Appdendix. *Multi-Level Random Walk for Software Test Suite Reduction*

Zongzheng Chi, School of Software, Dalian University of Technology, Dalian, CHINA
Jifeng Xuan, State Key Lab of Software Engineering, Wuhan University, Wuhan, CHINA
Zhilei Ren, School of Software, Dalian University of Technology, Dalian, CHINA
Xiaoyuan Xie, State Key Lab of Software Engineering, Wuhan University, Wuhan, CHINA
He Guo, School of Software, Dalian University of Technology, Dalian, CHINA

I. Appendix

We conduct the Mann-Whitney U test [1] on the results to compare the differences between MultiWalk with the other six algorithms.Table I shows the detailed p-values in comparison. Among all comparisons but one, MultiWalk is statistically significant ($p\text{-}value < 0.01$), comparing with other algorithms. The result of the Mann-Whitney U test shows that the size difference of test suites by MultiWalk is acceptable.

As mentioned in [2], the capability of fault detection is another measurement of test suite reduction. In our work, we count the loss of the fault detection rate. In each project, we randomly select 300 faults based on the program mutation technique [3]. Faults in a project are randomly seeded by slightly changing the original source code; the program mutation technique leverages such changes to generate faulty source code and records the tests that can detect the faulty one.

Table II show the loss of the fault detection rate by algorithms in comparison. Algorithms of test suite reduction slightly lose the fault detection rate on most of the projects. MultiWalk loses less than 5% of the fault detection rate after test suite reduction on 8 out of 10 projects. Note that this paper focuses on evaluating test suite reduction with the size change of test suites and we do not further explore the loss of fault detection rate.

A local optima can be affected by the trials of random walk in WalkTest. We further evaluate the size of test suites by changing the number of trials of WalkTest. Fig. 1 takes the project

Corresponding Author: Jifeng Xuan (jxuan@whu.edu.cn)

TABLE I

P-VALUE IN THE MANN-WHITNEY U TEST ON THE SIZE OF TEST SUITES BETWEEN MULTIWALK AND OTHER ALGORITHMS ON TEN LARGE PROJECTS

Project	ILP	GRE	HGS	RAPS	RWLS	GA	WalkTest
Camel	3.4e-13	3.4e-13	3.4e-13	3.4e-13	3.4e-13	3.8e-10	3.4e-13
AssertJ	3.4e-13	3.4e-13	3.4e-13	1.7e-14	3.4e-13	8.0e-13	3.4e-13
Configuration	3.4e-13	3.4e-13	1.7e-14	6.2e-04	3.4e-13	8.0e-13	3.4e-13
JGit	3.4e-13	3.4e-13	3.4e-13	1.7e-14	1.7e-14	8.0e-13	1.7e-14
Closure	8.0e-13	8.0e-13	8.0e-13	1.7e-14	1.7e-14	3.4e-13	3.4e-13
Collections	-	1.6e-11	8.0e-13	1.7e-02	3.4e-13	1.6e-11	1.6e-11
JFreeChart	-	8.0e-13	3.4e-13	1.7e-14	3.4e-13	3.4e-13	3.4e-13
Lang	3.9e-12	3.9e-12	3.9e-12	6.2e-04	3.4e-13	8.1e-12	3.9e-12
JodaTime	-	8.1e-12	3.4e-13	8.1e-12	8.1e-12	8.1e-12	8.1e-12
Math	-	8.1e-12	8.1e-12	3.9e-12	8.1e-12	8.1e-12	8.1e-12

TABLE II

LOSS OF THE FAULT DETECTION RATE FOR MULTIWALK AND THE OTHER ALGORITHMS ON TEN LARGE PROJECTS IN PERCENT

Project	MultiWalk	ILP	GRE	HGS	RAPS	RWLS	GA	WalkTest
Camel	0.00	0.83	1.11	1.11	0.67	1.39	0.11	1.11
AssertJ	3.67	3.89	3.00	3.33	4.22	4.35	2.56	3.67
Configuration	0.78	1.48	1.00	0.89	1.00	1.67	0.89	0.89
JGit	3.00	3.33	2.67	2.00	2.56	2.96	3.00	2.11
Closure	2.67	3.06	2.00	3.00	2.78	3.15	2.78	2.11
Collections	12.56	-	10.33	12.56	12.33	11.11	11.22	11.44
JFreeChart	0.00	-	0.00	0.00	0.11	0.93	0.00	0.00
Lang	3.00	3.33	2.56	3.78	3.56	3.80	2.78	2.56
JodaTime	7.44	-	8.44	6.44	7.67	7.22	7.89	9.11
Math	4.78	-	3.56	3.67	3.67	3.89	2.89	3.67

JodaTime as an example to illustrate the size of test suites. The size of test suites of a solution is measured by the difference with the best known solution: given a solution S and a known best solution O of the original test suite T, the difference is defined as $difference(S, O) = |S| - |O|$.

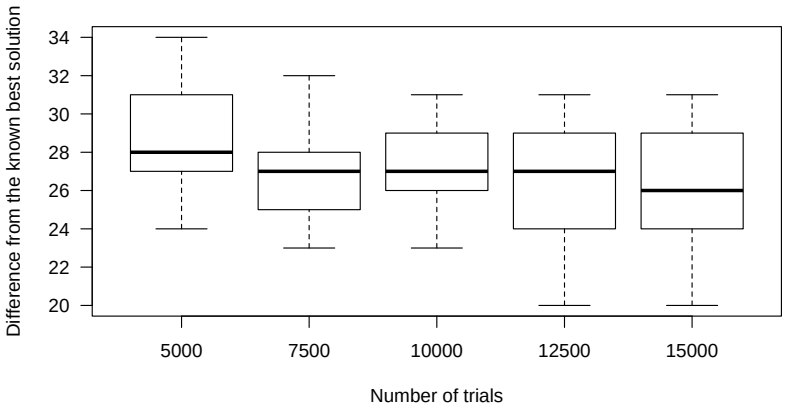

Fig. 1. Size of test suites by changing the number of trials of WalkTest in JodaTime.

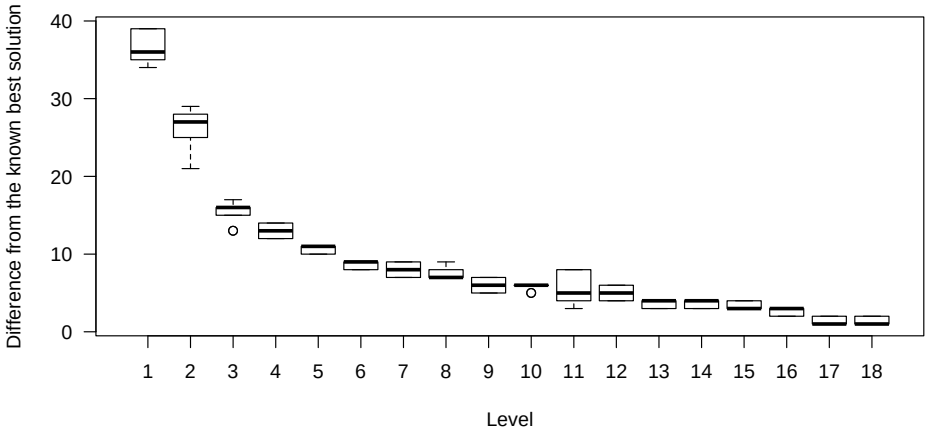

Fig. 2. Size of test suites by changing the levels of MultiWalk in JodaTime.

As shown in Fig. 1, the size of test suites is insensitive with the number of trials in random walk. The size of test suite slightly decreases while the number of trials increases. In our work, we choose a large one, i.e., 10,000 trials in WalkTest, as in [2].

We further analyze the size of test suites by MultiWalk when changing the maximum number of levels, i.e., α. Fig. 2 illustrates the size of test suites with the project JodaTime as an example. Similar to Fig. 1, we evaluate the size of test suites via measuring the difference with the best known solution. As shown in Fig. 2, the size of test suites by MultiWalk decreases while the maximum number of levels increases.

REFERENCES

[1] L. Dinneen and B. Blakesley, "Algorithm as 62: A generator for the sampling distribution of the mann-whitney u statistic," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 22, no. 2, pp. 269–273, May 1973.

[2] Z. Chi, J. Xuan, Z. Ren, X. Xie, and H. Guo, "Multi-level random walk for software test suite reduction," http://cstar.whu.edu.cn/p/multi-walk/techreport.pdf, State Key Lab of Software Engineering, Wuhan Unviersity, Tech. Rep., 2017.

[3] J. Xuan and M. Monperrus, "Test case purification for improving fault localization," in *Proc. of 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. Hong Kong, China: ACM, November 16–22, 2014, pp. 52–63.