Identifying ATTCK Tactics in Android Malware Control Flow Graph Through Graph Representation Learning and Interpretability

Cameron Noakes

February 2022

1 Sentence Overview

The literature provides an automation methodology to locate TTP in the control flow graph that describes the execution flow of a malware executable. This methodology that comes with this proposed solution can start to merge graph representation learning and machine learning explanations.

The researchers propose an approach to help identify ATTCK tactics, techniques and procedures (TTPs) in a Control Flow Graph (CFG) by applying techniques from Graph Machine Learning to Android Malware applications. Using the Graph Neural Network and SIR-GN node representation learning approach to process the CFG.

Comparisons of the procedures are done by combining SIR-GN and standard classification models with Graph Attention Networks (GAN).

2 Introduction

A Control Flow Graph (CFG) describes the actions of a specific program during the execution of the application and can examine internal and external function calls that are called my the potential malicious application (malware).

The ATTCK framework is used to locate TTP in the control flow graph and is also mapped to malware analysis to help provide significant mitigation steps since the researchers outline how it is crucial to have a deep understanding of

malware and what the application does.

The researchers identified the aim of this literature which was to use the control flow graph (CFG) to identify which subset of the actions in the graph have a high likelihood of being responsible for the tactics, techniques and procedures.

The researchers developed a proposition as an approach to identify ATTCK TTPs in a Control Flow Graph (CFG) by applying Graph Machine Learning techniques on Android Malware applications.

To develop the CFG the researchers made use of graph neural networks and node representation.

3 Related Works

The first mentioned literature cited detailed the use of machine learning and a control flow graph to correctly identify if an application was malicious but does not identify where the part of the CFG that represent these techniques within ATTCK. This literature gave a solid foundation to the research conducted and developing this papers CFG.

CFG definition: A control flow graph (CFG) gives a graphical representation of the overall structure of a given program, this is done by graphing the functions that the program executes. The CFG represents how each function interacts with the other functions.

4 Methodology

The methodology described in the research and proposed solution of the CFG graphing system that was developed by the researchers has 3 different segments.

- Data collection and processing for training a graph.
- SIR-GN graph representation learning procedure.
- An attribution procedure that explains the classification of tasks.

The researchers needed a way to identify the android APK applications as malicious and thus used the Virus Total API to collect the human-curated lists of TTP from the ATTCK framework for each uploaded application and then convert each android malware into its corresponding control flow graph (CFG).

A sandbox is used to collect the android malware TTPs from the ATTCK framework as the sandbox allows the raw APK file samples to be passed in, and the TTPs extracted from ATTCK.

5 Conclusion

The proposed solution is an automated procedure based on graph representation learning and interpretation used to detect their related subgraphs in the CFG and map to ATTCK TTPs.

6 What problem does this solve?

The problem outlined and expressed is one of how malware executable files need a description of how to map/locate TTPs in malware scenarios from the ATTCK framework.

7 How do they solve it?

The proposed research and solution development solves the outlined problem by having an automation methodology to locate TTP in the control flow graph that describes the execution flow of a malware file.

8 How do they assess it?

The assessment for the overall functionality of the proposed solution was given by comparisons being made in terms of an F1-score and accuracy in classifying the TTP. The F1-score is a measurement evaluating classification performances that is robust to unbalanced data.