Searching for Multi-fault Programs in Defects4J

Gabin An, Juyeon Yoon, and Shin Yoo(✉)

KAIST, Daejeon, Republic of Korea
{agb94,juyeon.yoon,shin.yoo}@kaist.ac.kr

Abstract. Defects4J has enabled numerous software testing and debugging research work since its introduction. A large part of its contribution, and the resulting popularity, lies in the clear separation and distillation of the root cause of each individual test failure based on careful manual analysis, which in turn allowed researchers to easily study individual faults in isolation. However, in a realistic debugging scenario, multiple faults can coexist and affect test results collectively. Study of automated debugging techniques for these situations, such as failure clustering or fault localisation for multiple faults, would significantly benefit from a reliable benchmark of multiple, coexisting faults. We search for versions of Defects4J subjects that contain multiple faults, by iteratively transplanting fault-revealing test cases across Defects4J versions. Out of 326 studied versions of Defects4J subjects, we report that over 95% (311 versions) actually contain from two to 24 faults. We hope that the extended, multi-fault Defects4J can provide a platform for future research of testing and debugging techniques for multi-fault programs.

Keywords: Software faults · Multiple faults · Bug database

1 Introduction

Defects4J [9] is one of the most popular real-world Java fault datasets in the field of software engineering, with over 650 citations as of June 2021 since its publication in 2014. Defects4J provides a number of software faults, along with a clearly separated and isolated set of test cases that can reveal each fault, making it easier for researchers to study individual faults in isolation. Due to both the ease of use and the realism of the curated faults, it has been broadly adopted in the empirical validation of numerous automated debugging work such as Fault Localisation (FL) [2,13,17] and Automated Program Repair (APR) [4,11,15].

However, in realistic debugging scenarios, multiple faults can coexist in software and affect the test results together. For example, a Continuous Integration (CI) process of large-scale industry software can produce hundreds of failing test cases that are caused by distinct root causes [7]. The isolation of individual faults that made Defects4J compatible with the Single Fault Assumption (SFA) ironically prevents it from being used to study the debugging of multiple faults.

© Springer Nature Switzerland AG 2021
U.-M. O'Reilly and X. Devroey (Eds.): SSBSE 2021, LNCS 12914, pp. 153–158, 2021.
https://doi.org/10.1007/978-3-030-88106-1_11

According to a systematic literature review of multiple faults localisation [19], the majority (33) of the 55 selected studies used only C faults for the evaluation. Only ten studies are reported to consider Java programs, five out of which employ Defects4J [12,14,18,20,21]. Only Zheng et al. [21] combined separate multiple Defects4J faults; since the procedure of creating the multiple faults was manual, only 46 have been created. The remaining work either concern multi-hunk faults, i.e., a single fault that can only be fixed by changing multiple locations [16] and consequently use Defects4J as it is [18,20], or actually concern neither multiple faults nor multi-hunk faults [12,14]. Note that, in this paper, we use the term *multiple faults* to denote the faults that can be fixed independently of each other.

Given the contributions to the automated debugging research made by Defects4J under SFA, we believe that the study of automated multi-fault debugging techniques [19], such as failure clustering [5,7,8] or fault localisation for multiple faults [1,6,21], would significantly benefit from the construction of a reliable dataset of realistic multi-fault Java programs. In this paper, we build a **real-world** Java **multi-fault** dataset by extending Defects4J. Instead of artificially injecting mutation or manually grafting faults, we use iterative search to systematically detect the existence of multiple faults in each version via fully automated transplantation and execution of the fault-revealing test cases. We report that 311 out of 326 studied faulty versions (95.4%) contain multiple faults, ranging from two to 24. The result data and replication package are publicly available[1].

2 Proposed Approach

The faults in Defects4J are extracted from the actual development history of various projects. Since every fault has a different life span [3,10], even a fault that was recently fixed may have existed in the project for a long time. In this work, we check if a specific fault N in version P of a Defects4J subject exists in an older version P' containing another fault M. If N exists in P', we regard P' as a multi fault program that includes both N and M. Note that we modify neither P nor P': the check is performed by test transplantation, and therefore we only reveal what already exists in P'. The following sections present the motivating example and our proposed method to search for multi-fault programs.

2.1 A Motivating Example

Listing 1.1 shows the fault Math-5 in Defects4J and its developer patch changing the return value from `NaN` to `INF`.[2] This fault is revealed by the test case `testReciprocalZero` (Listing 1.2) that checks if the return value is equal to `INF`. Each Defects4J fault is similarly provided with a set of fault-revealing test cases that reveals a single fault.

[1] https://github.com/coinse/Defects4J-multifault.

[2] http://program-repair.org/defects4j-dissection/#!/bug/Math/5.

```
--- a/src/main/java/org/apache/commons/math3/complex/Complex.java
+++ b/src/main/java/org/apache/commons/math3/complex/Complex.java
@@ -304,7 +304,7 @@ @@ public Complex reciprocal() {
         if (real == 0.0 && imaginary == 0.0) {
-            return NaN;
+            return INF;
         }
```

Listing 1.1. The developer patch for Math-5

```
public void testReciprocalZero() {
    Assert.assertEquals(Complex.ZERO.reciprocal(), Complex.INF);
    // Error message: junit.framework.AssertionFailedError: expected:<(NaN,
        NaN)> but was:<(Infinity, Infinity)>
}
```

Listing 1.2. The fault-revealing test case of Math-5

We note that, with few exceptions of recently added subjects and versions, the majority of faulty versions in Defects4J are indexed chronologically based on their revision dates, so that a lower ID refers to a more recently fixed fault: for instance, Math-5 was fixed later than Math-6. Therefore, the faulty source code version of Math-6 (referred to as Math-6b) may also contain the fault Math-5. Listing 1.3 confirms that Math-5 does exist in Math-6b, but is simply not revealed due to the absence of the fault-revealing test case, `testReciprocalZero`. When transplanted to Math-6b, the test fails with the same error message as in Math-5b, showing that Math-6b contains at least two faults, Math-5 and Math-6.

2.2 Searching for Multiple Fault Versions

Let B_M be the Defects4J faulty source code version that corresponds to the fault M.[3] As shown in our motivating example, if a fault N is fixed after a fault M, the fault N may *already exist* in B_M. Consequently, to build a multi-fault dataset, we check which faults exist in which preceding faulty versions.

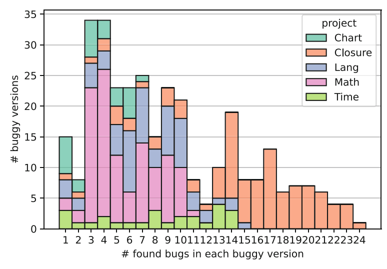
Search Strategy. For each fault N in a project, we sequentially check whether the fault exists in each previous faulty version B_M, such that $M.id > N.id$, from the latest version to the older version. The search stops once N is not revealed in B_M. For example, the fault Lang-3 is revealed in Lang-[4,16]b, but not in Lang-17b. In this case, the search immediately stops and moves to the next iteration with a new N (Lang-4). This is because if B_M does not contain the fault N, it is likely that versions older than B_M do not include N either.

Existence Check. To determine the presence of a fault N in B_M, we *transplant* all fault-revealing tests of N to B_M. We confirm that N exists in B_M if and only if (1) all target test class files to where test case methods are transplanted exist in B_M, (2) all transplanted test cases are successfully compiled and fail against B_M, and (3) the error messages in B_M are the same as those in B_N. If

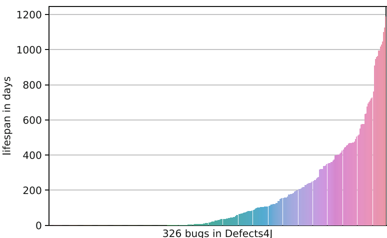
[3] `defects4j checkout -p Math -v 6b -w <dir>` checks out B_{Math-6} into `<dir>`.

```
304         if (real == 0.0 && imaginary == 0.0) {
305             return NaN; // Math-5b
306         }
```

Listing 1.3. In Math-6b, `Complex.java` (line 305) contains the fault Math-5



(a) The number of faulty versions in Defects4J with each number of faults



(b) The sorted life span of faults in days (average=154, standard deviation=246)

Fig. 1. The summary of search results

the fault-revealing test cases of the faults N and M overlap with each other, we further execute the fault-revealing tests of N on the fixed version of M to ensure that the overlapped test cases still fail due to N without the presence of M.

Building Multi-fault Subjects. When the above search is done, we obtain the set of pairs E such that $(N, M) \in E$ if and only if N exists in B_M. For every fault M in Defects4J, the set of *found faults* in B_M, $F(B_M)$, is defined as $F(B_M) = \{M\} \cup \{N|(N, M) \in E\}$. If $|F(B_M)| > 1$, B_M is a multi-fault subject.

2.3 Implementation Details

The process in Sect. 2.2 is dockerised and automated. We use `javaparser`[4] to detect the line range of the target test methods during transplantation. In the docker container, one can simply checkout to the multi-fault version by invoking `python3.6 checkout.py Math-1-2-3 -w /tmp/Math-1-2-3`, after which the *same* source code with Math-3b, augmented with the fault-revealing test cases of Math-1 and Math-2, is checked out.

3 Results

Multiple Fault Subjects. Figure 1a shows how many faults are contained in the faulty versions of five projects[5]. The x-axis shows the number of faults found

[4] https://github.com/javaparser/javaparser.

[5] Defects4J Bug IDs: Lang 1-65, Chart 1-26, Math 1-106, Time 1-27, and Closure 1-106. Note that Lang-2, Time-21, Closure-63 and -93 are excluded since they are either no longer reproducible under Java 8 or the duplicate bugs.

in each faulty version, and the y-axis shows the number of faulty versions. Out of 326 faulty programs, 95.4% (=311/326) of them contain multiple faults (i.e., # found faults >1). Furthermore, 126 and 22 faulty versions have ≥10 and ≥20 faults, respectively. For example, Closure-90b contains 24 faults. Our repository contains the full results of the found multi-fault versions.

Lifespan of Faults. To confirm whether lifespans of Defects4J faults vary similarly to existing findings [3,10], we calculate the lifespan of each fault. Let us define the *lifespan* of fault N as the number of days between the date of the oldest previous faulty version where fault N is detected and the revision date of N when the patch is applied. If there is no preceding version where the fault N is revealed, the lifespan is zero. Figure 1b shows that lifespans of faults range from 0 days up to longer than three years (e.g., Lang-41 has the lifespan of 1,187 days). The variance in lifespan suggests that the probability of having multiple faults at any given time can be nontrivial.

4 Conclusion

The paper presents a multi-fault Java dataset based on Defects4J, for which subjects with multiple real faults are constructed by transplanting tests without modifying the source code. Exploiting the chronological indexing of Defects4J, we propose a systematic search strategy to find co-existing faults that have not yet been revealed by failing tests. The results show that 311 out of 326 versions in Defects4J actually contain multiple faults. We hope that our extension of Defects4J can aid future research on search-based automated debugging under the existence of multiple faults.

Acknowledgement. This work is supported by National Research Foundation of Korea (NRF) Grant (NRF-2020R1A2C1013629), Institute for Information & communications Technology Promotion grant funded by the Korean government (MSIT) (No.2021-0-01001), and Samsung Electronics (Grant No. IO201210-07969-01).

References

1. Abreu, R., Zoeteweij, P., Van Gemund, A.J.: Spectrum-based multiple fault localization. In: 2009 IEEE/ACM International Conference on Automated Software Engineering, pp. 88–99. IEEE (2009)
2. Le, T.D.B., Lo, D., Le Goues, C., Grunske, L.: A learning-to-rank based fault localization approach using likely invariants. In: Proceedings of the 25th International Symposium on Software Testing and Analysis, pp. 177–188 (2016)
3. Canfora, G., Ceccarelli, M., Cerulo, L., Di Penta, M.: How long does a bug survive? An empirical study. In: Proceedings of Working Conference on Reverse Engineering, pp. 191–200. IEEE (2011)
4. Chen, Z., Kommrusch, S.J., Tufano, M., Pouchet, L.N., Poshyvanyk, D., Monperrus, M.: Sequencer: sequence-to-sequence learning for end-to-end program repair. IEEE Trans. Softw. Eng. (2019)

5. Dang, Y., Wu, R., Zhang, H., Zhang, D., Nobel, P.: ReBucket: a method for clustering duplicate crash reports based on call stack similarity. In: 2012 34th International Conference on Software Engineering (ICSE), pp. 1084–1093. IEEE (2012)
6. Ghosh, D., Singh, J.: Spectrum-based multi-fault localization using chaotic genetic algorithm. Inf. Softw. Technol. **133**, 106512 (2021)
7. Golagha, M., Lehnhoff, C., Pretschner, A., Ilmberger, H.: Failure clustering without coverage. In: Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 134–145 (2019)
8. Jones, J.A., Bowring, J.F., Harrold, M.J.: Debugging in parallel. In: Proceedings of the International Symposium on Software Testing and Analysis, pp. 16–26 (2007)
9. Just, R., Jalali, D., Ernst, M.D.: Defects4J: a database of existing faults to enable controlled testing studies for Java programs. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis, pp. 437–440 (2014)
10. Kim, S., Whitehead Jr., E.J.: How long did it take to fix bugs? In: Proceedings of the International Workshop on Mining Software Repositories, pp. 173–174 (2006)
11. Koyuncu, A., et al.: FixMiner: mining relevant fix patterns for automated program repair. Empirical Softw. Eng. **25**, 1–45 (2020)
12. Laghari, G., Murgia, A., Demeyer, S.: Fine-tuning spectrum based fault localisation with frequent method item sets. In: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, pp. 274–285 (2016)
13. Li, X., Li, W., Zhang, Y., Zhang, L.: DeepFL: integrating multiple fault diagnosis dimensions for deep fault localization. In: Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 169–180 (2019)
14. Li, X., d'Amorim, M., Orso, A.: Iterative user-driven fault localization. In: Bloem, R., Arbel, E. (eds.) HVC 2016. LNCS, vol. 10028, pp. 82–98. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49052-6_6
15. Liu, K., Koyuncu, A., Kim, D., Bissyandé, T.F.: TBar: revisiting template-based automated program repair. In: Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 31–42 (2019)
16. Saha, S., et al.: Harnessing evolution for multi-hunk program repair. In: International Conference on Software Engineering, pp. 13–24 (2019)
17. Sohn, J., Yoo, S.: FLUCCS: using code and change metrics to improve fault localization. In: Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 273–283 (2017)
18. Xia, X., Bao, L., Lo, D., Li, S.: "Automated debugging considered harmful" considered harmful: a user study revisiting the usefulness of spectra-based fault localization techniques with professionals using real bugs from large systems. In: International Conference on Software Maintenance and Evolution, pp. 267–278 (2016)
19. Zakari, A., Lee, S.P., Abreu, R., Ahmed, B.H., Rasheed, R.A.: Multiple fault localization of software programs: a systematic literature review. Inf. Softw. Technol. **124**, 106312 (2020)
20. Zhang, M., Li, X., Zhang, L., Khurshid, S.: Boosting spectrum-based fault localization using pagerank. In: Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 261–272 (2017)
21. Zheng, Y., Wang, Z., Fan, X., Chen, X., Yang, Z.: Localizing multiple software faults based on evolution algorithm. J. Syst. Softw. **139**, 107–123 (2018)