MISSISSIPPI STATE UNIVERSITY

HIGH PERFORMANCE DATA ANALYTICS

DASI

Cyber Phylogeny

CHALLENGE

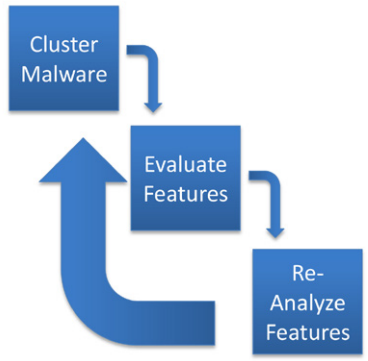
In the current environment of advanced nation-state and highly organized criminal actors creating malicious software, attribution is a problem that must be addressed in order to respond to these threats. We recognize the need to improve methods for identifying and attributing attackers' intent on disrupting network operations. Although some attackers use hands-on techniques, most successful attacks use at least some automation in the form of malware, to facilitate the invasions. We need to better understand this malware in order to best prevent future attacks.

CURRENT PRACTICE

In Year 1, we have developed a procedure for gathering the dynamic data needed to perform machine learning on malware samples. In this procedure, memory images are created at runtime for large numbers of samples. These memory images are then processed to extract the needed data. In Year 1, this data included indicators of DLL injection, which was used with existing machine learning algorithms to classify malicious DLLs in memory images.

Throughout Year 1 we've continued to look into new bio-based technologies that can be applied to our malware characterization and/or detection. We've discovered several new methods that could be applicable to getting "better" results and have started to incorporate some of the "easier" ones such as using HMMR (Hidden Markov Models) for motif generation. We've also discovered some new mathematical and statistical methods that can be used to optimize our process in both speed and precision.

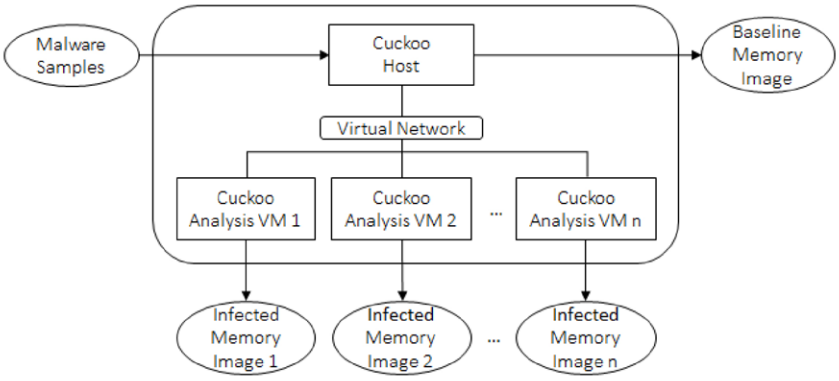
Developing a phylogeny of malware samples, characterizing their lineage, and informing attribution



Iterative process for identifying features for machine-learning clustering

TECHNICAL APPROACH

We are identifying features of malicious software that can be used by machine learning algorithms to help identify the presence of malicious software in the memory images of sandbox environments, and classify that malware based on the feature set. With a collection of these features, relationships between malware samples can be identified and a "family tree" may be built. We are performing the collection of data, identification of features, and application of that information to new samples in a high-performance computing environment.



By generating memory image data for feature extraction, we can identify relationships between malware samples and build a "family tree."

Overall, our approach for Year 2 revolves around the cooperation between teams at MSU and PNNL and integrating the techniques developed at both while still moving forward. We intend to explore and implement options for utilizing the output of the PNNL techniques as features for classification in the MSU process, as well as use the MSU dynamic data capture as a potential way to bypass issues with packed and obfuscated code.

Biosequence analysis will enhance the malware analysis process in operational environments by organizing the space of malware into a rigorous categorical structure. This has two immediate benefits. First, it will significantly reduce the computation required to recognize malware because the family tree allows us to represent many individuals using a single family recognizer. Second, because it is tolerant to variation, it introduces the possibility to catch previously unseen attacks simply by their similarity to known attacks.

IMPACT

With the ability to leverage high-performance computing resources, analysts will be able to take current sources of potential new malware samples and quickly classify those samples by potential authorship or common source. New malware samples may be grouped with existing samples that have a high degree of similarity, giving analysts more information on specific threat actors. This benefits those who wish to attribute cyberattacks by giving them more data that can be examined for operational mistakes by the attackers. This raises the likelihood of finding the source of an attack or widespread malware infection.

Contacts

Wesley McGrew
Task Lead
662.325.8278
mcgrew@cse.msstate.edu

Funding provided by a grant from Pacific Northwest National Laboratory.