Group

Sayli Karnik, Shubhangi Kishore, Binayak Ranjan Das, Atharva Urdhwareshe

Title

BotBucket - Static Analysis and Classification of Botnets and Malwares

Problem

With the rapidly growing amount of software in use around the world, it has become impossible to manually audit it all for vulnerabilities, so it is imperative that we develop efficient systems to automate this process. Malware and Bots have been a pressing threat to the internet and computational systems. Due to the extensive use of the binary obfuscation, it is very difficult to analyze and classify the bots. Various techniques including intrusion detection uses algorithms that helps to detect and classify these bots and eradicate them. These algorithms must have a good sense of understanding of a bot binary to make accurate classification and filtering. The ability of Machine learning to extract meaningful decision-making based on wide ranging features of data has been applied to wide ranging fields. We want to develop an intrusion detection system which analyses according to host and network to successfully train and classify the obfuscated bot binaries using different machine learning models. This system will be designed to effectively deal with the malware and the obfuscated bot binaries.

Context

The paper- "Automatic analysis and classification of obfuscated bot binaries"[1], the authors create a framework to analyze a bot binary's runtime system call trace and uses the longest common subsequences between system call traces for the classification of bot binaries. The develop heuristics to work with the relocation of instructions in the bot binary. In "Scalable, Behavior-Based Malware Clustering"[2] , the authors use a clustering approach to identify and group malware samples that exhibit similar behavior by performing dynamic analysis to obtain the execution traces of malware programs. And generalizing these traces into behavioral profiles. There also exists tools to examine and analyse malware such as CWSandbox [3], Norman Sandbox [4]. We have seen the use of machine learning models for intrusion detection system and would like to use the same capabilities to classify the botnets and other malware.

Approach

We want to track the system calls made when the binary is executed as well as perform a static analysis. Obfuscation does not change the original system call sequence in a bot binary. To identify and group malware samples that exhibit similar behavior we look at their system call sequences. For static analysis, we will focus primarily on open source Botnets like Mirai and Hajime and Windows Portable Executables (PEs). Radare2 is a complete open source framework for reverse-engineering and analyzing binaries. We want to analyse multiple dimensions of malware executables. The greatest limitation to static analysis is obfuscation of code. Most malware code is obfuscated. We would be using Radare2 for Reverse Engineering and analysing bot binaries. We will be analysing the bot binaries and looking at the trends in malware detection to engineer features, like the sequence of system calls, text in the file, or presence of another language. We would use this features to fit out model and analyse the similarity between the botnets. Previous approach for this task used "CaLCS Continuously Approximating LCS ", a differentiable surrogate of LCS, to predict the match between the malware code. We want to use an ensemble approach where we would be surveying various machine learning models that have tried to categorise malware and using them as features or trees in our model. "Malware Images: Visualization and Automatic Classification" converts the malware binary to an 8 bit vector grayscale image and use images classification CNN models to classify the malware variants belonging to the same family. We would be using the image vector as our feature as Images can capture small changes yet retain the global structure.

Evaluation

We plan to present a framework for classification of several obfuscated bot binaries and define a similarity metric based on LCS values. The classification accuracy(true positive rate and true negative rate) is one of the criteria for evaluation.

Scope

Phase 1: Binaries selection and experimenting with Radare2.

Phase 2: Build the classifier.

Phase 3: Inferences and evaluation techniques. We would also be analysing the trends in Malware and how they have changed over time.

Reference

1. Lin, Y.-D & Chiang, Y.-T & Wu, Y.-S & Lai, Y.-C. (2014). Automatic analysis and classification of obfuscated bot binaries.
2. Bayer, Ulrich, et al. "Scalable, behavior-based malware clustering." *NDSS*. Vol. 9. 2009.
3. CWSandbox. http://www.cwsandbox.org/, 2008.
4. Norman Sandbox. http://www.norman.com/ microsites/nsic/, 2008.
5. CALCS: Continuously Approximating Longest Common Subsequence for Sequence Level Optimization [Semih Yavuz, Chung-Cheng Chiu, Patrick Nguyen, YonghuiWu]
6. S Ranjan, “Machine learning based botnet detection using real-time traffic features”