Approach for Estimating Similarity between Procedures in Differently Compiled Binaries

S. Stojanović, Z. Radivojević, M. Cvetanović

Abstract:
Context: Detection of an unauthorized use of a software library is a clone detection problem that in case of commercial products has additional complexity due to the fact that only binary code is available.
Objective: The goal of this paper is to propose an approach for estimating the level of similarity between the procedures originating from different binary codes. The assumption is that the clones in the binary codes come from the use of a common software library that may be compiled with different toolsets.
Method: The approach uses a set of software metrics adapted from the high level languages and it also extends the set with new metrics that take into account syntactical changes that are introduced by the usage of different toolsets and optimizations. Moreover, the approach compares metric values and introduces transformers and formulas that can use training data for production of measure of similarities between the two procedures in binary codes. The approach has been evaluated on programs from STAMP benchmark and BusyBox tool, compiled with different toolsets in different modes.
Results: The experiments with programs from STAMP benchmark show that detecting the same procedures recall can be up to 1.44 times higher using new metrics. Knowledge about the used compiling toolset can bring up to 2.28 times improvement in recall. The experiment with BusyBox tool shows 43% recall for 43% precision.
Conclusion: The most useful newly proposed metrics are those that consider the frequency of arithmetic instructions, the number and frequency of occurrences for instructions, and the number of occurrences for target addresses in calls. The best way to combine the results of comparing metrics is to use a geometric mean or when previous knowledge is available, to use an arithmetic mean with appropriate transformer.

Keywords:
Software clone, Software metric, Binary code analysis, Semantic clone, Clone detection

References: