

Mind the Gap: The Difference Between Coverage and Mutation Score Can Guide Testing Efforts

Kush Jain

Carnegie Mellon University
Pittsburgh, Pennsylvania

Goutamkumar Tulajappa Kalburgi

Northern Arizona University
Flagstaff, Arizona

Claire Le Goues

Carnegie Mellon University
Pittsburgh, Pennsylvania

Alex Groce

Northern Arizona University
Flagstaff, Arizona

Abstract—An “adequate” test suite should effectively find all inconsistencies between a system’s requirements/specifications and its implementation. Practitioners frequently use code coverage to approximate adequacy, while academics argue that mutation score may better approximate true (oracular) adequacy coverage. High code coverage is increasingly attainable even on large systems via automatic test generation, including fuzzing. In light of all of these options for measuring and improving testing effort, how should a QA engineer spend their time? We propose a new framework for reasoning about the extent, limits, and nature of a given testing effort based on an idea we call the *oracle gap*, or the difference between source code coverage and mutation score for a given software element. We conduct (1) a large-scale observational study of the oracle gap across popular Maven projects, (2) a study that varies testing and oracle quality across several of those projects and (3) a small-scale observational study of highly critical, well-tested code across comparable blockchain projects. We show that the oracle gap surfaces important information about the extent and quality of a test effort beyond either adequacy metric alone. In particular, it provides a way for practitioners to identify source files where it is likely a *weak* oracle tests *important* code.

Index Terms—code coverage, oracle strength, mutation testing

I. INTRODUCTION

Which cryptocurrency project has better testing practices, Bitcoin Core or Algorand? Both projects’ tests cover their critical core transaction verification logic reasonably well: Algorand’s tests achieve 90% statement coverage on the files in question, while Bitcoin’s cover an astonishing 98.7% of the statements in equivalent core code. Using this (very common, in practice [12], [19]) test suite adequacy metric, although Algorand is certainly well-tested, Bitcoin has the edge. Of course, the question being asked is not specific to cryptocurrency codebases, but can more generally be asked of any two testing efforts, including different tests for the same project, or the same project at different times.

It is well known that *executing* code and actually fully functionally testing it are not the same thing. Code coverage is informative, but is really a one-way test rather than a true measure of test suite quality: low coverage is bad, but high coverage does not mean a test suite is truly adequate. Another long-studied way to measure adequacy relies on *mutation analysis*, which checks how well tests detect syntactic changes injected into source code files. Mutation analysis *incidentally* measures code coverage (you cannot detect changes you do

not run) but *fundamentally* measures oracle quality/power: the ability to tell “good” from “bad” code [34].

Returning to our cryptocurrency project example, armed with this expensive but arguably more precise metric we can again ask: who is winning, Algorand or Bitcoin? Using an any-language mutation testing framework [13] on Bitcoin’s transaction verification code we find a more than respectable 75.85% mutation score. Meanwhile, using comparable settings, Algorand’s tests boast a remarkable 99.7% mutation score! Perhaps Algorand is “winning” after all? Modern test adequacy research focuses on correlating either coverage, mutation score, or both, with the prevalence of faults, claiming e.g., that coverage is “good enough” or that mutation’s expensive “more informative number” is needed [38], [11], [4]. Here, though, mutation score is not a “refinement” of coverage, but yields a substantially different assessment of effectiveness.

In this paper, we present evidence in favor of a new way of thinking about these metrics, and in particular how to use the relationship between them to inform testing efforts. Our opening question is a red herring: in practice, which of two different projects has better tests is not nearly as important a question as “how should Bitcoin’s (or Algorand’s) test engineers spend their time?” Neither coverage nor mutation score alone explains *where a test effort has been effective thus far, and where effort should now be directed*.

The structure of testing advice to date, regardless of metric, is most easily summed up as “write more tests.” As automated testing grows, however, improving test coverage and actual *oracle* power often diverge into separate engineering efforts and research topics. For example, changing fuzzers or using symbolic execution to cover additional obscure paths is a poor use of time if assertion weakness is the primary limit of a given test effort. Meanwhile, adding new checks is of limited value when a testing effort is not able to find most outright *crashes*. The naive approach, to focus on improving coverage until it is nebulously “high” before working on improving assertions, ignores the fact that in many code bases, once the most critical functionality is covered, there may be more utility in improving the existing tests than in “racking up” coverage gains by simply running uninteresting code.

We define the *oracle gap* as the difference between source code coverage and mutation score for a given software element. Naive, automatically-generated tests that cover code without strong assertions straightforwardly produce positive

gaps; basic coverage is easier to achieve than high mutation scores. Conversely, code that is well tested but not well covered produces negative gaps: coverage is low, but the mutation score is high because the provided oracles for covered code are strong, and the covered/tested elements contribute a large portion of the mutants (e.g., where complex arithmetic and logical code with many mutants is well tested, but surrounding initialization or logging calls, where only statement deletion applies, are poorly tested).

Previous efforts to establish test adequacy measures tend to focus on a single number corresponding to “testedness” [1]; monolithic scores will always struggle to distinguish very different kinds of test effort and are thus difficult to *act on* cost-effectively. Making such metrics actionable is the goal of this paper, and in our experiments we show that the oracle gap is more useful for this purpose than either metric alone, or alternatives like pseudotesting [25], [35], [36]. In many real-world projects, the code that *is covered* is well-tested, perhaps because finite resources are best devoted to code where the impact of bugs is higher, but seldom-used or merely “cosmetic” code is not executed by tests. It is not obvious that developers are doing the wrong thing, in such cases. However, it is still helpful for developers to *know* that that’s what their test effort looks like, to help allocate finite QA resources.

In particular, the primary use of the oracle gap is to identify code (files or even functions) with unusually high coverage relative to mutation score. In some, obvious instances, these may be logging or other code that is easy to execute and unimportant to test. However, such unusual *positive oracle gaps* should be prioritized for examination by developers or test engineers, in that they likely represent cases where a portion of code considered important (hence the high coverage) has not been sufficiently tested. Such cases may arise due to lack of specification effort, or due to bugs in testing (e.g., an assertion with an implication that never holds, an apparently useful but in practice vacuous check). Previous approaches to similar problems do not address these problems. Checked coverage [31] and related methods [17], for example, cannot detect cases where a statement is covered and flows to an assertion, but the assertion is written in such a way that the outcome is never false. Prioritizing code *to examine for oracle weaknesses* by code coverage alone is obviously useless. Prioritizing by mutation score alone similarly will usually simply point out poorly-covered (and often unimportant) code. We therefore propose a simple, primary use for the oracle gap: *examine source elements (usually files) exhibiting unusual positive oracle gaps, as these are the likely locations of missing or buggy assertions.*

Our contributions are of four kinds. First, fundamentally, we define the *oracle gap* and the *covered oracle gap* to measure the tendency of a test effort to 1) favor executing code over checking its behavior for correctness, 2) favor checking correctness of executed code over covering structural code elements or 3) balance these goals. Second, we gather empirical data from real-world Java projects, investigating the distribution of raw and covered oracle gap, showing that

knowing the oracle gap adds useful information over either coverage or mutation score alone. Third, a synthetic examination of oracle gap on a subset of Java projects demonstrates the oracle gap’s ability to identify gaps in testing efforts or, put differently, to provide accurate, actionable testing advice. Finally, we present a qualitative investigation of how oracle gap varies across equivalent implementations from several comparable projects that should be thoroughly tested, and thus surfaces differences in test approach and quality not otherwise visible.

II. THE ORACLE GAP

The most common approximation for test adequacy in practical use is *code coverage*, or the percentage of the code base (measured in terms of lines or branches, typically) the test suite executes. Although cheap to compute, it measures execution, not testedness. *Mutation analysis* checks how well a project’s tests detect syntactic changes injected into project source code [8]. Mutation analysis incidentally measures code coverage, but more fundamentally evaluates the test suite’s ability to detect bad code.

Mutation analysis is well established in the academic literature [23], [3], [20], [37]. As a practice, however, it has only recently begun to achieve industrial penetration [29], [2], in large part due to its computational expense, and seldom as an actual adequacy metric (vs. a pinpoint for test issues in *newly committed code*). Prior work has extensively examined the correlation between various mutants generated by mutation analysis and real world faults in code, finding that in many cases mutation analysis can mimic such faults in code [21]. As a result, researchers recommend using mutation score to measure adequacy in terms of a test suite’s likelihood to detect bugs. In contrast, defenses of code coverage often amount to claims that, in practice, it is predictive enough of mutation score that it is reasonable to use it in its place.

We argue in favor of reasoning about coverage and mutation score in tandem to inform both assessments of current testing efforts, and decisions about where to expend future effort. Current practice, focusing on mutants of “new code,” fails especially when future effort should re-visit older, but weak, tests. We define the *oracle gap* for a given test effort as:

$$\text{oracle_gap}_{X,T} = \text{cov}(X,T) - \text{mut}(X,T) \quad (1)$$

Where X corresponds to the unit of analysis (file, project, etc) and T to its test suite. Computing the difference between two different measures (i.e., statements covered and mutants killed) that are themselves measured as fractions of different units may at first appear improper; however, we could formally map any two metrics into an abstract adequacy measure, using fixed coefficients of 1.0, or by defining a suitable transfer function. Informally, the key is that the difference is only interesting in terms of sign and approximate magnitude. In practice, a comparison of whether code is “more covered” or “more checked” is straightforward enough for conception and application. In the research literature, informal comparisons

of, e.g., branch and statement coverage are not infrequent [9], [5]. Indeed, PIT and other mutation tools arguably report a limited kind of “binary” gap, by noting (only) individual lines of code that are covered, but whose mutations are not detected. In a sense we simply extend this notion to files and other larger coverage units including entire projects, with a finer granularity for mutation score. Our first research question is therefore:

RQ1: What is the empirical relationship between mutation score and coverage?

We study this question on a large dataset of Java projects (Section III). We find that mutation score and coverage are indeed positively correlated (as shown in many previous studies [28], [18], [1]), although with high variance. Despite this, there should not be (and in fact, we do not find) a perfect correlation; otherwise, it would never make sense to use the computationally more expensive mutation score. Instead, we find that the two correlate particularly well at low coverage values (where mutation score rarely exceeds file coverage). Better covered files show significantly more variation, with frequent large positive gaps between coverage and mutation score—indicating code that is executed, but poorly checked.

On the face of it, then, divergences between coverage and mutation score might be most informative when the implicit measure of coverage contained in mutation score is removed. We therefore distinguish between raw oracle gap and *covered oracle gap*, or the gap between code coverage and mutation score *over covered code only*. Prior work [27], [4] runs mutation testing on all code, giving a score strongly related to coverage. Recent work by Google [29] only mutates covered code, because mutations to uncovered code are obviously not going to be detected, even if they induce outright crashes. This informs our second research question (also Section III):

RQ2: How does the correlation between mutation score and coverage differ when only considering covered code (lines)?

Covered oracle gap is a more novel point of view on a testing effort. The covered oracle gap directly speaks to the quality of the developer-written oracles, taking into account that one cannot detect mutations of unexecuted code.

Both types of oracle gap speak to the quality and form of a testing effort. However, our first two questions are fundamentally descriptive, retroactively using the oracle gap to evaluate testing efforts post facto. For the oracle gap to be useful, it must be better able to precisely identify actionable inadequacies in testedness. Therefore, we ask the question:

RQ3: Can oracle gap clarify testing *problems*?

We answer this question via a synthetic experiment where we artificially vary the coverage and oracle power of well tested real-world Java code (Section IV). We find that the oracle gap does “move” as expected, increasing as assertions are removed or coverage is added, decreasing otherwise, corresponding to expected advice to focus attention on either coverage or test oracle power, respectively. We moreover find that it does so more precisely than alternative metrics.

TABLE I: Descriptive Statistics for Large Java Corpus.

Statistic	Mean	Median	Std. Dev.	Total
<i>Per project</i>				
#files	267.8	156.0	447.0	12051.0
mutation analysis runtime (mins)	51.8	30.3	236.6	54758.9
lines of code	10298.2	5767.0	11715.0	463417.0
# tests	1277.6	483.0	1881.2	42161.0
<i>Per file</i>				
line coverage	63.8	81.3	39.5	-
#mutants	41.9	30.5	36.0	44320.0

Finally, we look at a smaller dataset of comparable programs where we could identify central functionality that *should* be well tested to answer:

RQ4: What are the implications (and causes) of a small or large, and positive or negative, oracle gap, across comparable real-world test efforts?

To answer this question, we performed a case study of high market cap cryptocurrencies’ transaction and block validation code (Section V). We manually examined test suites and determined the reason(s) for different oracle gaps.

III. ORACLE GAP ON LARGE JAVA PROJECTS

Our first two research questions ask:

RQ1 What is the empirical relationship between mutation score and coverage?

RQ2 How does the correlation between mutation score and coverage differ when only considering covered lines?

We conduct an observational study of large Java projects, providing a high level view of how mutation score and coverage are related, as well as general oracle gap trends.

A. Experimental Setup

1) *Dataset:* We aimed to construct a dataset of large, well-maintained, popular real-world Java projects. Our analysis requires both coverage computation and mutation analysis, and so we sought projects that used the Maven build system and provided coverage reports. We collected Java projects from GitHub ordered by star count (a proxy for popularity) as well as repositories from top open source organizations with more than 20 stars. Our final corpus contains 45 projects, totalling 463,417 lines of code across 12,051 files. Table I lists descriptive statistics. Although there is significant variation, the average file is fairly well covered (63.8%) by the projects’ extant test suites.

2) *Setup and Analysis:* We use the `universalmutator` [14], a multi-lingual regular-expression-based tool for mutant generation, to compute mutation score. `universalmutator`’s operators are the usual combination of arithmetic (e.g., replace + with -), logical (replace && with ||), statement deletion, and control flow (e.g., adding `break` or `continue`) changes. We

used `universalmutator` instead of the popular PIT tool for Java [7] first because `universalmutator` can handle C and C++ as well as Java, allowing us to use consistent settings and operators (modulo a few language-specific operators) across all research questions. Second, we manually investigated certain mutants to better understand our results. PIT mutates code at the bytecode/ASM level, while `universalmutator` produces easily interpreted source-level mutants. `universalmutator` has been validated against other widely used mutation tools for Java, C++, Rust, and Python mutation, and has a good combination of low equivalency rates and diversity of mutation operators.

Computing mutation score on all files in every project in our corpus is computationally intractable. We therefore sampled up to 100 mutants per file, for up to 100 files per project. We perform stratified sampling by statement coverage to select the files for analysis per project, bucketing all files into 0-25% coverage, 25-50% coverage, 50-75% coverage and 75-100% coverage. We randomly select 25 files in each bucket for buckets with more than 25 files (and all files otherwise). We performed no analysis to remove equivalent mutants, as our use of mutation scores is focused only on comparing differences across projects, where we can assume equivalence rates are sufficiently similar across projects to not substantially impact our results. Our corpus contains a median of 236 mutants/file, while we sample a median 30.5 mutants/file, giving us a 12.9% sample rate overall. This is in line with prior work [10] that analyzes mutation score in light of sampling; they obtain a 7% error when sampling less than 5% of total mutants.

B. Results

Figure 1 shows a linear regression fit between mutation score over all lines of code and code coverage; Figure 2 shows the same over covered lines of code. Figure 3 shows boxplots of raw and covered oracle gaps at different coverage thresholds, and in aggregate.

1) *Raw Oracle Gap (RQ1)*: Figure 1 shows a weak positive correlation between mutation score and coverage. The correlation coefficient of the regression line is 0.7479, with an r-value of 0.757. Variation is moderately explained by the line, with an R^2 value of 0.573. For low coverage files, mutation scores tend to be very low, with only a couple of outliers. Figure 3 plots oracle gap by coverage range, showing substantial variance (603.92) and a Pearson correlation coefficient of 0.4591 between coverage and raw oracle gap. Importantly, the residual plot indicates that there is likely *not* a strictly linear relationship between mutation score and coverage, as residual magnitude increases with coverage. Instead, the plot suggests that at lower levels of coverage, the relationship is somewhat linear, but as coverage increases it becomes a much weaker predictor of mutation score. This is also visible in Figure 3.

This increased variance at higher coverage values (for files in the 75%–100% coverage bucket, variance is 823.31, compared to a variance of 25.35 for files in the 0%–25% bucket) suggests that mutation testing and oracle gap are more

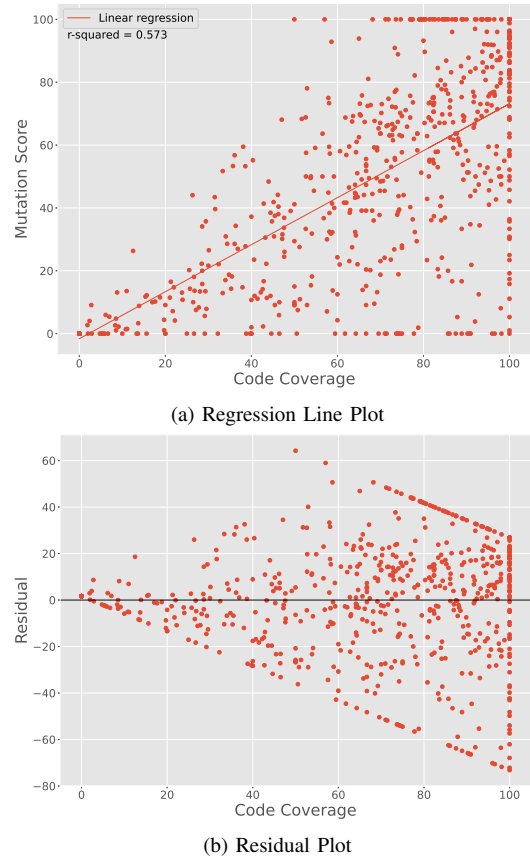


Fig. 1: Regression and residuals of coverage vs mutation score for all lines. Coverage and mutation score are weakly positively correlated.

informative at higher coverage levels (motivating mutation analysis only once a certain amount of test effort has been expended).

2) *Covered Oracle Gap (RQ2)*: Unlike the relationship between overall mutation score and coverage, the relationship between covered lines and mutation score *over covered lines only* is much noisier overall, with weaker correlation. The regression in Figure 2 shows a correlation coefficient of 0.2066 and an r-value of 0.145. The R^2 of 0.021 indicates that very little of the variation can be explained by the regression line.

The residual plot, however, depicts an interesting phenomenon: at lower coverages the relationship tends towards a negative oracle gap. This suggests that for poorly-covered files, the outliers tend to be cases where few lines are being tested, but in general these lines are well tested. At higher coverage, the oracle gap trends towards positive oracle gaps, as would usually be expected given that code is easier to execute than to check. Figure 3 also supports this conclusion, with a slight upwards trend in covered oracle gaps, as coverage increases (correlation coefficient of 0.5661) and an increased aggregate variance of 1142.74. Additionally, and by contrast with the results for raw oracle gap, variance in covered gaps *decreases* at higher coverage levels (for files with 75%–100% coverage, variance is 742.30 compared to a variance of 1655.82 for

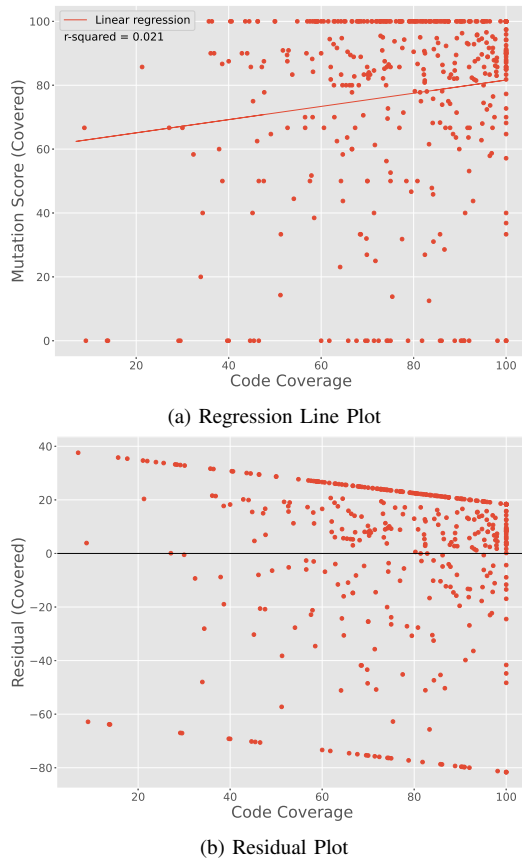


Fig. 2: Regression and residuals of coverage vs mutation score for covered lines, showing significant noise and no clear correlation.

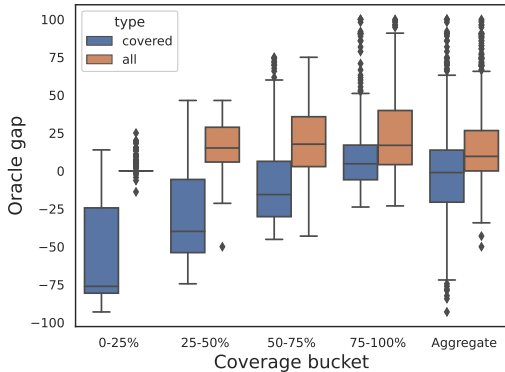


Fig. 3: Boxplot of oracle gaps, both raw and covered, for the Java dataset. Covered oracle gaps tend to increase with coverage, but variance decreases; there is high variance overall, though it is higher in aggregate for covered gaps.

files with 0%–25% coverage). That is, it appears that better-tested files tend towards more consistent testing behavior (trending slightly positive in terms of gap). We examine these phenomena further in Section V, on a different, more directly comparable dataset of projects.

To validate our observations, we manually examined the 26 files in our corpus with greater than 80% coverage and less

than 20% mutation score. Of these files, at least 12 clearly had multiple missing assert statements in their tests, failing to catch obvious defects introduced by “gross” mutation. Another 7 files had one specific type of missing assertion, corresponding to a specific mutation operator. Finally, the remaining 7 files were false positives caused by, e.g., message string or log statement mutations, leading to spurious low mutation scores.

3) *Discussion:* Based on these results, we observe:

- As coverage increases, oracle gap variance also increases, suggesting more value in running mutation analysis for test suites that achieve higher coverage.
- The *covered oracle gap* exhibits significant variance, in aggregate, though it decreases substantially as coverage increases. Moreover, while the basic limit that low coverage tends to make a high mutation score impossible causes raw oracle gaps to tend positive, *covered* oracle gaps are both negative and positive. They cluster around zero, indicating a kind of “default” balanced test effort. However, many projects have much larger gaps, suggesting more effort should be spent in improving the “lagging” measure.
- Our manual review suggested that large positive covered oracle gaps often clearly indicated an actionable lack of important assertions.

Since both Figures 1 and 2 showed significant noise, we also computed the variance of the oracle gap within projects as compared to overall variance. Average project variance was 402, compared to the overall variance of 604. Projects tend to have a higher degree of similarity with regards to their oracle quality even across differently-covered files as compared to the aggregate variance across all files (i.e, projects do, even within large variance, have a “testing style”).

Finally, we note that a rank listing of files by mutation score may likely be very similar to a ranking by coverage. While our data does not show this for individual projects (we did not collect enough files per project), we note that coverage and mutation score rankings of all files have a Spearman correlation of 0.56. In contrast, coverage and covered oracle gap have correlation of 0.53 and mutation score and covered oracle gap have a correlation of only 0.28. Considering only covered mutation score lowers the correlation further to 0.13. Of course, end users will seldom wish to rank files across many projects, but the relative relationship (where coverage and mutation score provide fairly redundant information about testing, but oracle gap provides an orthogonal ranking) seems likely to exist within projects as well. A rank listing by oracle gap will highlight cases where mutation score may technically be “good” but lags coverage by an unusual amount; this is never clear from mutation score or coverage *alone*.

IV. DETECTING TESTING PROBLEMS

Our analysis of trends in Section III suggests that oracle gaps *may* be informative, especially in projects with better coverage, and on covered code only. However, it is a purely descriptive study; being retroactive, it does not fully establish

that the oracle gap can pinpoint problems in a testing effort *as they occur*. This motivates our third research question:

RQ3: Can oracle gap clarify testing *problems*?

To answer this question, we artificially induce “problems” in testedness by synthetically varying the coverage and oracle power of test suites for well-tested Java files, and observe the corresponding behavior of the oracle gap. We do this to confirm that the (covered) oracle gap does, in general, “advise” the testers of these files to focus on either coverage or oracle power, respectively, in the appropriate direction. We also compare to pseudotesting [25], an alternative test adequacy metric, to show that the oracle gap is a more effective framework for test evaluation. We supplement this with a qualitative discussion of how the oracle gap could have informed testing efforts in real-world code and changes.

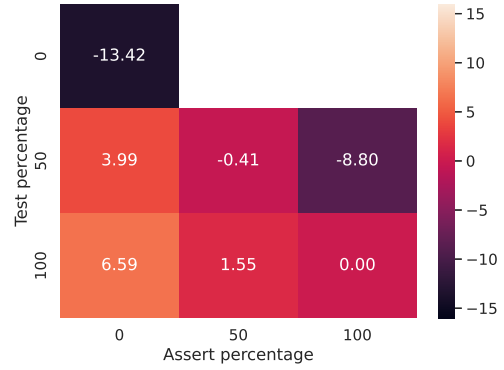
Dataset. These experiments require high quality test suites, and significant computation time. We therefore selected 25 files (and their tests) from 5 projects in our Java dataset with both high statement coverage and high mutation score; The left side of Table II summarizes this smaller corpus. We manually examined these files’ unit tests to ensure they could be manipulated as necessary.

A. Synthetic Experiment

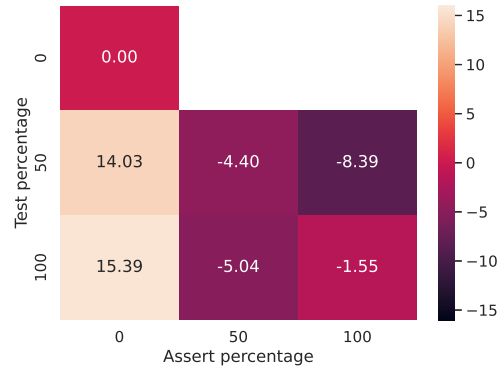
The goal of this experiment is to artificially inject testedness problems into a test effort for a file. We confirm, in general, that the covered oracle gap detects these synthetic induced issues and provides the appropriate actionable advice.

1) *Setup and Analysis:* For each file, we programmatically find all related tests and `assert` statements. We generate synthetic test suites for these 25 well-tested Java files by sampling from the starting pool of assertions and tests for each file (effectively by commenting out some number of assertions and tests to produce a smaller suite). Our configurations consist of 0%, 50%, and 100% of both assertions and tests, in combination. The intuition here is that we construct, at different levels, a range of test suites from high-coverage but low-quality, to low-coverage but high-quality, and high-coverage and high-quality (the original). For example, the 50% tests, 100% assertions configuration consists of half the tests but all of those tests’ assertions; 100% tests, 0% assertions includes all tests, but with all assert statements removed. We exclude configurations with 0% of the tests and more than 0% of assertions. For any configuration involving 50% of either assertions or tests, we randomly sample five times, unless the number of assertions per test in a file is less than 1.5 on average. We compute coverage and mutation analysis on the source code files in question using the resulting test suites to generate both the regular oracle gap and the covered oracle gap. The mutation analysis over this corpus for these experiments took 20 CPU weeks over 19 AWS EC2 instances.

We exclude integration or system tests in these experiments, and unit tests that check for exceptional behavior. We do this because we modify tests by removing `assert` statements; in tests checking for exceptional behavior, removing the `assert`



(a) 100:100 difference: element config - 100:100 config



(b) Previous difference: element config - left of element (wrapping to row above if leftmost in current row) config

Fig. 4: Difference in mean covered oracle gap against either the full unit test suite, or the “previous” configuration. Covered oracle gaps fall as more asserts are added and rise fastest when tests with no asserts are added.

means that thrown exceptions are not checked, leading the test to always (incorrectly) fail. System and integration tests are not as simply structured as unit tests, making it difficult-to-impossible to programatically systematically vary test coverage/oracle strength. The 0% tests provide baselines, measuring adequacy of the system/integration tests and unit tests for exceptional behavior alone.

2) *Results:* Figure 4a shows how the covered oracle gap changes with respect to the *full* suites (100% of tests and 100% of asserts); Figure 4b shows how the covered oracle gap changes with respect to the *previous* test suite configuration (the configuration to the left). These differences are averaged over all configuration executions on all 25 files. To understand this matrix better, consider one of the files in our corpus, `ShardingService.java` (raw numbers for this file are available in our replication package). At 0% of unit test cases (0% assertions), the tests for this file show a covered oracle gap of 21.10%. This means that the integration and exceptional behavior unit tests are substantially biased in the direction of coverage over oracle strength. This gap grows as more tests are added without their assert statements: 100% of the tests with 0% of the asserts have a covered oracle gap of 40.63%.

We also investigate what happens when a hypothetical tester added more asserts. For example, 100% of tests with 50% of available assertions consists of all the system and integration tests along with 100% of the test cases in the unit test file with 50% of their constituent assertions. Note that any configuration including 50% of tests or assertions are sampled to account for variance. When averaged across all five samples, the tests for our example file achieve a covered oracle gap of 31.75% at the 100% of tests with 50% of asserts, which is lower than the covered oracle gap of 40.53% at the 100% of tests with no assertions. Adding assert statement reduces overall oracle gap. In other words, when assertions are “missing” the oracle gap straightforwardly indicates a tester should add assertions/oracle power.

Figures 4a and 4b, show that the covered oracle gap always decreases as unit test assertions are added. Practically, this means that when covered oracle gaps are strongly positive, one can directly add more asserts to decrease the covered-oracle gap, and improve overall test oracle power. Conversely, directly adding coverage with minimal asserts (for example 0-0 to 100-0), directly increases the covered oracle gap — coverage itself increases by construction, and the oracle gap precisely pinpoints the fact that oracle power remains weak.

While the average trends of oracle gaps are in line with our hypotheses, the trend does not hold universally for all files. Specifically, while in all files adding asserts increases mutation score, coverage is not always fixed as assert statements are added. This is because asserts sometimes contain invocations to source methods, increasing coverage and oracle power simultaneously. In rare cases coverage from these asserts is greater than the corresponding increase in mutation score. That said, one way to think about the synthetic experiment is to note that on average, *following the advice of covered oracle gap* tends towards producing the final test suites. If we assume the final suites for popular mature projects are often also high quality, then the advice is generally good.

B. Comparison with Pseudotesting

Our observations so far demonstrate that the oracle gap is more actionable than mutation score or coverage alone. We next compare it against pseudotesting, which uses very coarse-grained mutants to more efficiently estimate mutation score. This is done by replacing each source method body with a single line returning some default value (e.g., replacing the body of a method that returns a `String` with `return "";`) and checking whether test behavior changes. This fundamentally means that it cannot identify cases where the lack of oracle power is nuanced; even such large mutations as statement deletions are generally under its threshold. We use the same corpus as above to ensure an apples-to-apples discussion; we used `descartes` [36] (a PIT pseudotesting tool) to measure each file’s pseudotesting score.

The right-hand columns of Table II show that for these well-tested files, pseudotesting and mutation scores vary dramatically. We suspect that this is because these are well-covered files from highly starred Java projects, meaning that it’s rare

```

1 public boolean get(final T key, final T data,
2     final SeekOp op) {
3     if (SHOULD_CHECK) {
4         requireNonNull(key);
5         requireNonNull(op);
6         checkNotClosed();
7         txn.checkReady();
8     }
9     kv.keyIn(key);
10    // ... elided happy path ...

```

Fig. 5: The function in `Cursor.java` in the `lmbdjava` newly covered by a test added in commit `a781b`

for methods to be completely untested. For such files, these results indicate that pseudotesting does not serve as a valid substitute to computing the oracle gap, since the pseudotesting numbers suggest that the test suites are more than adequate.

Pseudotesting and the oracle gap are complementary metrics. Due to the significantly lower compute cost required for pseudotesting, it may be useful early in a testing effort, when entire methods are untested; however, when oracle omissions are less blatant, it may offer little actionable information.

C. Qualitative Discussion

Our injection of “testedness problems” into high-quality test suites shows that the oracle gap would give the right “advice” to test engineers. But, the injection is by definition artificial. We therefore additionally manually examined recent real-world commits that significantly changed tests on these projects to see how the oracle gap applies in current development practice. For many commits, coverage and mutation score change in roughly equal proportion. However, here we highlight two (of several) interesting cases we found as anecdotal examples of how the oracle gap could assess and inform current testing efforts.

In `lmbdjava` project, commit `a781b` adds a new test to cover the `get` method from the `Cursor` class, shown in Figure 5. As expected, both raw coverage and raw mutation score increase (from 88.6% to 99.6% and 41.1% to 47.9%, respectively). The new test covers new code and as a result kills more mutants. However, the covered oracle gap of this file actually *increases* from 47.5% to 51.6%, because the test only checks the *happy* execution path—none of the checks on lines 3–6 in Figure 5 fail. Although both coverage and mutation score go up—substantiating that adding this new test was good—the change in the oracle gap more precisely identifies the weaknesses that remain, and could have motivated the developer to more closely examine the new tests.

In `shardingsphere-elasticjob`, commit `8e20d` adds a series of new tests that check various failure cases in the underlying source class, such as checking that jobs are paused when the computer is shutdown or that a `JobFailureException` is appropriately thrown under exceptional cases. As expected, coverage does not change much, going from 97.5% to 97.6%. However, raw mutation score

TABLE II: Synthetic experiment corpus, consisting of 25 well-tested Java files over 5 projects. Right-hand columns show covered mutation score, gaps, and descartes pseudotesting scores.

Project	Tested File	Test LOC	Mutants	Tests	Asserts	Pseudo Score	Covered Mut. Score	Covered Gap	Pseudo Gap
datasketches	DirectQuickSelectSketch	683	366	22	151	100%	61.3%	34.1%	-4.6
	ReservoirItemsSketch	458	604	14	85	96%	68.5%	30.2%	2.7%
	Util	279	1133	17	74	93%	73.4%	25.0%	5.4%
	BoundsOnBinomialProportions	104	702	4	12	100%	92.0%	8.5%	0.5%
	MurmurHash3Adaptor	297	741	17	53	74%	50.7%	48.0%	24.7%
apollo	StubClient	242	44	17	24	88%	66.7%	26.1%	4.8%
	JsonSerializerMiddlewares	70	10	5	9	100%	100%	-5.0%	-5.0%
	RuleRouter	273	53	29	60	100%	91.4%	-18.9%	-27.5%
	Headers	102	35	9	21	100%	83.9%	16.1%	0.0%
	ServiceImpl	531	259	41	77	83%	60.3%	33.8%	11.1%
lmdbjava	CursorIterable	271	51	21	47	100%	82.5%	12.2%	-5.3%
	KeyRange	233	35	19	41	100%	77.8%	22.2%	0.0%
	ByteBufferProxy	118	187	7	11	87%	76.4%	15.1%	4.5%
	ResultCodeMapper	124	32	5	7	100%	96.2%	3.8%	0.0%
	Txn	257	113	7	19	100%	79.1%	19.6%	-1.3%
github-api	GHGist	100	51	3	45	75%	76.3%	20.3%	21.6%
	GHLicense	120	29	10	40	86%	50.0%	44.2%	8.2%
	GHCheckRunBuilder	114	131	6	23	100%	97.9%	-2.9%	-5.0%
	GHWorkflow	122	44	5	22	80%	80.0%	9.1%	9.1%
	GHBranchProtection	87	14	6	18	60%	50.0%	21.2%	11.2%
ss-elasticjob	ElasticJobExecutor	205	69	8	20	100%	69.6%	21.5%	-8.9%
	ShardingService	246	146	18	37	85%	71.0%	29.0%	15.0%
	AverageAllocationJobShardingStrategy	56	88	6	6	100%	88.8%	11.3%	0.1%
	AppConstraintEvaluator	146	127	7	15	100%	64.2%	35.8%	0.0%
	TaskContext	92	110	13	27	100%	94.7%	4.4%	-0.9%

increases from 63.9% to 72.2%. Covered oracle gap therefore decreases substantially from 33.7% to 25.4%.

These examples show how coverage, while informative, is limited on its own as an adequacy metric. Mutation score alone, however, does not fully supplant it and, when looked at alone, may not be cost effective in terms of developer time [30]. For `lmdbjava`, mutation score *does* increase with the new tests. Given finite time and the fact that 100% mutation score is intractable, when should a developer spend more time looking at the remaining unkilld mutants? On this commit, the oracle gap gives a much clearer indication that the developer might fruitfully consider whether the new tests are as strong as they could be. Meanwhile, coverage does not increase meaningfully in the `shardingsphere-elasticjob` example, but the narrowing oracle gap shows that the new tests are adding significant strength to the overall suite. In other cases a low mutation score might, on its own, indicate a weak oracle, when in fact the problem is entirely due to poor coverage.

V. ORACLE GAP CASE STUDY

Having investigated the oracle gap in the large, we ask:

RQ4: What are the implications (and causes) of a small or large, and positive or negative, oracle gap, across comparable real-world test efforts?

Our Java studies show general trends and examples of oracle gaps within projects, and so speak to the use of finding files with actionable gaps in a project, but do not address “higher level” uses comparing testing “styles.” The diversity in kinds

of code makes drawing conclusions across different testing efforts difficult. We thus sought an example of high-criticality code of similar functionality and complexity *across projects*.

A. Experimental Setup

In particular, we examined the transaction validation code for several cryptocurrency projects. Cryptocurrencies are essentially the sum of the operations of code executed by many independent nodes, especially nodes that mine cryptocurrency. The code validating blocks and transactions for blockchains implementing cryptocurrencies is therefore of the utmost importance: arguably, checking transaction correctness is the *raison d’être* of any blockchain.

For Bitcoin Core [24], we chose the core `tx_verify.cpp` file, which verifies all transactions. We also performed mutation analysis of transaction-verification-related code for three other high market cap cryptocurrencies: `ethereum`, `avalanche`, and `algorand`. Our selection was influenced by both the market cap of each project and the availability of code coverage reports for the projects. We identified candidate files roughly comparable to Bitcoin’s `tx_verify.cpp` by searching for keywords like `transaction`, `verify`, `sign`, and `validate`. We manually inspected functions and test coverage for these functions (where applicable) to identify 1–2 files focusing on this similar core functionality per project to mutate. We ran these mutants against each project’s default test suite (determined by consulting READMEs and build/CI

documentation) using `universalmutator`. The projects are written in C++ and Go.

B. Results

Table III shows the range of raw and covered oracle gaps for code in these cryptocurrencies. Our framework exposes key differences within otherwise superficially (by coverage or mutation score) similar testing efforts on comparable code. To return to the example from the introduction, Bitcoin and go-algorand both have what would be usually considered “good” coverage, at $> 90\%$. And the 75% and 99% mutation scores are, respectively, solid and remarkable. In isolation, however, the numbers either suggest “Bitcoin core is somewhat better tested” (coverage) or “Algorand is much better tested” (mutation score). Using both numbers, however, we can see that these projects have chosen different trade-offs in testing. The Bitcoin core tests emphasize covering all code, and in fact do cover all but extremely obscure (and possibly impossible to encounter in practice) behavior. Meanwhile, Algorand leaves more code uncovered, including code that appears to correspond to obscure but still possible conditions. However, Bitcoin core has taken much less pain to construct strong oracles. In practice, 75% is generally considered a very good mutation score for real-world code, but the 99%+ mutation coverage of Algorand reflects an even greater attention to ensuring every behavior’s impact is checked, and moreover suggests that Algorand chose to cover almost all code with serious semantic impact. It is likely that to some extent Bitcoin sacrificed some oracle-improvement effort to focus on coverage, which is closely monitored by the project, while efforts to perform frequent mutation analysis on Bitcoin Core are only now beginning.

Similarly, Ethereum’s `block_validator.go` would benefit from stronger oracles. The tests for `validation.go` simply fail to cover a large portion of the file, though do a good job of checking the covered code. For `avalanchego`, the covered oracle gap is what we could consider “normal” reflecting a somewhat weaker oracle combined with reasonable code coverage. Note that by mutation score, `validation.go` from Ethereum and `add_subnet_validator_tx.go` appear to be very similar: using the conventional “best practice” of applying mutation testing alone as an adequacy measure when possible, these files would both seem to be poorly tested. The mutation score would not indicate that in one case, the failure is mostly due to poor coverage (a large negative oracle gap) and in the other case, the result is due to an oracle that significantly lags coverage. As with Bitcoin and Algorand, considering the gap tells a much fuller story about the testing efforts, and the best mitigations for weaknesses. Scores for all projects show the importance of usually focusing on *covered* oracle gap for overall understanding: raw oracle gaps are all positive, since at project level limited coverage dominates the factors leading to unkillable mutants.

C. Bitcoin and Fuzzing: Acting on the Gap

Bitcoin Core includes a complex, well-designed, set of fuzzing tests that are run on OSS-Fuzz [15]. Two particularly relevant fuzz targets in `tx_verify.cpp` are (`process_message_tx` and `coins_view`); we used these to explore qualitatively how oracle gap is particularly informative for highly automated testing methods.

Fuzzing `tx_verify.cpp` does not affect code coverage either way: fuzz test coverage is approximately as high as for functional tests, though with minor changes in exact code covered. However, overall, fuzzing could detect just under 12% of all the generated mutants, resulting in very large positive oracle gaps. Fuzzing adds only two unique mutant kills beyond those that Bitcoin Core’s existing functional tests can find.

So: why fuzz? The answer is that fuzzing uncovers subtle bugs that functional tests designed by humans will almost never detect, e.g. <https://github.com/bitcoin/bitcoin/issues/22450>. Neither fuzzing nor functional/unit tests replace one another. Consider two mutants detected by `coins_view` fuzzing alone. These correspond to two (hypothetical) bugs not detectable by any other means; in the real world, if one such bug is exploitable, detecting it may “pay for” all the fuzzing effort, and there will seldom be just one such bug (see an approximate list of fuzzer-detected, fixed bugs in Bitcoin Core at <https://github.com/bitcoin/bitcoin/issues?q=is%3Aissue+fuzz+is%3Aclosed+label%3ABug>).

While the oracle gap does not show fuzzing is useless, it *does* likely point to the most effective way to improve the fuzzing: manual, expert developer effort to improve the *oracles* used by existing fuzz targets, or efforts to craft custom, more restricted, fuzz targets with stronger oracles. Building fuzz harnesses with complex correctness checks is hard, of course; the functional tests know exactly what inputs are being provided to APIs, and can check for expected behavior. Trying to inject this kind of check into fuzz harnesses ranges from non-trivial to impossible. When applicable, more generic, “mathematical” constraints such as are used in property-based testing [6] can help, but these are often hard to devise. The most promising route to solve these problems may be to manually add high-quality invariants and assertions to source, which can be executed by both functional and fuzz tests. At present, the Bitcoin Core code has about 1,800 `assert` statements, scattered among 180KLOC of C and C++. The resulting ratio of about one assertion per 100 lines of code is not terrible, but is certainly at the lower limit of acceptable. Given that Bitcoin Core defines at least 4,000 functions, the code fails to meet the NASA/JPL proposal of an average of two assertions per function [16].

VI. THREATS TO VALIDITY

The primary threats to validity of which we are aware are ones common to all studies using mutation testing, e.g., equivalent mutants; however, as our interests are in correlations between coverage and mutation score, rather than absolute mutation scores, we do not believe these are significant. A second potential threat is that the cryptocurrency

TABLE III: Covered Mutation Scores and Oracle Gaps For Selected Files

Project	File path	File coverage	Mutation score	Covered mutation score	Raw oracle gap	Covered oracle gap
bitcoin	src/consensus/tx_verify.cpp	98.7%	75.8%	83.1%	22.9%	15.5%
go-ethereum	core/block_validator.go	81.0%	66.7%	78.0%	14.3%	3.0%
	signer/fourbyte/validation.go	60.0%	42.2%	76.4%	17.8%	-16.4%
avalanchego	vms/platformvm/add_subnet_validator_tx.go	79.1%	44.3%	67.1%	34.8%	12.0%
go-algorand	data/transactions/logic/eval.go	90.0%	99.7%	99.7%	-9.7%	-9.7%

study only examined a small number of projects and a particular functionality, so may not generalize. The threat of a focus on Java/Maven code in most of our results is however mitigated by similar findings over C++ and Go code in the cryptocurrency projects. Our replication package (<https://figshare.com/s/06cbb520b30751ebb1b2>) provides full details on all experiments and methods.

VII. RELATED WORK

While there is foundational work on the oracle/test distinction [34] at a theoretical level, most previous work assumes a basic framework of trying to determine correlation of mutation testing or coverage alone with fault detection and/or coverage with mutation score [28], [27], [4]. In such work, the oracle/test distinction is not considered in the light of whether code is “more executed” or “more checked.” The most similar approach, which does not involve mutants, is the notion of *checked coverage* [31], which, however, still results in a single score, as does pseudotesting [25], [35], [36]. Checked coverage is inherently unable to identify *buggy* assertions that do not in practice detect mutants, if there is any dynamic flow from coverage to the faulty assertion. The study of coverage gaps (in a different sense than ours) of Hossain et al. [17] extends checked coverage to further coverage criteria, and provides *recommendations* (at the individual statement level only, unlike our approach) for code that may not be checked, but does not escape this fundamental limitation. Their study does, however, show that gaps of the type we identify more completely often cause faults to go undetected.

Just et al. [21] found that 73% of real-world faults can be associated with common mutation operators. Beller et al. [2] examined how to feasibly implement mutation testing at Facebook. These studies support the notion that lagging mutation score indicates lagging fault *detection* capability despite high coverage, as also suggested by work by Sina et al. on whether automatically generated unit tests help find bugs [32]. Smith et al. [33] briefly discussed the relationship between coverage and mutation score, but examined only two open source projects. Li et al. [22] also examined when it is best to use various testing methods, including mutation testing. Recent industrial uses of mutants [29], [2] do not include gap analysis, but informally hint that large-scale users may be amenable to the type of analysis we provide, given how they use mutants to pinpoint test issues now. Finally, it

is interesting to consider that large positive covered oracle gap can be alternatively thought of as the presence of many *stubborn* mutants in a file [26].

VIII. CONCLUSION AND IMPLICATIONS

Current approaches to measuring test adequacy focus on either simply reporting structural coverage, or on reporting a complex mix of oracle power and structural coverage, in the form of mutation score. Neither approach yields easily practicable advice on the degree to which a test’s coverage and oracle efforts are “in balance.” We propose the *oracle gap* or *the difference between source code coverage and mutation score* as a new metric to help researchers and practitioners improve their understanding of test adequacy. We show that in practice the oracle gap varies widely in real code, and connect that variance to real differences in testing. Our core immediate *use* for the oracle gap, identifying cases of missing or *faulty* assertions *in important code* is supported by showing that removal of assertions creates substantially larger gaps.

This way of looking at test adequacy has implications for both testing researchers and practitioners. For researchers, while there exists the expected correlation between coverage and mutation score, the relationship is subtle: mutation score is not a “refined” coverage score, and reporting one number without the context of the other paints a partial picture, especially for automatic test generators. There is already strong impetus for research in improving or constructing oracles for existing or generated tests; our results show that the recent growth and impressiveness of fuzz testing efforts provides further motivation for examination of gaps.

This speaks to the implications of our reasoning for practitioners. Covering more code probably *is* good, but there is a balance in how finite testing effort should be allocated. Practitioners can look at the oracle gap from time to time in their test efforts to understand where their efforts should next be allocated. Companies are already using mutants of covered lines that are not killed to improve testing. In addition to localized advice, oracle gap might give visibility into what might be called *oracle debt* by analogy with technical debt: methods, files, or entire submodules where policies about coverage have yielded testing whose coverage exceeds its ability to probe for faults, or where buggy assertions make effort expended fail to pay off.

A portion of this work was supported by the National Science Foundation under awards CCF-2129446 and CCF-2129388.

REFERENCES

- [1] Iftekhar Ahmed, Rahul Gopinath, Caius Brindescu, Alex Groce, and Carlos Jensen. Can testness be effectively measured? In *International Symposium on Foundations of Software Engineering, FSE 2016*, page 547–558, New York, NY, USA, 2016. Association for Computing Machinery.
- [2] Moritz Beller, Chu-Pan Wong, Johannes Bader, Andrew Scott, Mateusz Machalica, Satish Chandra, and Erik Meijer. What it would take to use mutation testing in industry - A study at facebook. In *International Conference on Software Engineering: Software Engineering in Practice*, pages 268–277. IEEE, 2021.
- [3] Timothy Budd, Richard J. Lipton, Richard A DeMillo, and Frederick G Sayward. *Mutation analysis*. Yale University, Department of Computer Science, 1979.
- [4] Thierry Titchou Chekam, Mike Papadakis, Yves Le Traon, and Mark Harman. An empirical study on mutation, statement and branch coverage fault revelation that avoids the unreliable clean program assumption. In *International Conference on Software Engineering*, pages 597–608, 2017.
- [5] Thierry Titchou Chekam, Mike Papadakis, Yves Le Traon, and Mark Harman. An empirical study on mutation, statement and branch coverage fault revelation that avoids the unreliable clean program assumption. In Sebastián Uchitel, Alessandro Orso, and Martin P. Robillard, editors, *International Conference on Software Engineering*, pages 597–608. IEEE / ACM, 2017.
- [6] Koen Claessen and John Hughes. QuickCheck: a lightweight tool for random testing of Haskell programs. In *International Conference on Functional Programming (ICFP)*, pages 268–279, 2000.
- [7] Henry Coles, Thomas Laurent, Christopher Henard, Mike Papadakis, and Anthony Ventresque. Pit: A practical mutation testing tool for java (demo). In *Proceedings of the 25th International Symposium on Software Testing and Analysis, ISSTA 2016*, page 449–452, New York, NY, USA, 2016. Association for Computing Machinery.
- [8] R. A. DeMillo, R. J. Lipton, and F. G. Sayward. Hints on test data selection: Help for the practicing programmer. *Computer*, 11(4):34–41, April 1978.
- [9] Milos Gligoric, Alex Groce, Chaoqiang Zhang, Rohan Sharma, Mohammad Amin Alipour, and Darko Marinov. Guidelines for coverage-based comparisons of non-adequate test suites. *Transactions on Software Engineering and Methodology*, 24(4):22:1–22:33, 2015.
- [10] Rahul Gopinath, Amin Alipour, Iftekhar Ahmed, Carlos Jensen, and Alex Groce. How hard does mutation analysis have to be, anyway? In *International Symposium on Software Reliability Engineering*, pages 216–227. IEEE Computer Society, 2015.
- [11] Rahul Gopinath, Carlos Jensen, and Alex Groce. Code coverage for suite evaluation by developers. In *International Conference on Software Engineering, ICSE 2014*, page 72–82, New York, NY, USA, 2014. Association for Computing Machinery.
- [12] Alex Groce, Mohammad Amin Alipour, and Rahul Gopinath. Coverage and its discontents. In *International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, Onward 2014, page 255–268, New York, NY, USA, 2014. Association for Computing Machinery.
- [13] Alex Groce, Josie Holmes, Darko Marinov, August Shi, and Lingming Zhang. An extensible, regular-expression-based tool for multi-language mutant generation. In *International Conference on Software Engineering: Companion Proceedings, ICSE 2018*, pages 25–28, New York, NY, USA, 2018. ACM.
- [14] Alex Groce, Josie Holmes, Darko Marinov, August Shi, and Lingming Zhang. An extensible, regular-expression-based tool for multi-language mutant generation. In *International Conference on Software Engineering: Companion*, pages 25–28, 2018.
- [15] Alex Groce, Kush Jain, Rijnard van Tonder, Goutamkumar Tulajappa Kalburgi, and Claire Le Goues. Looking for lacunae in bitcoin core’s fuzzing efforts. In *International Conference on Software Engineering: Software Engineering in Practice*, 2022.
- [16] Gerard J Holzmann. The power of 10: Rules for developing safety-critical code. *Computer*, 39(6):95–99, 2006.
- [17] Soneya Binta Hossain, Matthew Dwyer, Sebastian Elbaum, and Anh Nguyen-Tuong. Measuring and mitigating gaps in structural testing. In *International Conference on Software Engineering*, 2023.
- [18] Laura Inozemtseva and Reid Holmes. Coverage is not strongly correlated with test suite effectiveness. In *International Conference on Software Engineering, ICSE 2014*, page 435–445, New York, NY, USA, 2014. Association for Computing Machinery.
- [19] Marko Ivankovic, Goran Petrovic, René Just, and Gordon Fraser. Code coverage at google. In Marlon Dumas, Dietmar Pfahl, Sven Apel, and Alessandra Russo, editors, *Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 955–963. ACM, 2019.
- [20] Yue Jia and Mark Harman. Constructing subtle faults using higher order mutation testing. In *International Working Conference on Source Code Analysis and Manipulation*, pages 249–258. IEEE, 2008.
- [21] René Just, Darioush Jalali, Laura Inozemtseva, Michael D. Ernst, Reid Holmes, and Gordon Fraser. Are mutants a valid substitute for real faults in software testing? In *Symposium on the Foundations of Software Engineering*, pages 654–665, Hong Kong, November 18–20 2014.
- [22] Nan Li, Upsorn Praphamontriphong, and Jeff Offutt. An experimental comparison of four unit test criteria: Mutation, edge-pair, all-uses and prime path coverage. In *International Conference on Software Testing, Verification, and Validation Workshops*, pages 220–229, 2009.
- [23] Richard J. Lipton, Richard A DeMillo, and Frederick G Sayward. Hints on test data selection: Help for the practicing programmer. *Computer*, 11(4):34–41, 1978.
- [24] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>, 2008.
- [25] Rainer Niedermayr, Elmar Juergens, and Stefan Wagner. Will my tests tell me if i break this code? In *International Workshop on Continuous Software Evolution and Delivery, CSED ’16*, page 23–29, New York, NY, USA, 2016. Association for Computing Machinery.
- [26] Mike Papadakis, Thierry Titchou Chekam, and Yves Le Traon. Mutant quality indicators. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 32–39. IEEE, 2018.
- [27] Mike Papadakis and Nicos Malevris. An empirical evaluation of the first and second order mutation testing strategies. In *International Conference on Software Testing, Verification and Validation*, pages 90–99. IEEE Computer Society, 2010.
- [28] Mike Papadakis, Donghwan Shin, Shin Yoo, and Doo-Hwan Bae. Are mutation scores correlated with real fault detection? a large scale empirical study on the relationship between mutants and real faults. In *International Conference on Software Engineering*, pages 537–548. IEEE, 2018.
- [29] Goran Petrovic and Marko Ivankovic. State of mutation testing at google. In Frances Paulisch and Jan Bosch, editors, *International Conference on Software Engineering: Software Engineering in Practice*, pages 163–171. ACM, 2018.
- [30] Goran Petrovic, Marko Ivankovic, Bob Kurtz, Paul Ammann, and René Just. An industrial application of mutation testing: Lessons, challenges, and research directions. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 47–53, 2018.
- [31] David Schuler and Andreas Zeller. Assessing oracle quality with checked coverage. In *International Conference on Software Testing, Verification and Validation*, pages 90–99. IEEE Computer Society, 2011.
- [32] Sina Shamshiri, René Just, José Miguel Rojas, Gordon Fraser, Phil McMinn, and Andrea Arcuri. Do automatically generated unit tests find real faults? an empirical study of effectiveness and challenges (t). In *International Conference on Automated Software Engineering*, pages 201–211, 2015.
- [33] Ben H. Smith and Laurie Williams. An empirical evaluation of the mujava mutation operators. In *Testing: Academic and Industrial Conference Practice and Research Techniques*, pages 193–202, 2007.
- [34] Matt Staats, Michael W. Whalen, and Mats Per Erik Heimdahl. Programs, tests, and oracles: the foundations of testing revisited. In Richard N. Taylor, Harald C. Gall, and Nenad Medvidovic, editors, *International Conference on Software Engineering*, pages 391–400. ACM, 2011.
- [35] Oscar Luis Vera-Pérez, Benjamin Danglot, Martin Monperrus, and Benoit Baudry. A comprehensive study of pseudo-tested methods. *Empirical Software Engineering*, 24(3):1195–1225, jun 2019.
- [36] Oscar Luis Vera-Pérez, Martin Monperrus, and Benoit Baudry. Descartes: A pitest engine to detect pseudo-tested methods: Tool demonstration. In *International Conference on Automated Software Engineering, ASE 2018*, page 908–911, New York, NY, USA, 2018. Association for Computing Machinery.

- [37] Jie Zhang, Ziyi Wang, Lingming Zhang, Dan Hao, Lei Zang, Shiyang Cheng, and Lu Zhang. Predictive mutation testing. In *International Symposium on Software Testing and Analysis*, ISSTA 2016, page 342–353, New York, NY, USA, 2016. Association for Computing Machinery.
- [38] Peng Zhang, Yang Wang, Xutong Liu, Yibiao Yang, Yanhui Li, Lin Chen, Ziyuan Wang, Chang-ai Sun, and Yuming Zhou. Test suite effectiveness metric evaluation: what do we know and what should we do?, 2022.