UNIFUZZ: A Holistic and Pragmatic Metrics-Driven Platform for Evaluating Fuzzers

Yuwei Li, Shouling Ji, Yuan Chen, Sizhuang Liang, Wei-Han Lee, Yueyao Chen, Chenyang Lyu
Chunming Wu, Raheem Beyah, Peng Cheng, Kangjie Lu, Ting Wang

USENIX Security 2021
A plethora of fuzzing works have emerged in both industry and academia.

GitHub has over 3,000 fuzzing related repos.

Dblp, a famous computer science bibliography website, contains more than 200 fuzzing related papers since 2010.
Open questions about fuzzing technique

- How do these fuzzers perform in practice?
- How to compare different fuzzers under a fair and comprehensive set of performance metrics?
- Which fuzzing primitives or techniques are promising and should be promoted?
The previous works cannot answer these questions

- Many existing works do not conduct appropriate and sufficient experiments to provide trustworthy results.
  - Insufficient repetitions, not using statistical test
  - Inconsistency of environments
The previous works cannot answer these questions

- Many existing works do not conduct appropriate and sufficient experiments to provide trustworthy results.
  - Insufficient repetitions, not using statistical test
  - Inconsistency of environments

- The evaluations of the existing fuzzers are often biased due to the lack of uniform benchmarks.
  - The choice varies widely.

- The existing metrics may not be suitable nor comprehensive for evaluating fuzzers.
  - Never consider “overhead”
Challenges for conducting comprehensive evaluations

- Challenge 1: usability issues of the existing fuzzers
- Challenge 2: lack of pragmatic real-world benchmarks
- Challenge 3: lack of proper and comprehensive performance metrics
Our Solution
UNIFUZZ: a holistic and pragmatic metrics-driven platform for evaluating fuzzers

Overview of UNIFUZZ
We conducted large-scale tests on the usability of the existing fuzzers.

- **15+ serious flaws**
- **35+ usable fuzzers**
  - Dockerfile
  - Detailed documents
Pragmatic benchmark suite

<table>
<thead>
<tr>
<th>Type</th>
<th>Program</th>
<th>Version</th>
<th>Arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>exiv2</td>
<td>0.26</td>
<td>@@</td>
</tr>
<tr>
<td></td>
<td>gdk-pixbuf-pixdata (gdk)</td>
<td>gdk-pixbuf 2.31.1</td>
<td>@@ /dev/null</td>
</tr>
<tr>
<td></td>
<td>imginfo</td>
<td>jasper 2.0.12</td>
<td>-f @@</td>
</tr>
<tr>
<td></td>
<td>jhead</td>
<td>3.00</td>
<td>@@</td>
</tr>
<tr>
<td></td>
<td>tiffsplit</td>
<td>libtiff 3.9.7</td>
<td>@@</td>
</tr>
<tr>
<td>Audio</td>
<td>lame</td>
<td>lame 3.99.5</td>
<td>@@ /dev/null</td>
</tr>
<tr>
<td></td>
<td>mp3gain</td>
<td>1.5.2-r2</td>
<td>@@</td>
</tr>
<tr>
<td></td>
<td>wav2swf</td>
<td>swftools 0.9.2</td>
<td>-o /dev/null @@</td>
</tr>
<tr>
<td>Video</td>
<td>ffmpeg</td>
<td>4.0.1</td>
<td>(-y -i @@ -c:v mpeg4 -c:a copy -f mp4 /dev/null)</td>
</tr>
<tr>
<td></td>
<td>flvmeta</td>
<td>1.2.1</td>
<td>@@</td>
</tr>
<tr>
<td></td>
<td>mp42aac</td>
<td>Bento4 1.5.1-628</td>
<td>@@ /dev/null</td>
</tr>
<tr>
<td></td>
<td>cflow</td>
<td>1.6</td>
<td>@@</td>
</tr>
<tr>
<td></td>
<td>infotocap</td>
<td>ncurses 6.1</td>
<td>-o /dev/null @@</td>
</tr>
<tr>
<td></td>
<td>jq</td>
<td>1.5</td>
<td>. @@</td>
</tr>
<tr>
<td></td>
<td>msys</td>
<td>1.0.2</td>
<td>@@</td>
</tr>
<tr>
<td></td>
<td>pdftotext</td>
<td>xpdf 4.00</td>
<td>@@ /dev/null</td>
</tr>
<tr>
<td></td>
<td>sqlite3</td>
<td>3.8.9</td>
<td>(sidin)</td>
</tr>
<tr>
<td>Binary</td>
<td>nm</td>
<td>binutils 5279478</td>
<td>(-A -a -l -S -s \ - -special-syms \ - -synthetic \ - -with-symbol-versions \ - -D @@)</td>
</tr>
<tr>
<td></td>
<td>objdump</td>
<td>binutils 2.28</td>
<td>-S @@</td>
</tr>
<tr>
<td>Network</td>
<td>tcpdump</td>
<td>4.8.1 + libpcap 1.8.1</td>
<td>-e -vv -nr @@</td>
</tr>
</tbody>
</table>

- **20 real-world benchmark programs**
  - 6 functionality types
  - 12+ vulnerability types

- **convenient offline results analysis**
  - bug triage
  - severity analysis
  - CVE matching
Six categories of performance metrics

- Quantity of unique bugs
  - Statistical test
- Quality of the bugs
  - Severity of the bugs, rareness of the bugs
- Speed of finding the bugs
- Stability of finding the bugs
- Coverage
- Overhead
Evaluations
We conducted large-scale evaluations on the state-of-the-art fuzzers.

- 8 state-of-the-art fuzzers: AFL, AFLFast, Angora, Honggfuzz, MOPT, QSYM, T-Fuzz, VUzzer64.
- large-scale: 200,000+ CPU hours
- 6 categories of performance metrics
Summary of interesting findings

- No fuzzer outperformed the others among all the benchmark programs.
  - Fuzzers may have preference over some specific programs.
- The synthetic benchmark programs may not be able to reflect a fuzzer’s performance on the real-world programs.
- A single metric may lead to unilateral conclusions.
- More factors can affect the fuzzing evaluation results than what we thought.
Summary of interesting findings

- No fuzzer outperformed the others among all the benchmark programs.
  - Fuzzers may have preference over some specific programs.

- The synthetic benchmark programs may not be able to reflect a fuzzer’s performance on the real-world programs.

- A single metric may lead to unilateral conclusions.

- More factors can affect the fuzzing evaluation results than what we thought.
No fuzzer outperformed the others among all the benchmark programs.

Angora performed the best on *exiv2*.

QSYM performed the best on *tcpdump*.

The number of unique bugs found by each fuzzers.
Summary of interesting findings

- No fuzzer outperforms the others among all the benchmark programs.
  - Fuzzers have preference over some specific programs.
- The synthetic benchmark programs may not be able to reflect a fuzzer’s performance on the real-world programs.
- A single metric may lead to unilateral conclusions.
- More factors can affect the fuzzing evaluation results than what we thought.
Synthetic benchmark VS. Real-world benchmark

The #unique bugs on the real-world programs.

T-Fuzz and VUzzer64 had better performance on the synthetic benchmark programs than on the real-world benchmark programs.

The #unique bugs on the synthetic programs (LAVA-M).
Summary of interesting findings

- No fuzzer outperforms the others among all the benchmark programs.
  - Fuzzers have preference over some specific programs.
- The synthetic benchmark programs may not be able to reflect a fuzzer’s performance on the real-world programs.
- A single metric may lead to unilateral conclusions.
- More factors can affect the fuzzing evaluation results than what we thought.
A single metric may lead to unilateral conclusions.

For **#unique bugs** metric, QSYM performed the best on *jhead*.
For **#line coverage** metric, Angora performed the best on *jhead*.
Summary of interesting findings

- No fuzzer outperforms the others among all the benchmark programs.
  - Fuzzers have preference over some specific programs.
- The synthetic benchmark programs may not be able to reflect a fuzzer’s performance on the real-world programs.
- A single metric may lead to unilateral conclusions.
- More factors can affect the fuzzing evaluation results than what we thought.
Fuzzers usually have different instrumentation methods.

- Compile-time instrumentation, e.g., AFL, Angora.
- Dynamic binary instrumentation, e.g., VUzzer.

Thus, the same tested benchmark program are compiled into different binaries!

We found that Angora cannot find some bugs on the program infotocap due to its instrumentation method, not its capability in finding bugs.
Fuzzers usually have different instrumentation methods.
- Compile-time instrumentation, e.g., AFL, Angora.
- Dynamic binary instrumentation, e.g., VUzzer.

Thus, the same tested benchmark program are compiled into different binaries!

We found that Angora cannot find the bugs on the program *infotocap* due to its instrumentation method, not its capability in finding bugs.

Using *cross validation* when analyzing the crash samples, e.g., re-executing the crash samples with different complied binaries to check whether these crash samples can be reproduced on all the binaries.
Factor2: crash analysis tools

- Different crash analysis tools are used in validating the bugs triggered by the crash samples.

Using different analysis tools may lead to different evaluation results, e.g., #unique bugs.

Table 13: Validated crash samples by different tools.

<table>
<thead>
<tr>
<th>Bug Type</th>
<th>Number</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither ASAN or GDB can validate</td>
<td>40,122</td>
<td>12.2%</td>
</tr>
<tr>
<td>Only GDB can validate</td>
<td>47,910</td>
<td>14.5%</td>
</tr>
<tr>
<td>Only ASAN can validate</td>
<td>40,267</td>
<td>12.2%</td>
</tr>
<tr>
<td>Both ASAN and GDB can validate</td>
<td>201,558</td>
<td>61.1%</td>
</tr>
<tr>
<td>Total</td>
<td>329,857</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 11: The number of unique bugs discovered on ffmpeg with GDB and ASan.

When using ASAN to validate the crashes, the result is that only Honggfuzz can discover bugs.
Different crash analysis tools are used in validating the bugs triggered by the crash samples. Using different analysis tools may lead to different evaluation results, e.g., #unique bugs.

Using multiple analysis tools may mitigate these biases!

When using ASAN to validate the crashes, the result is that only Honggfuzz can discover bugs.

Table 13: The number of unique bugs discovered on ffmpeg with GDB and ASan.

<table>
<thead>
<tr>
<th>Bug Type</th>
<th>Total</th>
<th>#Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither ASAN nor GDB validate</td>
<td>1,588</td>
<td>25.7%</td>
</tr>
<tr>
<td>Only GDB validate</td>
<td>92,178</td>
<td>58.2%</td>
</tr>
<tr>
<td>Only ASAN validate</td>
<td>74,580</td>
<td>44.5%</td>
</tr>
<tr>
<td>Both ASAN and GDB can validate</td>
<td>201,558</td>
<td>61.1%</td>
</tr>
<tr>
<td>Total</td>
<td>329,857</td>
<td>100%</td>
</tr>
</tbody>
</table>
Conclusion
We proposed and implemented UNIFUZZ, a holistic, and pragmatic metrics-driven platform for evaluating fuzzers in a comprehensive and fair manner.

UNIFUZZ has incorporated 35 usable fuzzers, 20 real-world benchmark programs and 6 categories of performance metrics.

We conducted extensive evaluations on the 8 state-of-the-art fuzzers and got many interesting findings.

We have open sourced UNIFUZZ to facilitate the future fuzzing research and welcome the community contributions.
https://github.com/unifuzz/overview