Initial Results on Counting Test Orders for Order-Dependent Flaky Tests using Alloy

Wenxi Wang[1], Pu Yi[2], Sarfraz Khurshid[1], Darko Marinov[3]

[1] The University of Texas at Austin, USA
[2] Peking University, China
[3] University of Illinois Urbana-Champaign, USA

Abstract. Flaky tests can seemingly nondeterministically pass or fail for the same code under test. Flaky tests are detrimental to regression testing because tests that pass before code changes and fail after code changes do not reliably indicate problems in code changes. An important category of flaky tests is order-dependent tests that pass or fail based on the order of tests in the test suite. Prior work has considered the problem of counting test orders that pass or fail, given relationships of tests within a test suite. However, prior work has not addressed the most general case of these relationships. This paper shows how to encode the problem of counting test orders in the Alloy modeling language and how to use propositional model counters to obtain the count for test orders. We illustrate that Alloy makes it easy to handle even the most general case. The results show that this problem produces challenging propositional formulas for the state-of-the-art model counters.

1 Introduction

Flaky tests [17] can seemingly nondeterministically pass or fail for the same code under test. They are detrimental to regression testing because tests that pass before code changes and fail after code changes do not reliably indicate problems in code changes. For example, Harman and O'Hearn point out problems of flaky tests at Facebook [6], and several other companies point out similar problems, including Apple [12], Google [4, 18, 24], Huawei [10], and Microsoft [7, 8, 14, 15].

An important category of flaky tests is order-dependent tests that pass or fail based on the order of tests in the test suite. More specifically, the tests deterministicaly pass in some test orders and deterministically fail in other test orders. Before establishing that the tests depend just on the order, the developers may view them as nondeterministically passing or failing in various runs. While running tests in a random order is not the default for most test frameworks, it can be used to find order-dependent tests [5, 16].

Shi et al. [19] have categorized several roles for order-dependent tests. Each order-dependent test itself can be either a *victim*, which passes when run by itself but fails when run after some other tests in the test suite, or a *brittle*, which fails when run by itself but passes when run after some other test in the test suite. Each victim test fails when run after (not necessarily immediately after) a *polluter* test, unless a *cleaner* test runs between the polluter and the victim.

Each brittle test passes when run after (not necessarily immediately after) a *state-setter* test. We focus on victim tests, because the analysis for brittle tests comes out as a special case.

Wei et al. [22] have recently considered the problem of counting the number of test orders for which a victim fails. This problem is important because it allows computing the *flake rate*, i.e., the probability that a test fails if the test order is a uniformly sampled permutation of the test suite. In turn, the flake rate allows researchers to compare various algorithms for detecting order-dependent tests [22]. Wei et al. [22] have derived analytical formulas for some cases of victims, namely when all polluters have the same set of cleaners, but have not addressed the most general case, namely when two or more polluters have different sets of cleaners.

We show how to encode the problem of counting test orders in the Alloy modeling language [9], and we use propositional model counters [13, 20] to count the test orders. Alloy has been used for many software analysis and testing tasks [3, 11], and model counters have seen wide applications in various domains [1, 2]. The Alloy toolset automatically translates the Alloy models into propositional formulas that are fed into model counters to solve the counting problems. Yang et al. [23] have presented AlloyMC that connects Alloy with model counters. However, no prior work has used Alloy to count test orders.

We illustrate how Alloy makes it easy to handle even the most general case of victims with polluters that may have different cleaners. We show a general *skeleton* model to encode the problem of counting test orders; the skeleton can be instantiated with the specific sets of polluters and cleaners. To evaluate correctness and scalability of our approach, we use 24 models as our benchmarks. The benchmarks consider a real scenario from the flaky-test dataset published by Wei et al. [22] with two polluters where one has a subset of cleaners of the other. We instantiate our skeleton with an increasing number of cleaners.

We choose Alloy to translate the test-order problem into SAT formula, because the Alloy analyzer uses the highly optimized constraint-solver Kodkod [21], which efficiently translates Alloy models into SAT formulas. The SAT formulas can be counted using any off-the-shelf model counters. We apply state-of-the-art model counters for both exact counting (ProjMC [13]) and approximate counting (ApproxMC4 [20]).

The results show that the problem of counting test orders provides *challenging* propositional formulas for model counters. In addition, we found that the exact counter generally runs *faster* than the approximate counter for all our non-trivial benchmarks, which is a surprising result because it is unusual that an exact model counter outperforms an approximate model counter [20].

In summary, this paper makes the following contributions:

- **Encoding:** We show how to encode the problem of test orders in Alloy.
- **Evaluation:** We evaluate our encoding on a number of challenging problems. The initial results are promising but point out to scalability issues.
- **Challenges:** We obtain a number of interesting and challenging problems for propositional model counters.

```
1.  open util/ordering[Test]

2.  abstract sig Test {}
3.  one sig Victim extends Test {}
4.  abstract sig Cleaner extends Test {}
5.  abstract sig Polluter extends Test { cleaners: set Cleaner }
6.  fact { Polluter.cleaners = Cleaner }
7.  pred Pollutes[p: Polluter] {
8.      p in prevs[Victim]
9.      and no p.cleaners & prevs[Victim] & nexts[p] }
10. pred Fail[] {
11.     some p: Polluter | Pollutes[p]
12.     and no p': nexts[p] & prevs[Victim] & Polluter | Pollutes[p'] }
13. pred Pass[] {
14.     !Fail[] }
15. pred Pass2[] {
16.     all p: Polluter & prevs[Victim] |
17.         some p.cleaners & prevs[Victim] & nexts[p] }

18. one sig c_1, c_2, c_3 extends Cleaner {}
19. one sig p_1, p_2 extends Polluter {}
20. fact Matrix {
21.     p_1.cleaners = c_1 + c_2
22.     p_2.cleaners = c_1 + c_2 + c_3 }

23. run Fail
24. run Pass
25. run Pass2
```

Fig. 1: An example of modeling flaky test orders in Alloy

2 Modeling Flaky Test Orders Using Alloy

We illustrate our approach for modeling flaky test orders in Alloy using an example. Through the example we also introduce the aspects of the Alloy language required to understand the modeling. Figure 1 shows an example Alloy model which encodes the problem of counting test orders.

We order of tests in a test suite using the Alloy library `util/ordering`, which defines a linear order (line 1). The signature (`sig`) `Test` declares a set of atoms that represent tests (line 2). The set of tests is partitioned (using keyword `extends`) into three subsets: a singleton (`one`) set for the victim (line 3), a set of cleaners (line 4), and a set of polluters (line 5). We do not model neutral tests, which have no impact on the victim, because their presence does not affect the flake rate [22]. The *field* `cleaners` in sig `Polluter` introduces a binary relation `cleaners: Polluter x Cleaner` to represent the matrix that relate polluters to cleaners (line 5). A `fact` introduces a constraint that must be satisfied in all models; the stated `fact` uses relational composition ('`.`') to require that the relational

image of `Polluter` under the relation `cleaners` equals the set of all cleaners, i.e., models have no extraneous cleaners (line 6).

A predicate (`pred`) introduces a parameterized formula that can be *invoked* elsewhere. The predicate `Pollutes` enforces two constraints on its parameter `p` that is a polluter (lines 7-9). One, `p` appears before the victim in the test order; `prevs[i]` (likewise, `nexts[i]`) is a library function that represents the set of atoms in the linear order before (likewise, after) `i`, and `in` is the subset operator. Two, no cleaner for `p` is between `p` and the victim; the quantifier `no` is the negation of existential quantifier `some`, and '`&`' is set intersection.

The predicate `Fail` defines the failing test order using the existential quantification: there is some atom in the set of polluters such that it pollutes (lines 10-12); in quantified formulas, '`|`' separates variable bounding from the subformula using the variable. The additional constraint (line 12) requires that `p` be the *last* such test before the victim, which rules out duplicate solutions where the difference is not based on the test order but based on which polluter is a *witness* to failure. (More formally, this constraint ignores from the model the new variable arising from Skolemization.) The predicate `Pass` defines the passing test order as the negation of the constraints for failing test order (lines 13-14). We also evaluate the predicate `Pass2` that defines another encoding for the passing test order, stating more directly that all polluters before the victim have a cleaner between the polluter and the victim; we expect this encoding to enable faster model counting.

The test suites in a general model contain 1 victim, n polluters, and k cleaners. Figure 1 (lines 18-22) shows one example containing 2 polluters and 3 cleaners. The `Matrix` (`fact`) states the cleaners for each polluter. In this example, polluter `p_1` has two cleaners `c_1` and `c_2`, and polluter `p_2` has three cleaners `c_1`, `c_2`, and `c_3`. We can change the number of polluters and cleaners simply by changing the declarations in lines 18-19, and change the relations between polluters and cleaners by changing the `Matrix`. The `run` command (lines 23-25) defines the constraint-solving problem, which is to solve the predicate `Fail`/`Pass`/`Pass2` subject to all applicable constraints on the sets and relations declared in the Alloy model. Each model represents one test order, and counting the number of models thus counts the number of test orders.

3 Experimental Evaluation

```
18. one sig c_1, c_2, ..., c_k extends Cleaner {}
19. one sig p_1, p_2 extends Polluter {}
20. fact Matrix {
21.     p_1.cleaners = c_1 + c_2 + ... + c_(k/2)
22.     p_2.cleaners = c_1 + c_2 + ... + c_k }
```

Fig. 2: The template for our benchmark generation

Table 1: Counting results of our benchmarks ('-' denotes timeout)

Benchmarks	ProjMC		ApproxMC		
	time	count	time	count	error (%)
k=1, Fail	0.01	15	0.00	15	0.00
k=2, Fail	0.02	56	0.01	56	0.00
k=3, Fail	0.09	270	0.21	260	3.85
k=4, Fail	0.87	1800	1.22	1856	3.11
k=5, Fail	6.79	12096	15.27	13056	7.94
k=6, Fail	93.21	104832	204.67	110592	5.49
k=7, Fail	937.31	907200	1040.32	884736	2.54
k=1, Pass	0.01	9	0.00	9	0.00
k=2, Pass	0.02	64	0.01	64	0.00
k=3, Pass	0.13	450	0.45	496	10.22
k=4, Pass	1.23	3240	5.21	3456	6.67
k=5, Pass	13.71	28224	72.80	31744	12.47
k=6, Pass	256.93	258048	547.60	278528	7.94
k=7, Pass	3197.09	2721600	>5000	-	-
k=1, Pass2	0.01	9	0.00	9	0.00
k=2, Pass2	0.02	64	0.01	64	0.00
k=3, Pass2	0.14	450	0.44	496	10.22
k=4, Pass2	1.22	3240	4.74	3456	6.67
k=5, Pass2	12.33	28224	80.05	31744	12.47
k=6, Pass2	217.80	258048	555.24	278528	7.94
k=7, Pass2	2905.76	2721600	4700.48	2359296	15.36

3.1 Setup

Model counters. We study how both exact model counting and approximate model counting perform on the generated propositional formulas. We apply ProjMC [13], which is the state-of-the-art exact model counter, and ApproxMC4 [20], which is the state-of-the-art approximate model counter.

Benchmarks. As our benchmarks, we want to generate propositional formulas for `Fail`, `Pass`, and `Pass2` predicates introduced in the above Alloy model (Figure 1) with various test combinations. To do so, we replace lines 18-22 of our Alloy model with a simple template shown in Figure 2. The template introduces 2 polluters and k cleaners: the first polluter has half of all cleaners as its cleaners,

and the second polluter has all the cleaners as its cleaners. In our experiments, we range k from 1 to 8, generating 8 propositional formulas for each predicate. In total, we generate 24 benchmarks for our evaluation. We choose this template because it resembles a case in the real-world flaky-test dataset by Wei et al. [22].
Metrics. The two key metrics we use in our evaluation are the model counts and the actual wall time to compute them. In line with ApproxMC, we report the error rate of the approximate model counting as $max(\frac{approx}{exact}, \frac{exact}{approx}) - 1$, based on multiplicative guarantees. We use timeout of 5000sec, as commonly done in work on model counting [20].
Platform. All the experiments are conducted on a machine with Intel Core i7-8700k CPU (12 logical cores in total) and 32-GB RAM.

3.2 Results

We apply both model counters on all the generated propositional formulas (with k ranging from 1 to 8) for all the three predicates (i.e., `Fail`, `Pass`, and `Pass2`). Our experimental results show that the limit of ProjMC for all three predicates is $k = 7$; the limit of ApproxMC for `Fail` and `Pass2` is $k = 7$ and for `Pass` is $k = 6$. Thus, the propositional formulas generated with our Alloy model are generally difficult, providing a challenging dataset for future work on model counters. For many other domains, model counters can count orders of magnitude more models [20], so the counting is not limited simply by the number of models.

Table 1 shows the detailed results of the benchmarks encoding all the three predicates with k up to 7. Our results provide a way to sanity check the correctness of our proposed Alloy model: the exact counts of `Pass`/`Pass2` should be the same, and the sum of the exact count of `Fail` and the exact count of `Pass`/`Pass2` should be $(k + 3)!$, the total number of permutations of all the tests (i.e., k cleaners, 2 polluters, and 1 victim). With the exact counts reported by ProjMC, we confirm that our proposed model passes the sanity check. We also manually check that correct models (not just count) are generated for $k = 1$.

Moreover, the results show that ProjMC generally runs *faster* than ApproxMC for all the non-trivial benchmarks ($k \geq 3$), which is quite surprising. It is unusual that an exact model counter outperforms an approximate model counter. Hence, our non-trivial benchmarks pose additional value for model-counting community. Our hypothesis is that SAT formulas for counting test orders offer few opportunities for optimization, not favoring the advanced steps performed by the approximate model counters such as ApproxMC. We also find similar results in our initial experiments using a direct SAT encoding, confirming that the interesting results are not simply due to the encoding done by Alloy.

Besides, we can also observe that ApproxMC sometimes over-approximates and sometimes under-approximates, with the error rate ranging from 0.00% to 15.36%; interestingly, ApproxMC approximates with lower error for the `Fail` predicate than for the `Pass`/`Pass2` predicates. Lastly, the results show that the `Pass` predicate is generally harder to solve than the `Pass2` predicate for both counters. Therefore, we confirm that different encoding for the same problem can result in different counting efficiency.

4 Conclusions

This paper presents a general way of encoding the problem of counting flaky test orders using the Alloy modeling language. We illustrate how Alloy makes it easy to handle even the most general case of victims with polluters that have different cleaners. We provide a general skeleton Alloy model that can be instantiated with the specific sets of polluters and cleaners. To evaluate our encoding, we generate 24 problems of different sizes. The results show that our Alloy encoding provides interesting and challenging propositional formulas that could help advance development of the state-of-the-art model counters.

Our future work is to evaluate more encodings of passing and failing test orders to try to improve scalability of the approach. We hope that we can push the approach to $k = 10$. Such values would cover many real cases [22] and also provide inspiration to develop analytical formulas for the number of test orders.

Acknowledgements

We thank Wing Lam and Anjiang Wei for discussions on counting test orders. This work was partially supported by NSF grants CCF-1718903 and CCF-1763788, and a grant from the Army Research Office accomplished under Cooperative Agreement Number W911NF-19-2-0333. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We also acknowledge support for research on flaky tests from Facebook and Google.

References

1. Abdulbaki Aydin, Lucas Bang, and Tevfik Bultan. Automata-based model counting for string constraints. In *CAV*, 2015.
2. Fahiem Bacchus, Shannon Dalmao, and Toniann Pitassi. Algorithms and complexity results for # SAT and Bayesian inference. In *FOCS*, 2003.
3. Fabian Büttner, Marina Egea, Jordi Cabot, and Martin Gogolla. Verification of ATL transformations using transformation models and model finders. In *ICFEM*, 2012.
4. Google. Avoiding flakey tests, 2008. `http://googletesting.blogspot.com/2008/04/tott-avoiding-flakey-tests.html`.
5. Martin Gruber, Stephan Lukasczyk, Florian Kroiß, and Gordon Fraser. An empirical study of flaky tests in Python. In *ICST*, 2021.
6. Mark Harman and Peter O'Hearn. From start-ups to scale-ups: Opportunities and open problems for static and dynamic program analysis. In *SCAM*, 2018.
7. Kim Herzig, Michaela Greiler, Jacek Czerwonka, and Brendan Murphy. The art of testing less without sacrificing quality. In *ICSE*, 2015.
8. Kim Herzig and Nachiappan Nagappan. Empirically detecting false test alarms using association rules. In *ICSE*, 2015.

9. Daniel Jackson. *Software Abstractions: Logic, Language, and Analysis*. The MIT Press, 2006.
10. He Jiang, Xiaochen Li, Zijiang Yang, and Jifeng Xuan. What causes my test alarm? Automatic cause analysis for test alarms in system and integration testing. In *ICSE*, 2017.
11. Eunsuk Kang and Daniel Jackson. Formal modeling and analysis of a flash filesystem in Alloy. In *ABZ*, 2008.
12. Emily Kowalczyk, Karan Nair, Zebao Gao, Leo Silberstein, Teng Long, and Atif Memon. Modeling and ranking flaky tests at Apple. In *ICSE-SEIP*, 2020.
13. Jean-Marie Lagniez and Pierre Marquis. A recursive algorithm for projected model counting. *AAAI*, 33:1536–1543, 2019.
14. Wing Lam, Patrice Godefroid, Suman Nath, Anirudh Santhiar, and Suresh Thummalapenta. Root causing flaky tests in a large-scale industrial setting. In *ISSTA*, 2019.
15. Wing Lam, Kivanç Muşlu, Hitesh Sajnani, and Suresh Thummalapenta. A study on the lifecycle of flaky tests. In *ICSE*, 2020.
16. Wing Lam, Reed Oei, August Shi, Darko Marinov, and Tao Xie. iDFlakies: A framework for detecting and partially classifying flaky tests. In *ICST*, 2019.
17. Qingzhou Luo, Farah Hariri, Lamyaa Eloussi, and Darko Marinov. An empirical analysis of flaky tests. In *FSE*, 2014.
18. Atif Memon, Zebao Gao, Bao Nguyen, Sanjeev Dhanda, Eric Nickell, Rob Siemborski, and John Micco. Taming Google-scale continuous testing. In *ICSE-SEIP*, 2017.
19. August Shi, Wing Lam, Reed Oei, Tao Xie, and Darko Marinov. iFixFlakies: A framework for automatically fixing order-dependent flaky tests. In *FSE*, 2019.
20. Mate Soos, Stephan Gocht, and Kuldeep S Meel. Tinted, detached, and lazy CNF-XOR solving and its applications to counting and sampling. In *CAV*, 2020.
21. Emina Torlak and Daniel Jackson. Kodkod: A relational model finder. In *TACAS*, 2007.
22. Anjiang Wei, Pu Yi, Tao Xie, Darko Marinov, and Wing Lam. Probabilistic and systematic coverage of consecutive test-method pairs for detecting order-dependent flaky tests. In *TACAS*, 2021.
23. Jiayi Yang, Wenxi Wang, Darko Marinov, and Sarfraz Khurshid. AlloyMC: Alloy meets model counting. In *ESEC/FSE Demo*, 2020.
24. Celal Ziftci and Jim Reardon. Who broke the build?: Automatically identifying changes that induce test failures in continuous integration at Google scale. In *ICSE*, 2017.