747</td>
<td>866</td>
<td>87,498</td>
<td>6,420</td>
<td>2,347</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14s</td>
<td>2s</td>
<td>245s</td>
<td>64s</td>
<td>14s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>68/79</td>
<td>69/79</td>
<td>70/79</td>
<td>70/79</td>
<td>70/79</td>
</tr>
<tr>
<td>full_bad</td>
<td>WM</td>
<td>5,180</td>
<td>2,563</td>
<td>19,330</td>
<td>10,610</td>
<td>5,414</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7s</td>
<td>5s</td>
<td>15s</td>
<td>40s</td>
<td>19s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9/16</td>
<td>12/16</td>
<td>12/16</td>
<td>12/16</td>
<td></td>
</tr>
<tr>
<td>MCC</td>
<td></td>
<td>9,353</td>
<td>2,563</td>
<td>12,360</td>
<td>12,296</td>
<td>7,043</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19s</td>
<td>5s</td>
<td>19s</td>
<td>32s</td>
<td>19s</td>
</tr>
<tr>
<td>WM</td>
<td></td>
<td>9,425</td>
<td>2,563</td>
<td>19,330</td>
<td>10,610</td>
<td>5,414</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19s</td>
<td>5s</td>
<td>15s</td>
<td>40s</td>
<td>19s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29/46</td>
<td>34/46</td>
<td>34/46</td>
<td>34/46</td>
<td>34/46</td>
</tr>
</tbody>
</table>

TO: time-out (5,400s)  x: crash due to a bug in the underlying solver

Table 1: Experiments (1/2) for DSE* vs DSE and random testing
<table>
<thead>
<tr>
<th>get_tag-5</th>
<th>get_tag-6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CC</strong> 240 loc</td>
<td><strong>CC</strong> 240 loc</td>
</tr>
<tr>
<td>Random (witness)</td>
<td>DSE(^{\star}) (witness)</td>
</tr>
<tr>
<td>#paths</td>
<td>time</td>
</tr>
<tr>
<td>56,468</td>
<td>64s</td>
</tr>
<tr>
<td>42,380</td>
<td>48s</td>
</tr>
<tr>
<td>35,935</td>
<td>51s</td>
</tr>
<tr>
<td>1,305,971</td>
<td>1,512s</td>
</tr>
<tr>
<td>1,353,672</td>
<td>1,481s</td>
</tr>
<tr>
<td>1,287,924</td>
<td>1,463s</td>
</tr>
<tr>
<td>53,330</td>
<td>91s</td>
</tr>
<tr>
<td>68,691</td>
<td>80s</td>
</tr>
<tr>
<td>74,999</td>
<td>94s</td>
</tr>
<tr>
<td>2,785,951</td>
<td>2,945s</td>
</tr>
<tr>
<td>3,247,423</td>
<td>3,447s</td>
</tr>
<tr>
<td>2,064,394</td>
<td>2,232s</td>
</tr>
</tbody>
</table>

TO: time-out (5,400s)  x: crash due to a bug in the underlying solver

Table 2: Experiments (2/2) for DSE* vs DSE and random testing
the overhead of lifting DSE to LC and the gain in terms of coverage,

- the gain in terms of coverage over random testing.

**Protocol.** We consider 12 standard benchmark programs\[1\] taken from related works \[40, 67, 65\], mainly coming from the Siemens test suite (tcas), the Verisec benchmark (get_tag and fullbad from Apache source code) and MediaBench (gd from libgd). We also consider three classes of labels simulating standard coverage criteria of increasing difficulty: CC, MCC and WM (cf. Section 4.2). For WM, we use the operators AOIU, AOR, COR and ROR \[25\] (in a similar way to what is done by MuJava in a Java context \[55\]), which are considered very powerful in practice \[60, 73\]. In the end, we consider a total of 26 pairs of programs and coverage criteria, discarding the experiments with CC and MCC for programs without decision predicates involving more than one atomic condition (in this case CC and MCC indeed come down to DC).

We compare the following testing techniques (cf. the columns of Tables 1 and 2): random testing (witness) denoted Random, standard DSE (witness) denoted DSE(P), standard DSE on direct instrumentation denoted DSE(P'), standard DSE on tight instrumentation denoted DSE*(P*), and DSE with iterative label deletion run on tight instrumentation denoted DSE*(P*)). The DSE engine runs in deterministic mode, generating the same concrete values from one run to the other. All DSE variants are stopped either when a chosen time-out value has been reached or when all the DSE-reachable paths within a chosen upper bound in the total number of loop iterations have been covered. As for the variants of DSE, random testing time includes both test generation and test execution steps. For each benchmark program, random testing is allocated the same time as consumed by DSE*(P*). To avoid non-representative results, random testing is repeated at least 5 times for each example, and even more if necessary until the average number of generated tests is measured with a relative standard error less than 5%. Time-out for the solver is set to 1 min, time-out for test generation is set to 90 min. Experiments are performed on an Intel Core2 Duo 2.40 GHz, 4 GB of RAM.

For DSE derivatives, we record the following information: number of paths explored by the search, computation time and achieved coverage. The number of paths is a good measure for comparing the complexity of the different search spaces, and therefore to assess both the “cost” of lifting DSE to labels and the benefits of our optimizations. Coverage score together with computation time indicate how practical label-based DSE is. For random testing, we record the allocated computation time, the average number of tests generated and executed (indicated in the “#paths” rows of Tables 1 and 2), and the average coverage score.

**Results.** Results for our 26 examples are presented in Tables 1 and 2, and summaries on overhead and achieved coverage can be found in Tables 3 and 4.

\[http://micdel.fr/public/ltest/benchmarks-TAP2014.tar.gz\]
Table 3: Overhead (slow-down) with respect to DSE

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>DSE(P)</th>
<th>DSE(P')</th>
<th>DSE*(P)</th>
<th>DSE*(P')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>37%</td>
<td>61%</td>
<td>62%</td>
<td>62%</td>
<td>62%</td>
</tr>
<tr>
<td>Median</td>
<td>63%</td>
<td>90%</td>
<td>92%</td>
<td>93%</td>
<td>95%</td>
</tr>
<tr>
<td>Max</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Mean</td>
<td>70%</td>
<td>87%</td>
<td>88%</td>
<td>88%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Time-outs are excluded from the coverage computation.

Table 4: Label coverage ratios

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>DSE(P)</th>
<th>DSE(P')</th>
<th>DSE*(P)</th>
<th>DSE*(P')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td></td>
<td>×1.02</td>
<td>×0.92</td>
<td>×0.49</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>×1.79</td>
<td>×2.55</td>
<td>×1.37</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td>×122.50</td>
<td>×105.88</td>
<td>×7.15</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>×20.29</td>
<td>×10.50</td>
<td>×2.15</td>
<td></td>
</tr>
</tbody>
</table>

*Overheads take into account time-outs, counted as 5400s (90min).

First, note that when no time-out occurs, direct instrumentation and both variants of tight instrumentation achieve the same coverage, and that this coverage is high (>90% on 18/26 examples). We also observe that direct instrumentation yields a significant overhead, confirming previous work [51]: DSE(P') has four time-outs (TO) while DSE(P) has none, the time-overhead goes up to 122x (excluding TO), and the blow-up of the path-space reaches 50x.

On the other hand, tight instrumentation DSE*(P') yields only a very reasonable overhead w.r.t. standard DSE: no time-out is reported, the time-overhead is kept under 7x with an average of 2.15x, and the path-space growth is limited to 3x. On some examples, tight instrumentation performs remarkably better than direct instrumentation (94s vs TO on gd-5-WM). For programs where the execution time is not meaningful enough or very close for various techniques, the number of explored paths still clearly illustrates the benefits. Interestingly, DSE*(P') does perform better than standard DSE (up to 2x) on a few examples with very few additional paths (e.g. for get_tag_6 or gd-6).

We conjecture that additional label constraints may sometimes greatly simplify the solving process by introducing pertinent case considerations, but this point must be investigated further.

From a coverage point of view, DSE*(P*) performs better than standard DSE on 10/26 examples, with an increase in coverage from 3% up to 39% (trityp-MCC). Any increase in coverage coming at a reasonable cost is welcome, since hard-to-cover test objectives are considered more likely to detect bugs [17]. Recall also that DSE*(P*) offers relative completeness guarantees regarding label coverage, while standard DSE does not.

Finally, we observe that random testing obtains very unstable label coverage (varying between 44% to 100% of the coverable labels, with the average of 70%). Yet, it remains significantly lower than the DSE*(P*) coverage (91% in average).

**Conclusion.** These experiments confirm our formal predictions:
• DSE\(^\star\)(\(P^\star\)) performs dramatically better on difficult programs than the direct instrumentation, both in terms of search space and computation time;

• the overhead of DSE\(^\star\)(\(P^\star\)) w.r.t. standard DSE turns out to be always acceptable and often very low\(^3\) while label coverage is indeed increased in many cases—sometimes very significantly;

• finally, DSE\(^\star\)(\(P^\star\)) achieves considerably better coverage than random testing for the same time budget.

These results suggest that DSE can be efficiently lifted to LC coverage thanks to our optimizations.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Random} & \text{DSE}(P) & \text{DSE}^\star\(P^\star\) & \text{INF} \\
\hline
\text{Min.} & 37\% & 61\% & 62\% & 80\% \\
\text{Med.} & 63\% & 90\% & 95\% & 100\% \\
\text{Max.} & 100\% & 100\% & 100\% & 100\% \\
\text{Mean} & 70\% & 87\% & 90\% & 96\% \\
\hline
\end{array}
\]

For DSE\(^\star\)(\(P^\star\)-INF), ratios take into account the detected infeasible labels. Time-outs are excluded from the coverage computation.

Table 5: Label coverage ratios

8.2. Evaluating the impact of infeasible label detection

**Objectives.** Our framework enables detecting infeasible labels. In this section, we aim at evaluating how efficient this detection can be. We also measure the impact of infeasible label detection over the efficiency of our label-driven test generation. Indeed, if some infeasible labels have been pruned before instrumentation and test generation, DSE\(^\star\)(\(P^\star\)) can be stopped as soon as it has attempted covering all the remaining (presumably feasible) labels, instead of wasting time trying to cover the infeasible labels as well.

**Protocol.** We consider the same annotated benchmark programs and the same three coverage criteria (\(CC\), \(MCC\) and \(WM\)) as in the previous section.

We compare the following variants of symbolic execution (cf. Tables 6 and 7): the DSE\(^\star\)(\(P^\star\)) technique presented so far\(^4\) and DSE\(^\star\)(\(P^\star\)-INF) that exploits in addition the infeasible labels detected through a preliminary pass with our L\textsc{Uncov} module. The computation time of DSE\(^\star\)(\(P^\star\)-INF) does not include the computation time of L\textsc{Uncov}, whose results are reported in a separate column. Thanks to the knowledge of at least some infeasible labels, label coverage for

---

\(^3\)The overhead is even further reduced with the optimizations presented in Section 8.2

\(^4\)with the additional ability to stop when all the labels have been covered, instead of terminating only when all the DSE-reachable paths within a chosen upper bound for the total number of loop iterations have been explored.
<table>
<thead>
<tr>
<th>loc</th>
<th>#paths</th>
<th>time</th>
<th>cover</th>
<th>MW</th>
<th>loc</th>
<th>#paths</th>
<th>time</th>
<th>cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>7</td>
<td>2.1s</td>
<td>23</td>
<td>35</td>
<td>35</td>
<td>2.1s</td>
<td>23</td>
<td>35</td>
</tr>
<tr>
<td>108</td>
<td>100</td>
<td>3.4s</td>
<td>2.9s</td>
<td>35</td>
<td>101</td>
<td>3.4s</td>
<td>2.9s</td>
<td>35</td>
</tr>
<tr>
<td>108</td>
<td>102</td>
<td>3.9s</td>
<td>2.9s</td>
<td>102</td>
<td>102</td>
<td>3.9s</td>
<td>2.9s</td>
<td>102</td>
</tr>
<tr>
<td>108</td>
<td>103</td>
<td>2.7s</td>
<td>1.05s</td>
<td>103</td>
<td>103</td>
<td>2.7s</td>
<td>1.05s</td>
<td>103</td>
</tr>
<tr>
<td>219</td>
<td>161</td>
<td>2.56</td>
<td>5.76</td>
<td>161</td>
<td>161</td>
<td>2.56</td>
<td>5.76</td>
<td>161</td>
</tr>
<tr>
<td>121</td>
<td>121</td>
<td>2.56</td>
<td>5.76</td>
<td>121</td>
<td>121</td>
<td>2.56</td>
<td>5.76</td>
<td>121</td>
</tr>
<tr>
<td>111</td>
<td>111</td>
<td>2.56</td>
<td>5.76</td>
<td>111</td>
<td>111</td>
<td>2.56</td>
<td>5.76</td>
<td>111</td>
</tr>
<tr>
<td>50</td>
<td>24</td>
<td>1.6s</td>
<td>24</td>
<td>50</td>
<td>24</td>
<td>1.6s</td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>1.3s</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>1.3s</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>129</td>
<td>129</td>
<td>5.1s</td>
<td>1.3s</td>
<td>129</td>
<td>129</td>
<td>5.1s</td>
<td>1.3s</td>
<td>129</td>
</tr>
<tr>
<td>61/129</td>
<td>61/129</td>
<td>5.1s</td>
<td>1.3s</td>
<td>61</td>
<td>129</td>
<td>5.1s</td>
<td>1.3s</td>
<td>61/129</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2.3s</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2.3s</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 6: Experiments (1/2) for DSE with infeasible label detection
<table>
<thead>
<tr>
<th>loc: number of lines of code</th>
<th>f: number of labels</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Table 7: Experiments (2/2) for DSE* with infeasible label detection</th>
<th>Random (witness)</th>
<th>DSE(P) (witness)</th>
<th>DSE*(P*)</th>
<th>Uncov. detect.</th>
<th>INF</th>
</tr>
</thead>
<tbody>
<tr>
<td>get_2ag-5 240 loc</td>
<td>CC 20/1</td>
<td>#paths</td>
<td>56.468</td>
<td>11.833</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>64s</td>
<td>60s</td>
<td>20/20</td>
<td>20/20</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCC 26/1</td>
<td>#paths</td>
<td>42.380</td>
<td>11.833</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>48s</td>
<td>60s</td>
<td>26/26</td>
<td>26/26</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WM 47/1</td>
<td>#paths</td>
<td>35.935</td>
<td>11.833</td>
<td>11.856</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>51s</td>
<td>60s</td>
<td>43/47</td>
<td>43/47</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>get_2ag-6 240 loc</td>
<td>CC 20/1</td>
<td>#paths</td>
<td>1,385,571</td>
<td>76.456</td>
<td>306</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>1,512s</td>
<td>20/20</td>
<td>20/20</td>
<td>20/20</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCC 26/1</td>
<td>#paths</td>
<td>1,353,672</td>
<td>76.456</td>
<td>310</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>1,481s</td>
<td>20/20</td>
<td>20/20</td>
<td>20/20</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WM 47/1</td>
<td>#paths</td>
<td>1,287,924</td>
<td>76.456</td>
<td>76.48</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>1,463s</td>
<td>20/20</td>
<td>20/20</td>
<td>20/20</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>16/36</td>
<td>51s</td>
<td>34s</td>
<td>34s</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCC 36/1</td>
<td>#paths</td>
<td>68.891</td>
<td>14.516</td>
<td>15.20</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>14/36</td>
<td>51s</td>
<td>34s</td>
<td>34s</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WM 63/1</td>
<td>#paths</td>
<td>74.999</td>
<td>14.516</td>
<td>14.60</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>74/36</td>
<td>51s</td>
<td>34s</td>
<td>34s</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gd-6 319 loc</td>
<td>CC 36/1</td>
<td>#paths</td>
<td>2,785,551</td>
<td>107.410</td>
<td>37.40</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>23/36</td>
<td>17.40</td>
<td>36s</td>
<td>36s</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCC 36/1</td>
<td>#paths</td>
<td>3,247,423</td>
<td>107.410</td>
<td>111.20</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>23/36</td>
<td>3.44s</td>
<td>36s</td>
<td>36s</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WM 63/1</td>
<td>#paths</td>
<td>1,204,394</td>
<td>107.410</td>
<td>77.79</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>23/36</td>
<td>2.17s</td>
<td>36s</td>
<td>36s</td>
</tr>
<tr>
<td></td>
<td>cover</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Experiments (2/2) for DSE* with infeasible label detection
DSE\(^*\)\((P^*)\)-INF is computed w.r.t. the set of potentially coverable labels, i.e. labels detected as infeasible are discarded.

**Results.** A total of 84 infeasible labels were identified, spanning over 13/26 programs and topping at 35\% of the total number of labels in some examples. This yields a substantial improvement of the reported coverage ratios (see Table 5). In 6 out of 26 cases, the coverage ratios do reach 100\% of the actually feasible objectives.

When infeasible objectives are detected in the tested program, pruning out those infeasible objectives made DSE\(^*\) in average 2.18x faster, the speedup topping at 11.66x. Results on the whole benchmark are more mitigated (cf. Tables 6 and 7). Yet, note that the detection of infeasible labels takes an acceptable amount of time w.r.t. the test generation time on our benchmark programs (12\% of the total computation time in average), and induces almost no slow-down for the bigger examples (less than 3\% when test generation takes more than 10s).

**Conclusion.** The experiments confirm the ability of our framework to detect infeasible labels efficiently and the interest of doing so to speed up test generation.

8.3. Experiments with a complex coverage criterion

**Objectives.** We want to check if the results of Sections 8.1 and 8.2 still hold for GACC, which is a more complex criterion, known as challenging for test automation tools [25].

**Protocol and results.** We reuse the same protocols and benchmark programs that those in Sections 8.1 and 8.2. Yet, we restrict ourselves to the 6 programs with decision predicates involving more than one atomic condition. Results are reported in Tables 8, 9, and 10. DSE\(^*\)\((P^*)\) achieves good coverage results (>90\% on 4/6 examples, average 85\%), significantly better than random testing and DSE. Yet, the overhead is significantly more important than for the other criteria (max. of 60x vs 7x, average of 10.55x vs 2.15x). Compared with direct instrumentation, DSE\(^*\)\((P^*)\) does avoid a time-out and is twice as fast on average (excluding TO). Finally, infeasible label detection allows to recover a reasonable overhead (max 3.14x, average 1.85x) and to detect 37 uncoverable labels out of 91 uncovered ones.

**Conclusion.** Experiments confirm that GACC is a challenging criterion for automatic tools. Yet, while the overhead of DSE\(^*\)\((P^*)\) is more important than it is with other criteria, it is still affordable in most cases, and the technique achieves very good coverage—much better than random testing or standard DSE. Infeasible label detection keeps overhead low. Surprisingly, MCC seems significantly easier to cover on these examples than GACC, while it is more powerful. We conjecture that our GACC encoding may lead to too complex formula to solve. Though, this encoding is very succinct (linear) in the number of atomic conditions while the MCC encoding is exponential, which could make a huge difference on very complex branching conditions.
Table 8: Results for GACC

<table>
<thead>
<tr>
<th></th>
<th>Random (witness)</th>
<th>DSE($P$) (witness)</th>
<th>DSE($P$)</th>
<th>DSE($P$)</th>
<th>Uncov. detect.</th>
<th>INF</th>
</tr>
</thead>
<tbody>
<tr>
<td>trityp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34fl</td>
<td>1.438</td>
<td>2.3s</td>
<td>34/34</td>
<td>2.3s</td>
<td>34/34</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>16/34</td>
<td></td>
<td></td>
<td>34/34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tcas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46l</td>
<td>6.319</td>
<td>2.787</td>
<td>10.168</td>
<td>4.234</td>
<td>1.201</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46/46</td>
<td></td>
<td></td>
<td>46/46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fullbad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34l</td>
<td>17.562</td>
<td>2.563</td>
<td>10.086</td>
<td>5.619</td>
<td>3.041</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22/39</td>
<td></td>
<td></td>
<td>23/39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>get_tag</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>94l</td>
<td>33,877</td>
<td>7.456</td>
<td>116/121</td>
<td>8,077</td>
<td>7.742</td>
<td></td>
</tr>
<tr>
<td></td>
<td>59/94</td>
<td></td>
<td></td>
<td>60/94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>get_tag</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>94l</td>
<td>1,123,829</td>
<td>47.216</td>
<td>753,409</td>
<td>50,957</td>
<td>48,819</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60/94</td>
<td></td>
<td></td>
<td>61/94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g5-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76l</td>
<td>45,876</td>
<td>14.516</td>
<td>52,419</td>
<td>15,459</td>
<td>15,401</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40/76</td>
<td></td>
<td></td>
<td>66/76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g5-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76l</td>
<td>3,723,361</td>
<td>107.419</td>
<td>TO</td>
<td>113,944</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48/76</td>
<td></td>
<td></td>
<td>66/76</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Overhead (slow-down) w.r.t. DSE for GACC

<table>
<thead>
<tr>
<th></th>
<th>DSE($P$)</th>
<th>DSE($P$)</th>
<th>INF</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanilla</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>1.44x</td>
<td>1.41x</td>
<td>1.38x</td>
</tr>
<tr>
<td>Med</td>
<td>3.76x</td>
<td>1.81x</td>
<td>1.44x</td>
</tr>
<tr>
<td>Max</td>
<td>130.79x</td>
<td>59.40x</td>
<td>3.14x</td>
</tr>
<tr>
<td>Mean</td>
<td>21.99x</td>
<td>10.55x</td>
<td>1.85x</td>
</tr>
</tbody>
</table>

Time-outs* 1 0 0

* For DSE($P$), a time-out counts as a 5400s (90min).

Table 10: Label coverage ratios for GACC

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>INF</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanilla</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>47%</td>
<td>62%</td>
<td>64%</td>
</tr>
<tr>
<td>Med</td>
<td>55%</td>
<td>76%</td>
<td>88%</td>
</tr>
<tr>
<td>Max</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Mean</td>
<td>60%</td>
<td>78%</td>
<td>85%</td>
</tr>
</tbody>
</table>

For DSE($P$)-INF, ratios take into account the detected infeasible labels.
Time-outs are excluded from the coverage computation.
Overall, taking into account all the coverage criteria considered in Section 8.1 plus \textbf{GACC}, overhead is kept small (no time-out, average 3.93x for DSE\(^*(P^*)\), 1.57x with infeasible label detection), and reported coverage ratio is very high (average of 90%, median value 94%)—especially with the help of infeasibility detection (average 95%, median value 100%).

9. Threats to Validity

The validity of our results has been crosschecked in several ways:

- We verify that all the labels covered by DSE\((P)\) are also consistently covered by DSE\((P^*)\) and that all the labels covered by DSE\((P^*)\) are also consistently covered by DSE\(^*(P^*)\) (both properties are satisfied as all DSE variants are run in the same deterministic way, so that the optimizations of more advanced variants free up additional test budget to cover more labels).

- We replay the test suites generated by all DSE variants using the LRe-PLAY tool and check that the reported coverage ratios are consistent with those reported by DSE\((P^*)\) and DSE\(^*(P^*)\).

- We check that all the labels covered by any technique (including random testing) are disjoint from the labels identified as infeasible by our approach.

- We specifically test that DSE\(^*(P^*)\) does not try to branch on an already covered label and that all path constraints from DSE\((P^*)\) have at most one label constraint (by reviewing a fraction of the output logs for constraint solving and covered labels).

Another class of threats may arise because of the tool implementations we used, as it cannot be completely excluded that Frama-C or our implementation are defective. However, Frama-C is a mature tool with industrial applications in highly demanding fields (e.g., aeronautics) and thus, it is unlikely to cause important problems. Moreover, our sanity checks would have likely spotted such issues.

While we have not directly evaluated the bug-finding power of label-based DSE, we have relied on common coverage criteria, whose bug-finding capabilities have already been extensively studied.

Common to all studies relying on empirical data, the present study may be of limited generalizability. To diminish this threat we used 12 standard benchmark programs, mainly coming from the Siemens test suite, the Verisec benchmark and MediaBench, and explored overall more than 10 millions of execution paths. While dynamic symbolic execution might still face scalability challenges in some situations, due to path explosion and constraint solving, it is yet a popular and powerful test generation technique, successfully applied in many industrial contexts. Additionally, our label framework is not tied to symbolic execution but
could also be integrated with other test generation approaches. As demonstrated in [19], detecting infeasible labels can scale up to large programs, via a compositional and parallel applications of assertion checkers.

Our results might also have been affected by the choice of the considered coverage criteria and in particular the specific mutation operators we employ. To reduce this threat, we used popular coverage criteria (CC, MCC, GACC and WM) included in software testing standards [13] [12], and employed commonly used mutation operators included in recent work [11] [10].

Finally, other threats may be due to the form of the infeasible objectives that we target, as well as the significance of their number in the considered software. However, infeasible objectives are a well-known issue, usually acknowledged in the literature as one of the main cost factors of the software testing process [16] [15] [14]. To reduce this threat, we considered several coverage criteria and various examples of programs in the benchmarks.

10. Extensions and Applications

We summarize here some recent extensions to the framework for automated test generation presented in this paper.

10.1. Extending the expressive power of labels: hyperlabels

While labels can express a large range of criteria (including a large part of weak mutations WM, and the weak GACC variant of MCDC), they still face some limits in terms of expressiveness. For instance, labels cannot express strong variants of MCDC, higher-order mutations or most path and dataflow criteria directly. In [8] [9], we introduce five simple label combination or enrichment operators: value bindings, sequences, guards, conjunctions and disjunctions. By combining and enriching labels using these operators, one can build new and more complex objectives to be covered, called hyperlabels. Hyperlabels are able to encode all criteria from the literature but full mutations, and enable specifying test objectives aiming at detecting violations of complex security properties, such as non-interference. Lifting our framework to provide efficient test generation and infeasible objective detection for hyperlabels is interesting future work.

10.2. Pruning out redundant and infeasible labels, at scale

As full objective coverage is rarely reached in practice, testers rely on the ratio of covered objectives to measure the strength of their test suites. However, the working assumption of this practice is that all objectives are of equal value. Testing research demonstrated that this is not true, as duplication and subsumption can make a large number of feasible test objectives redundant. Such feasible redundant objectives may artificially deflate or inflate the coverage ratio, skewing the measurement, which may misestimate test thoroughness and fail to evaluate correctly the remaining cost to full coverage. The approach introduced in [19] is similar to the one used here to detect infeasible labels and
enables one to prune out many redundant pairs of labels. The proof of redundancy for a pair of labels is reduced to a proof of validity for a corresponding assertion in the code, which is delegated to an off-the-shelf assertion prover.

Infeasible and redundant label detection relies on using an assertion prover as a back-end, whose complex analyses may poorly scale to real world programs. Marcozzi et al. [19] also demonstrate that scalable detection of infeasible and redundant labels is possible, using local-scoped compositional analyses and a multicore implementation. Experiments over large real-world applications such as OpenSSL and SQLite show that one can process hundreds of thousands of labels in hundreds of thousands of lines of code in acceptable time, pruning out more than 10% of the labels as either infeasible or redundant.

10.3. Ongoing industrial adoption of label-based technology

Recent work [33] reports on an ongoing industrial adoption and further enhancements of label-based testing tools at MERCE, a research center of Mitsubishi Electric. MERCE performed experiments with an industrially adapted and extended version of LTest (including program annotation, detection of infeasible objectives and test generation) on a real industrial code (with about 1,300 functions and 80,000 lines of C code). The labels were generated automatically thanks to the program annotation module. The detection of infeasible objectives using static analysis was very efficient and classified 8.3% of objectives as infeasible, within a couple of minutes. Test generation for the remaining labels covered 86% of test objectives and took about 8 hours. MERCE roughly estimated the effective time benefit factor compared to the manual testing as more than 230x for test input generation.

An even more recent experiment combining label-based testing with genetic search-based test-generation techniques reported by MERCE [34] lead to the classification (that is, either covering by a test case or proving to be infeasible) of more than 99% of test objectives out of 20,000 objectives on a real industrial code (with about 82,000 lines of C code). Those very good results are very encouraging for further pushing the technology in industry.

10.4. Using labels for combinations of testing and proof

An interesting usage of label coverage appeared in recent research on combinations of testing and proof. In the context of certification of avionic software [12] (e.g., according the DO-178C norm, level A), the verification engineer must demonstrate sufficient testing coverage of both code (structural coverage) and specification (functional coverage). When some parts of code are formally proved, while others are tested, the requirements on the coverage should be modified. Recent work [35] proposed a new notion of verification coverage based on the notion of labels and mutation coverage. It also proposed a methodology to ensure that a verification campaign is complete with respect to this coverage. It allows the verification engineer to combine testing tools and provers in the verification process and to reduce the verification cost.
10.5. Using labels for detecting polluting test objectives for data-flow criteria

For dataflow criteria, test objectives are often expressed using the notion of a def-use pair, linking a statement where a variable is defined (written) to a statement where it is used (read) without being redefined in between. A def-use pair can be defined as a sequence of two labels with an additional condition on the path between them (that must be def-clear for that variable). Def-use pairs are combined to form more complex criteria such as all-defs and all-uses. They are examples of criteria expressible in hyperlabels (cf. Section 10.1).

Recent work [1] extended the LUNCOV module of LTest to the detection of polluting test objectives for dataflow criteria using various static analysis techniques: dataflow analysis, value analysis and weakest precondition calculus. The reported experiments (on programs of up to 11000 loc) show that 64% of objectives were identified as polluting (non-inapplicable, infeasible or redundant). The analysis time remained acceptable (taking at most 64min. on the longest example with \( \sim 45000 \) test objectives).

11. Related Work

Lifting DSE to various coverage criteria. The need for enhancing DSE with better coverage criteria has already been pointed out in active testing (also known as assertion-based testing) [40, 48, 53], Mutation DSE [65, 66] and Augmented DSE [51, 68, 77]. The present work generalizes these results and proposes ways of taming the potential blow-up, resulting in an effective support of advanced coverage criteria in DSE with only a small overhead. More precisely, we give a more generic view of the problem, identifying labels and annotated programs as the key concept underlying the approach. We also clearly state the limits and hypotheses of the method by introducing the notions of simulation and labelling functions, identifying the side-effect free fragment of WM, proving soundness of direct instrumentation and providing a formalization of the path space complexification induced by direct instrumentation. Most importantly, we propose the tight instrumentation which is proved to completely prevent complexification. Finally, our optimizations can be implemented in a pure black-box setting and we do not impose any specific search heuristics, keeping room for future improvements.

Active testing targets run-time errors by adding explicit branches into the program. It is similar to the Run-Time Error Coverage criterion presented in Section 4. Labels are a more general approach. The direct instrumentation \( P' \) for this criterion is mostly equivalent to \( P^* \) since additional branches can only trigger errors and stop the execution. Yet, active testing could benefit from the IDL optimization. Finally, since most test objectives are (hopefully!) uncoverable for Run-Time Error Coverage, some approaches aim at combining DSE with static detection of uncoverable targets [39, 40]. They can be reused for labels, and should be useful when many labels are uncoverable.

Following Offut et al. [45], Papadakis et al. show that WM can be reduced to branch coverage through the use of a variant of Mutant Schemata [72]. This
is pretty similar to the direct encoding $P'$ mentioned here. They propose essentially two variations of DSE for mutation testing: a black-box approach \cite{65} based on a direct encoding similar to our DSE($P'$) scheme, and a more ad hoc approach \cite{66} preventing reuse of existing DSE tools but offering several optimizations. Papadakis et al. propose a variant of IDL, a dedicated search heuristic based on shortest paths \cite{64} and an improvement of the direct encoding through the use of mutant identifiers (following exactly Mutant Schemata). On the one hand, it ensures that a given path cannot go through several different mutants, on the other hand there is still an exponential blow-up of the search space in the worst case, and IDL cannot cover more than one mutant at once.

Augmented DSE \cite{51} is a variant of direct instrumentation. Several coverage criteria are encoded, getting results similar to those of Section 4.2, yet the side-effect free subset of WM is not identified. Experiments \cite{51, Table 2} report an average time-overhead of 272x, going up to >2,000x. That confirms the strong benefits of our optimizations, that yield a maximal overhead of 7x (60x with GACC). Following the same approach, Pandita et al. show that GACC can be simulated through direct instrumentation \cite{68}. This result can be directly recast in terms of labels, cf. Section 8.3.

In the same line of ideas, JPF allows specifying complex coverage criteria through temporal logic formulas \cite{78}. This is slightly more expressive than labels because temporal formulas can express constrained reachability objectives (e.g. dataflow criteria) while labels are in principle limited to strict reachability. Yet, constrained reachability can be reduced to strict reachability with proper instrumentation. Labels describe test objectives in a somewhat less convenient way, yet we propose an in-depth analysis of their expressive power in terms of coverage criteria (especially mutations) and a very efficient support in symbolic execution. Note that JPF provides an encoding of Masking MCDC (a.k.a. CACC \cite{25}), a criterion standing in between GACC and MCDC. Their approach \cite{78} can be recast in terms of labels.

**Labels and optimized DSE.** The label-specific optimizations described here can be freely mixed with other DSE optimizations. It is left as future work to explore which optimizations turn out to be the most effective for labels. As already stated, combining static discovery of uncoverable labels with DSE \cite{39, 40} could be useful for often-uncoverable labels, such as those generated for Run-Time Error Coverage or MCC. First results in that direction \cite{26} are presented in Section 8.2, more recent results about uncoverability detection can be found in \cite{27}. Other promising directions are to adapt DSE search heuristics \cite{75} by taking advantage of the dissimilarities between labels and branches, possibly getting inspiration from \cite{64} and \cite{70}, or to distribute the DSE search thanks to a static pre-partitioning of test objectives \cite{70}.

The IDL optimization shows some similarities with Look-Ahead pruning (LA) \cite{29, 30}. Basically, LA takes advantage of (global) static analysis to prune partial paths which cannot reach any uncovered branches. In particular, on $P^*$, IDL prunes several label-terminated paths at once thanks to dynamic analysis, which is orthogonal to LA.
Automation of mutation testing. Mutation coverage \cite{44,61} has been established as a powerful criterion through several experimental studies \cite{24,61}. Yet, it is very difficult to automate. Even mutation score computation is expensive in practice if not done wisely. Weak mutations \cite{50} relax mutation coverage by abandoning the “propagation step”, making WM easier to compare with standard criteria and easier to test for. WM has been experimentally proved to be almost equivalent to strong mutations \cite{59}, and from a theoretical point of view WM subsumes many other criteria \cite{62}.

The few existing symbolic methods for mutation-based ATG rely on the encoding proposed by Offutt et al. and have already been discussed \cite{45,67,66}. The Mutation Schemata technique \cite{72} was originally developed in order to factorize the compilation costs of hundreds of similar mutants. Static analysis has been proposed for the “equivalent mutant detection” problem \cite{57,56} in a way similar to what is sketched in Section 7.

The side-effect free fragment of WM presented in this paper seems to be a sweet spot of mutation testing: it is amenable to efficient automation and still very expressive. It is left as future work to identify if something essential is lost within this fragment. Finally, our encoding of WM into LC is orthogonal to and can be combined with some of the many techniques developed for efficient mutation testing, such as operator reduction \cite{60,73} or smart use of operators \cite{52}.

Property-based testing. Property-based testing \cite{5} uses an automatic test generation tool to find violations of some semantic properties by the program under test. Contrary to labels, which are structural test objectives generated automatically and massively from a syntactic analysis of the program under test, property-based testing targets violations of a limited number of hand-written partial specifications of the program’s semantics. While some recent works have advocated the use of more directed test generation approaches in property-based testing \cite{6,7}, state-of-the-art tools rely on random test generation to find property violations and do not take advantage of static analyses to prove correct properties.

12. Conclusion

In this paper, we present a complete panorama of different achievements related to the label coverage criterion. Label coverage appears to be both expressive and amenable to efficient automation. Some of the ideas behind labels underlie previous work by other teams. We generalize them, propose ways of taming the potential complexification of the path space and provide both formal and experimental evidence. Especially, we have shown how to extend DSE for label coverage in a black-box manner with an acceptable overhead and how to efficiently detect infeasible test objectives by existing assertion checkers. Experiments show that our optimizations yield very significant improvements. Recent applications and ongoing adoption of the label coverage based techniques in industry \cite{34,33} demonstrate the practical interest of the technology.
This work also bridges part of the gap between symbolic ATG techniques and coverage criteria. On the one hand, we show that DSE techniques can be cheaply extended to support more advanced coverage criteria, including side-effect free weak mutations. On the other hand, we identify a powerful criterion amenable to efficient automation.

As a whole, label coverage forms the basis of a very generic and convenient framework for test automation, providing a powerful specification mechanism for test objectives and featuring efficient integration into symbolic ATG techniques.

Acknowledgment

The authors would like to thank the Frama-C and PathCrawler team members for providing the tools and support, François Cheynier for initial experiments with labels, as well as David R. Macier and Alastair F. Donaldson for useful discussions about property-based testing. Many thanks to the anonymous reviewers for many helpful comments and suggestions. This work was partially funded by EU FP7 (project STANCE, grant 317753) and French ANR (project BINSEC, grant ANR-12-INSE-0002).

References


