Opcode Sequences as Representation of Executables for Data-mining-based Unknown Malware Detection

Jeff Malavasi
Dept. of Computing Security
Rochester Institute of Technology
Email: jm3378@rit.edu

I. Review

The authors of the paper explore the limitations of signature based malware detection and propose a novel technique based on the prevalence of specific op-code sequences in the binary. During signature-based detection, malware is identified by cross referencing a database of known malicious files which makes these models especially vulnerable to novel attacks. The researchers combat this limitation by designing a malware detection system that looks at the sequence of two and three word sequences of machine code, which is then fed into a machine learning classifier [1].

The proposed solution uses a data-mining-based approach which relies on a dataset that is comprised of both malicious and benign samples. Previous work has shown some success in extracting features based on headers, strings and byte sequences but these methods can be easily evaded. In this paper, the researchers computed the frequency of each op code as it appeared in each dataset in order to determine its relevance to that specific class. While many different lengths of op code sequences were tested, the researchers found that longer sequences drastically increased the training time and resources required to generate the model. Additionally these longer sequences were more vulnerable to evasion, and they achieved a 95% accuracy rate with just two word sequences [1].

The researchers evaluated and validated four different algorithms: decision trees, support vector machines, k-nearest neighbors, and bayesian networks before concluding that decision trees were the most accurate. The datasets were broken up into ten subsets comprised of 90% learning and 10% testing. Additionally, due to the larger feature sets produced by two word sequences, the most relevant 1000 features were selected. Interestingly, the combination of one and two word sequences did not increase the accuracy of the model. The researchers found that bayesian networks also were accurate, but produced a high number of false positives [1].

In the proposed solution, the researchers successfully create a machine learning malware classifier that uses op code sequences as features. They provide a robust evaluation of multiple algorithms in order to show the relevance of using op code to detect novel malware. However, the solution is not practical with packed binaries, which is a common obfuscation technique used by malware designers. Additionally, the researchers were only able to successfully test one and two word sequences, while noting that longer sequences may improve accuracy but also vulnerability to evasion.

As malware becomes increasingly sophisticated, signature based detection models are becoming more and more inaccurate. Therefore, it is important for researchers to create models that are both resistant to evasion and can easily detect novel malicious binaries. The proposed solution advances the current models exponentially, while providing a low false positive rate.

While the researches were able to create a novel detection system, they noted a few limitations to their design. Firstly, the model does not scale well with larger op code sequences. By providing the model with more data it becomes more accurate, but requires a significant amount of feature reduction and training. Additionally, the model is unable to classifier packed binaries. The researchers outline a few techniques to unpack the binaries prior to evaluation, but these methods were not tested and would require dynamic analysis. While dynamic analysis can provide additional features to the analysis, it also requires a specialized environment to safely test and avoid evasion. Finally, the researchers omitted the operands from the code sequences. Although they showed that the generated model could still be accurate without them, further research is needed to determine whether or not feature selection could be improved with the inclusion of operands.

In conclusion the researchers highlight the current problems with signature-based detection and provide a novel machine learning model that analyzes the frequency of op code sequences in binaries. They were able to correctly classify malicious and benign samples, however future research is needed to scale and train the model to work with common obfuscation techniques.

References

[1] I. Santos, F. Brezo, X. Ugarte-Pedrero, and P. Bringas, "Opcode sequences as representation of executables for data-mining-based unknown malware detection," in *Elsevier*, 2011.