A Replication of results initially achieved by FlowDroid for DroidBench

Authors of original paper (FlowDroid - PLDI 2014)
Steven Arzt, Alexandre Bartel, Eric Bodden, Christian Fritz, Jacques Klein, Yves Le Traon, Patrick McDaniel, Damien Octeau, Siegfried Rasthofer

Authors of replication paper (ReproDroid - ESEC/FSE 2018)
Felix Pauck[1], Eric Bodden, Heike Wehrheim

[1]fpauck@mail.uni-paderborn.de

As steady as the number of devices and users of the Android operating system grows, new versions of it are released. With each release Android app analysis tools may have to be updated. This presents a huge insuperable challenge for many teams developing, for example, taint analysis tools. However, the team behind FlowDroid [1] frequently updates their tool. Not least because of that, it has become a reference in the field of Android taint analyses. Unsurprisingly its proposing paper is cited almost 250[1] times just in the ACM digital library. Nonetheless, the results proposed had never officially been replicated so far.

At the time the FlowDroid paper was published, no Android taint analyses benchmark existed. Thus, the authors developed one themselves: DroidBench[2]. Since then, this benchmark has been further extended by the community. It nowadays consists of 190 test cases comprising 19 categories such as Aliasing, Callbacks, Lifecycle or Inter-App Communication. Each of these test cases contains one or more expected results, for example, a taint flow originating from a sensitive source (e.g. the device's serial number) to a sink (e.g. sending a text message). Any analysis tool competing in the area of Android taint analyses attempts to beat FlowDroid in finding such privacy leaks. More precisely, to find more leaks (true positives) while producing less false warnings (false positives).

The structure of DroidBench has not changed since its first release. Hence, DroidBench possesses some limitations that hinder comparability and automatic-execution. These limitations have been overcome by ReproDroid, an Android benchmark reproduction framework proposed in our ESEC/FSE'18 paper [2].

During the evaluation of ReproDroid, as a side result, we replicated the DroidBench results that were proposed in the original FlowDroid paper. Therefore, FlowDroid has been executed automatically by means of ReproDroid. To evaluate the outcome, the metrics precision, recall and F-measure have been computed. Arzt et al. claimed, for instance, to reach an F-measure of 89% (see Section 6.1 in [1]) for a certain subset of the DroidBench test cases. With a reproduced F-measure value of 90% (see Section 5.1 in [2]) for the same set this claim could be confirmed. Thus, the manually performed evaluation could be replicated with ReproDroid. Minor results reported for an unspecified set of real-world apps and for an today unavailable vulnerable app could not and generally cannot be replicated, thus all suitable results have been replicated.

In the future, only replicable benchmark versions, such as the refined version of DroidBench[3], should be used. Thereby the comparability of results as well as a fair and unbiased evaluation is guaranteed.

[1] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, and Patrick D. McDaniel. Flowdroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. In Michael F. P. O'Boyle and Keshav Pingali, editors, *ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '14, Edinburgh, United Kingdom - June 09 - 11, 2014*, pages 259–269. ACM, 2014.

[2] Felix Pauck, Eric Bodden, and Heike Wehrheim. Do android taint analysis tools keep their promises? *CoRR*, abs/1804.02903, 2018.

[1] `https://dl.acm.org/citation.cfm?id=2594299`
[2] `https://github.com/secure-software-engineering/DroidBench/tree/develop`
[3] `https://FoelliX.github.io/ReproDroid/#droidbench`