Extending Big Data Fuzz Testing with Coverage Guidance

Author: Bo van den Berg b.vandenberg-6@student.tudelft.nl

Supervisor: Burcu Özkan b.ozkan@tudelft.nl

TU Delft

1 Background

DISC Systems - Data-Intensive Scalable Computing systems are often used for handling large data. Rare and buggy corner cases are often encountered.

Fuzz Testing - An automated software testing technique: Automatically generate malformed inputs, and see if tbreaks things.

Big Data Testing - Hard to apply traditional fuzzing, because:

1. DISC systems have long latence
2. most code comes from the framework implementation
3. random mutations rarely generate valid data

2 Aim

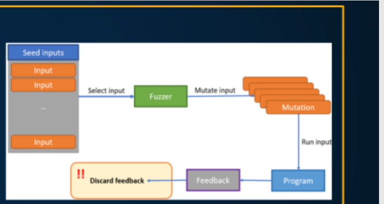
How does input selection based on coverage affect the performance of fuzz testing big data applications?

- How is the coverage information currently used by big data fuzzers?
- How can big data input selection be improved based on coverage information?
- How does the extended fuzzer compare to the current fuzzer?

3 Method

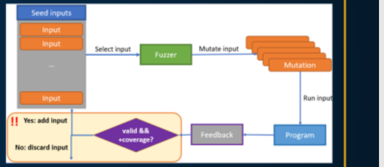
0. **Black-box fuzzer**

Does not use coverage information



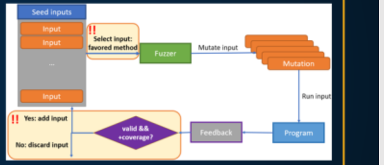
1. **Grey-box fuzzer**

Uses coverage information



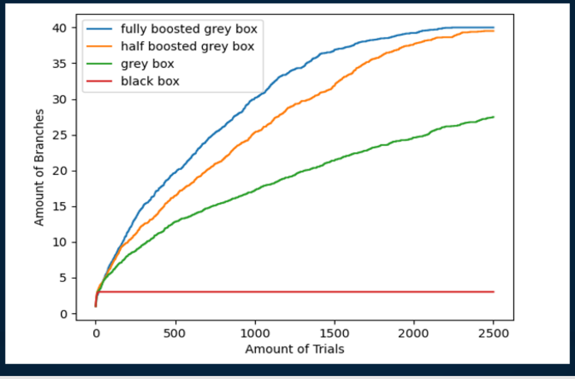
2. **Boosted grey-box fuzzer**

Uses coverage information
And favored input selection



4 Results

Coverage Improvement of Boosted Greybox Fuzzer Over Traditional Greybox- and Blackbox Fuzzer



5 Conclusion

- ➢ Both perform at least as good as black-box fuzzing on error detection
- ➢ Both extensions allow coverage exploration
- ➢ Boosted grey-box fuzzing is most efficient