CSCI-550 Presentation Summary

Andrew Johnson and Kemal Turksonmez

November 15, 2020

1 Topic

Our presentation was focused on introducing Anomaly Based Intrusion Detection Systems (A-NIDS) and the techniques that can be used to build these systems. We primarily focus on the paper "Anomaly-based network intrusion detection: Techniques, systems and challenges."[1]

2 Anomaly Based Intrusion Detection Systems

Intrusion Detection Systems (IDS) are used to detect potentially malicious activity on a network or system. In the paper that we cover, the authors introduce two different types of Intrusion Detection Systems. The first is a signature based system which uses previously known attack signatures to identify incoming malicious activity. The other system is an A-NIDS, which utilizes knowledge of a machine's behaviour to identify suspicious activity that deviates from the defined norm. The biggest difference between the two systems are that signature based systems require previous knowledge of attack signatures in order to identify malicious activity.

3 Anomaly Detection Techniques

The techniques used for anomaly detection in these systems can be largely categorized into three different types of techniques: statistical based, knowledge based, and machine learning based.

Statistical based techniques rely on analyzing data and determining the stochastic behavior of the network. This allows the comparison of new behavior to the previously created model to determine how anomalous current behavior is. Several types of models fall under this category. First are univariate models which look at single variables as gaussian random variables, multivariate models which look at the interaction between variables, and time series, which consider timing of events. Statistical based methods have the advantage of not needing to really know what normal behavior really looks like, as the model will determine what is 'normal' through its training phase. However, the drawbacks of

this model are that false positives can be very common if model parameters are not tuned properly and over-tuning the parameters to lower the false positive rate drastically will result in false negatives, and these models are vulnerable to being trained by adversaries. If the network traffic that the model is trained on is already compromised, then the behavior which should be classified as anomolous will become part of the data that is considered normal, and future malicious behavior will be overlooked as normal behavior. This is possible even more so when new data from network traffic is being added to the training data set.

Knowledge based methods rely on high-quality data and knowledge of what normal behavior is. Expert systems allow the classification of anomalous behavior based on a set of rules. These rules can be determined manually through an expert's knowledge or derived from high quality data. Formal methods allow the definition of normal behavior through well defined techniques such as finite state machines or UML models. These techniques can be particularly time consuming as it requires defining all standard behavior that would be expected in the network traffic.

The third type of technique is machine learning based. These models share a lot of their advantages and disadvantages with statistical methods. Several examples of these types of techniques are bayesian networks, markov models, neural networks, fuzzy logic, genetic algorithms, and clustering. Without going too in depth on all of these techniques, these are largely built upon the idea of statistical models. They allow for a set of data which is considered normal for training a model, and then using that model to identify behavior that did not appear in the normal behavior, or sequences of events that may not have appeared in the training data. Just like statistcal models, they are prone to being trained by an adversary and require parameter tuning for optimal detection rates.

As an example of what a machine learning method would look like, we will look at clustering specifically. Clustering is a technique used to partition data into groups with similar characteristics. We have gone over a few methods for clustering such as K means and DBScan, either of which could be used for this purpose. Clustering would be used to classify training data which consists of standard network traffic data that is considered normal. You may be wondering how clustering would be utilized at this point, as the function of clustering is primarily what has already been done: partitioning data. One way to utilize this partition is to select a point from each group which represents the cluster well. Then when we seek to determine if new traffic is anomalous, we would check if it belongs in one of these clusters by distance from the representative points. If the new data does not fit well into any of the clusters, we would classify it as anomalous.

References

[1] Garcia-Teodoro, Pedro, et al. "Anomaly-based network intrusion detection: Techniques, systems and challenges." *computers & security 28.1-2 (2009): 18-28.*