Dataset Research Proposal

Francis Birck Moreira

November 8, 2018

1 Dataset Definition

In my work proposal I plan to analyze a dataset which represents the behavior of program phases. Each program phase is identified by the "Critical Basic Block Transition" which identifies it. For each phase, the number of Basic Blocks executed is recorded, as well as how many unique Basic Blocks were executed. As shown elsewhere, there is a strong relationship between sequences of CBBTs. Therefore, the table has the following variables:

- Phase Identification, which consists of a pair of unsigned long integer representing the first address of the source BBL and the first address of the destination BBL of the CBBT.
- Number of Basic Block executions.
- Number of Unique Basic Block executions.
- Next Phase Identification, which consists of the same data type of Phase identification. It is used to represent what is the following phase. The last phase receives a value of $N0$, where N is the last BBL executed.

I have 20 datasets for each program executed, where each dataset represents how the phases behaved in executions with each input. I executed a total of 8 programs.

2 Questions Definition

What I want to know, generally, is whether the selected dataset features are stable long the inputs. Can a subset of the datasets represent what happens in all of them? How reliably can CBBTs identify whether a program is behaving normally?

To do this, I randomly select a subset of 5 inputs, S, and ask:

- Is there any phase in all of the inputs with a different combination of "Number of Basic Block executions" and "Number of Unique Basic Block executions" which is NOT found in S?

- Is there any phase sequence P1 - P2 in all of the inputs which is not seen in S?
- Is there any combination or correlation of the features which can reliably define normal behavior for all inputs?

I make these questions in the context of security research, as I already have shown that these characteristics can detect attacks such as Heartbleed and Shellshock. This dataset analysis aims to show and/or improve my research to avoid false positives.