Using a GCN to Classify Unknown Binaries with Extracted Control Flow Graph Data for Familial Attribution

Micah Flack  
CSC-786 “Cyber Problems”  
Dakota State UniversityMadison, South Dakota  
micah.flack@trojans.dsu.edu  
  
Rita Foster  
Infrastructure Security  
*Idaho National Laboratory*Idaho Falls, Idaho  
rita.foster@inl.gov

*Abstract*— The ability to accurately predict the threats that unknown binaries propose to an environment is an ongoing issue within cyber security. Malware analysis is typically performed through two general methods, static analysis and dynamic analysis. Through static analysis control flow graphs can be generated to represent the flow of a program between different code-blocks. Which further data can be generated to represent individual instructions as VEX commands. By organizing this data into StellarGraph objects, we can use GraphSAGE models to generate low-dimensional embeddings for familial attribution with greater resistance to unseen samples causing mis/overfitting.

Keywords—machine learning, GCN, malware, attribution, CFG, control flow graphs

# Introduction

The ability to accurately predict the threats that unknown binaries propose to an environment is an ongoing issue within cyber security. Because current intrusion detection systems (IDS) and intrusion prevention systems (IPS) rely on static, signature-based solutions, the most minute changes to code can cause the same signatures to be targeted or misclassify malicious binaries as otherwise benign. Where this research diverges from traditional methods is the type of static features extracted, their represented format, and the algorithms used to assess familial inferences. In doing so we hope to reduce the number of false positives produced by misfit models classifying unknown binaries; in short, reducing the number of networks or environments compromised by emerging malware or advanced persistent threats (APTs).

# Literature REview

## Practical Malware Analysis

Malware analysis is typically performed through two general methods, static analysis and dynamic analysis. Static analysis is done without running the binary and the same methods can be performed against source code as well. Whereas dynamic analysis requires running the binary and then inspecting different elements about the binary itself, it’s memory, or the environment it is running in. Most analysts will use these two methods of analysis depending on what kind of signatures they are trying to generate. Static analysis provides indicators that can be used for attribution before malware interacts with a victim’s environment or network. Indicators derived from dynamic analysis only provide indicators when they have already been deployed and are trying to interact with the host or another system in some way. Essentially, these are pre and post deployment indicators.

## Inductive Representation Learning on Large Graphs

Executables can have thousands of functions or code-blocks represented within a control flow graph. Because of this, it is important to determine whether there are any valid ways to efficiently generate embeddings and inferences from the nodes within those graphs. [2] presents GraphSAGE (Graph Sample and Aggregate) from the StellarGraph library [3], an inductive framework that leverages node attribute information to generate embeddings via forward propagation for previously undigested data.

1. Visual demonstration of GraphSAGE sampling and aggregation.

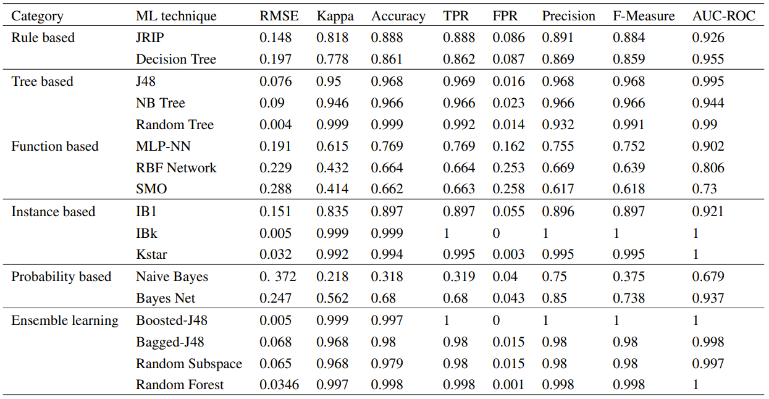
On average the traditional machine learning algorithms like, Random Tree, achieved an accuracy score of ~99%; the lowest score being Naïve Bayes, a probability based technique. This means that it has already been established that current models can accurately detect and categorize unknown binaries into malware families. One limitation of methods like these though could be the inability to account for new categories or variants of malware. Something which is constantly happening due to emerging threats.

## Evaluation of Supervised Machine Learning Techniques for Dynamic Malware Detection

The purpose of this paper is to evaluate the use of generative adversarial networks, however, it is good to have a baseline with which we can compare our models too; especially in terms of training time, testing time, and overall complexity. This paper however does use more traditional features determined at runtime versus control flow graph data. In particular, the 41 features chosen were:

1. [4] features built during data generation phase

|  |  |  |
| --- | --- | --- |
| Category | Feature | Data Type |
| Info | Duration | Integer |
| Network | UDP\_requests | Integer |
|  | IRC\_requests | Integer |
|  | http\_requests | Integer |
|  | smtp\_requests | Integer |
|  | tcp\_requests | Integer |
|  | hosts\_contacted | Integer |
|  | DNS\_requests | Integer |
|  | domains\_contacted | Integer |
|  | ICMP\_requests | Integer |
| Usage | CPU\_usage | Integer |
|  | mem\_usage | Integer |
| Dropped | Dropped | Integer |
| Processes | processes\_generated | Integer |
| API categories | Noti\_API | Integer |
|  | Certi\_API | Integer |
|  | Crypto\_API | Integer |
|  | exception\_API | Integer |
|  | file\_API | Integer |
|  | iexplore\_API | Integer |
|  | misc\_API | Integer |
|  | netapi\_API | Integer |
|  | network\_API | Integer |
|  | ole\_API | Integer |
|  | process\_API | Integer |
|  | registry\_API | Integer |
|  | resource\_API | Integer |
|  | services\_API | Integer |
|  | Syn\_API | Integer |
|  | system\_API | Integer |
|  | ui\_API | Integer |
|  | other\_API | Integer |
| API Summaries | files\_accessed | Integer |
|  | files\_written | Integer |
|  | files\_deleted | Integer |
|  | Mutexes | Integer |
|  | executed\_cmds | Integer |
|  | started\_services | Integer |
|  | files\_read | Integer |
|  | resolved\_APIs | Integer |
|  | created\_services | Integer |
| Malware Family | Family | Categorical |

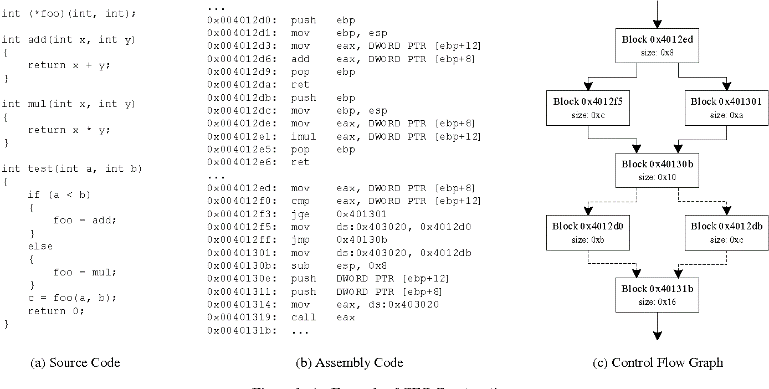
These features were gathered using Cuckoo sandbox which is a controlled environment that allows malicious executables to be safely detonated and then monitored to capture runtime data. The models used and the performance metrics to produced were:

1. [4] Performance metrics for each category of machine learning techniques used

On average the traditional machine learning algorithms like, Random Tree, achieved an accuracy score of ~99%; the lowest score being Naïve Bayes, a probability based technique. This means that it has already been established that current models can accurately detect and categorize unknown binaries into malware families. One limitation of methods like these though could be the inability to account for new categories or variants of malware. Something which is constantly happening due to emerging threats.

## Metamorphic Malware Detection using Control Flow Graph Mining

There are different ways that executables can be represented, but the three most common are: source code, disassembly, and control flow graphs. Each of these representations allows us to derive different information or indicators which will be the primary data source for our models. For instance, with source code we know exactly what the code originally looked like before the executable was compiled. We can see the libraries that were imported, the functions called from those libraries, as well as the structure of any methods or data types.

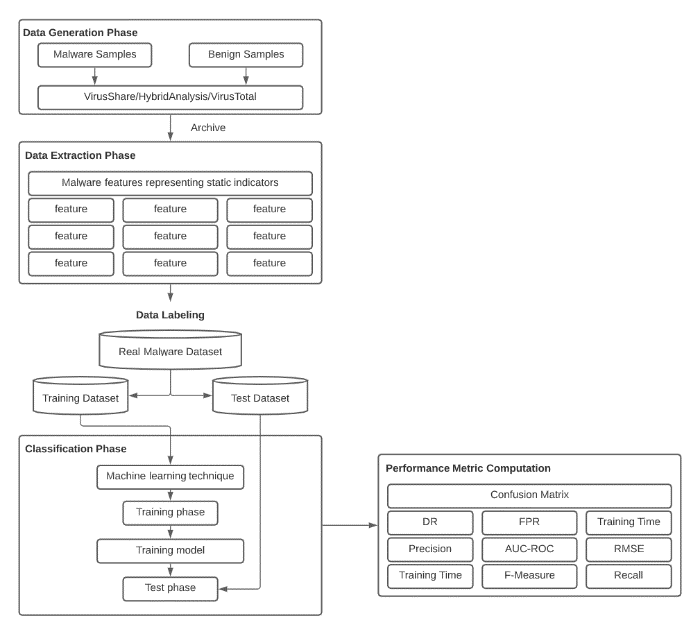


1. Different representations of an executable

We can disassemble an executable and use the raw assembly or opcodes produced from that disassembler to count opcode frequency, instruction count, or to even track references to addresses within memory. Similarly, we can use that same disassembly to create decompilations which are a form of pseudo-source code meant to improve the understanding of a program [6]. Depending on the language an executable was written in, decompilation will produce wildly different results. Like the difference between the decompiler used by dnSpy for .NET code versus the results produced by IDA with Hex-Rays and any C or C++ programs [7][8]. [5] discusses how we can use the third representation type, control flow graphs (CFG), to gather further data as features for our datasets.

Control flow is indicated by instructions using an unconditional jump **jmp**, a conditional jump **jcc**, a function call call, or a function return **ret** operator. Using this intuition, we can build CFGs defined by the entrypoint address of a function to where any of the previously indicated instructions are made. This will produce CFGs consisting of multiple blocks interconnected by arrows; an example of such a graph is shown in Figure 3.

# proposed solution



1. Four main phases of the proposed solution

The phases for this proposal are as defined by Figure 4. Under the Data Generation Phase malicious portable executable (PE32) samples will be gathered from the shown sources: VirusShare, HybridAnalysis, and VirusTotal. The benign samples will be gathered from the same sources as well as from verifiable Microsoft Windows images. The exact distribution for this sampling is 991 malicious and 991 benign samples.

## Data Extraction

The CFG data will be extracted from the samples using angr [9], a multi-architecture binary analysis toolkit, with the ability to perform dynamic symbolic execution, and various static analyses on binaries. For the purpose of this proposed solution, we will use a static CFG (CFGFast) to generate a CFG for each sample. The analytical data for these CFGs are stored using the VEX intermediary language of each node or block within Neo4j [10].

## Data Labeling

Awaiting data generation from Idaho National Laboratory.

## Classification

The GraphSAGE framework from Stellargraph is currently the preferred method for supervised learning and node classification of the CFGs and their node features generated by angr. Using this framework will allow typical splits of 70% training and 30% testing dataset. After several runs of training the models with a mix of benign and malicious PE32 binaries, the model will be used to predict the family and class of the binaries. The output of the classification phase is a confusion matrix and other details not limited to training/testing time.

## Performance Metric Computation

The performance metric computation phase will calculate the necessary performance metrics as indicated in Figure 4. The matrix will be used to calculate False Positive Rate (FPR), Precision, Accuracy, Root Means Square Error (RMSE), F-Measure, and AUC-ROC. Finally, a table of the performance metrics will be created containing the values from the proposed metrics.

# Results and evaluation

Awaiting data generation from Idaho National Laboratory.

# Conclusion

Awaiting data generation from Idaho National Laboratory.

##### References

1. M. Sikorski, A. Honig, and R. Bejtlich*, Practical malware analysis: the hands-on guide to dissecting malicious software*. San Francisco, CA: No Starch Press, 2012.
2. W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive Representation Learning on Large Graphs,” *31st Conference on Neural Information Processing Systems (NIPS 2017)*.
3. CSIRO's Data61, “StellarGraph Machine Learning Library,” *Github Repository*, 2018.
4. H. Zhao, M. Li, T. Wu, and F. Yang, “Evaluation of Supervised Machine Learning Techniques for Dynamic Malware Detection,” *International Journal of Computational Intelligence Systems*, vol. 11, no. 1, p. 1153, 2018.
5. M. Eskandari and S. Hashemi, “Metamorphic Malware Detection using Control Flow Graph Mining,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 11, no. 12, 2011.
6. *The Decompilation Wiki*. [Online]. Available: http://www.program-transformation.org/Transform/DeCompilation. [Accessed: 05-Nov-2020].
7. 0xd4d and Ki, *dnSpy - Github Repository*, 2016. [Online]. Available: https://github.com/dnSpy/dnSpy. [Accessed: 2020].
8. *Hex Rays*. [Online]. Available: https://www.hex-rays.com/products/decompiler/compare/. [Accessed: 05-Nov-2020].
9. “Introduction,” *angr Documentation*, 2018. [Online]. Available: https://docs.angr.io/. [Accessed: 05-Nov-2020].
10. “Intermediate Representation,” *angr Documentation*, 2018. [Online]. Available: https://docs.angr.io/advanced-topics/ir. [Accessed: 05-Nov-2020].