MALWARE DETECTION USING HONEYPOT AND MACHINE LEARNING

https://doi.org/10.5281/zenodo.6591373

Salimova Husniya Rustamovna
Master's degree, specialty "Information Security", Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
Ganiyev Asadullo Mahmud o’g’li
Bachelor degree, Faculty of Software engineering, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan

In our thesis, we explained honeypot systems in detail, and implemented low interaction, middle interaction and high interaction honeypots at laboratory. Our goal was to understand their strategy and how they are working in order to lure intruders towards the system. We discovered their security flaws in order to help researchers and organizations. Several companies are using honeypot systems to protect the whole organization’s network security, and researchers are making academic experiments on them at schools. As we all know network security is very significant for all computer systems because any unprotected machine in a network can be compromised in any minute. One may lose all the secret and important data of a company, which can be a great loss, and it is also very dangerous that someone else knows your important personal information. Thus, we tried to find answers for honeypots’ security using all interaction honeypots possible. Our main goal for our thesis was to see if honeypots are easy to hack and check if they are really isolated from other networks like a organization’s network. When a honeypot is compromised, is it possible to reach other systems and compromise them too ? After the system is compromised, is it possible to track the hacker by using necessary forensic science tools ? How efficient are they ? As we stated in results and analysis part,we easily hacked all the honeypots that we used for our thesis. Especially, low interaction honeypot Honeyd can be hacked easily without too much effort. As we stated before, any amateur hacker can seize the system and also can see that it is a trap system. Therefore, Honeyd is not a good honeypot as its features are not efficient to fool the hacker. As Honeyd is a deamon, it is just simulating a operating system’s services. So, it is not possible to a hacker to seize other systems using Honeyd. For the intruder, it will not take time to see that the system is not real, so he will not continue compromising it. He will leave the system. For forensic part, Honeyd’s log was sufficient to see the actions of the hacker. Next part was to try Nepenthes as medium interaction honeypots. The result was quite similar. Thus,we came up with this conclusion: Low interaction honeypots and medium interaction honeypots are just simulating the services of a real system, because of that it is not possible to capture significant data from intruders. They are slightly different from each other but the main idea is the same. As they are not real operating systems , it is not risky to build them. There is no need to mention about further attacks. So, we moved on to the last level. After working low interaction and medium interaction honeypots, we decided to deploy high interaction honeypots. We studied on Honeywall. Even though it is time consuming and difficult, we managed to create a structure and worked on it. Our result were more interesting than before.High interaction honeypots are not virtualizing the system. They are real systems.So, it is very risky but the captured information is important. After deploying the implementation correctly, we successfully hacked the honeynet, but not Honeywall itself. It was the result we were looking for. The development of Internet and social media contributes to multiplying the data produced on the Internet and the connected nodes, but the default installation and the configuration of variety of software systems represent some security holes and shortcomings, while the majority of Internet users have not really set up safety awareness, leading to huge security risks. With the development of network attack techniques, every host on the Internet has become the target of attacks. Therefore, the network information security cannot be ignored as a problem. To deal with 0-day and future attacks, the honeypot technique can be used not only passively as an information system, but also to reinforce the traditional defense systems against future attacks. In this paper, we present an introduction of machine learning and honeypot systems, and based on these technologies, we design a smart agent for cyber-attack prevention and prediction.

The purpose of this paper is to show, firstly, the strength of using machine learning and honeypots, as solutions for the cyber security purpose, through some related works and by introducing these technologies. The second purpose of this work is to discuss a cyber security solution based on honeypot and machine learning techniques. Our main objective is to design an intelligent agent for predicting new attack profiles by analyzing, automatically, the gathered data via the honeypot, using a combination of machine learning algorithms. The objective of the algorithms combination is to represent the data with a lot of accuracy and build an efficient predictive agent for cyber security, especially for the future and 0-day attacks prediction.

Anomalies characterization receives a lot of attention; all seek to protect themselves against fraudulent use of their data or against malicious intrusions into computer systems. A lot of security solutions were proposed in the last decade, but results still present some limitations, and the most recent works are based on machine learning algorithms to model anomalies activities using data collected by information technologies such as honeypots.

The authors in propose an intelligent honeypot which improves IoTs devices' security, based on machine learning. In order to store each device response, an IoT-scanner was proposed to probe accessible IoT devices on the Internet and scan the Internet for each malicious interaction, and a model called IoTLearner was trained to be used by the intelligent honeypot that can optimize a model to reply attackers.

The authors in propose an autonomous method for attacks characterization, based on unsupervised anomalies learning, using the collected information by honeypots. This approach is based on clustering techniques such as density-based clustering, subspace clustering, and evidence accumulation for classifying flow ensembles in traffic classes. The advantage of this method is that it does not require a training phase. The authors in propose an automatic classification of social spam-based machine learning (e.g., SVM), for network communities such as Facebook and MySpace using a social honeypot to gather information about malicious profiles.

The authors in propose a linkage defense system based honeypot to overcome the limitations of the traditional tools. The linkage technique will ensure management and communication between the honeypot and components of the defense system, constructing a linkage management module based on SNMP protocol for network management. In order to overcome the problem of new attacks, the system is centroid honeypot for treating suspicious flows arrived from the traditional defense system, and the decision to block or not will depend on the state of the honeypot. If the honeypot is damaged, then the correspondent intruder will be blocked by the firewall.

A Honeypot is a diversion of intruders' attention, in order for intruders to think that it has managed to break down and retrieve data from a network, when in fact the data is not important and the location is isolated. A way to trap or deny unauthorized use of effort in an information system. One type of honeypot is honeyd. Honeyd is a low interaction honeypot that has a smaller risk compared to high interaction types because the interaction with the honeypot does not directly involve the real system. The purpose of the implementation of honeypot and firewall, firewall is used on Mikrotik. Can be used as an administrative tool to view reports of Honeyd generated activity and administrators can also view reports that are stored in the logs in order to assist in determining network security policies.

Deployment of honeypots depends on whether the decoy system is intended to monitor external or internal attacks to the organization's network; hence, it can be installed in front of the firewall, in a demilitarized zone (DMZ), or behind the firewall .

A computer is not smart; it performs tasks described in a program form, as orders of what to do and how to do it, and this is called traditional programming. While writing a traditional program, the decision is made directly into the program. Machine learning is a subarea of artificial intelligence that aims to give computers the opportunity to learn; its techniques allow understanding the structure of the data and integrating them into models that can be understood and used to solve complex problems in real-life situations, and its techniques represent an efficient tool to address the significant challenges posted by the big data.

A machine learning model is designed in two phases: the first is to estimate the model from the available data, by executing practical tasks such as animal recognition in pictures, speech translation, or participating in autonomous vehicles driving, this is called the training phase, and it is generally performed before the practical use of the model. The second is the production phase, meaning the phase of passing new data to obtain the result corresponding to the desired task. According to the information available during the learning phase, learning is qualified in different ways; if the data is labelled, then the learning would be supervised, in a more general case, and without labels, the learning is unsupervised.

Companies have invested a great deal on time and money in manual networks reconfiguration, in order to protect information systems from infiltration. It is well known that the locks break and the keys can be copied; therefore, it is an illusion to think that a lock and a key represent perfect security. So, the real challenge in terms of cyber security is to accept the probability of an imminent attack and to understand what is really going on within complex information systems. Traditional security tools such as IDS, Firewalls, and IPS can protect systems against simple attacks that use the same tools and tactics repeatedly. They are implemented independently; hence, there is no contact between them to block intrusion detected in an IDS by the firewall, for example, they represent a passive solution when it is about 0-day attacks.

TheThe forensic science branch that we are interested in our thesis is computer forensics which is the same definition of forensic science but this time electronic devices are involved with our researches. The necessary data is obtained from the devices, and forensic investigators make deeper examination on them. There are several roles and responsibilities for forensic investigation. Forensic investigation is done with first responders, investigators, technicians, evidence custodians, forensic examiners and forensic analysts. (Kipper G., (2007)). The different honeypots we studied offered us several log files that a forensic party can analyze. The most common file to study when we talk about network security is the .pcap file that most honeypots are generating. This file contains all the packets exchanged between the attacker and its target. It can be opened with Wireshark and allow the forensic to see what communication happened. This file can be huge in size but contains very important information. The difficulty here is to sort the relevant information. In the case of a honeypot, we assume that all traffic is suspicious thus any IP address not within our network must be analyzed. This make the sorting easier than on a production network where the attack is harder to detect. Another part of the forensic work is called reverse engineering. When a hacker successfully compromises a system, he will most likely upload one or more malware. Reverse engineering take a closer look at these malware by decompiling it and trying to understand what are their purposes and how they work. Again this technique is very time consuming but can allow the forensics team to identify new threats. Honeypot system In the computer network is very important for network security, especially related to applications involving various interests, there will be many things that can disrupt the stability of the computer network connection, whether related to hardware (physical security, power resources) and related to software (System, configuration, access system, etc.). Disruption of the system can occur due to accidental factors performed by the manager (human error), but not least also caused by a third party. Disturbances can include destruction, infiltration, theft of access rights, misuse of data or systems, to criminal acts through computer network applications. Security of the system should be done before the system is enabled. The use of the system should be done before the actual system is enabled.

With a basic scan it is possible to find which ports are open but as soon as the attacker tries to actually connect on a port, he will realize the service is fake. For example the script used for a Web server, by connecting it using telnet, thew server should send back replies but nothing is happening. Another problem is one cannot understand if there is an incoming attack to the system or not. Because there is no such alarm system that can make you understand that there is an attack. Information gathering is not very smart either. As a result the hacker can understand quickly that there is something wrong with the target and will abort his attack. Even unprofessional intruders can compromise the honeypot without spending too much time on it. Because it is very popular and easy to use well known techniques such as Nmap. There is no additional approach needed for it. Our second step was to configure medium level interaction honeypot Nepenthes. We explained how it works and how we studied on it in implementation part. However, we found some problems with Nepenthes too. First of all, Nepenthes is for capturing malware over internet. It is mostly used for this aim. Thus, it must be implemented very rapidly since threats for users over internet are increasing dramatically day by day. Nepenthes could not keep up with new threats. As new threats are arriving and Nepenthes is not up to date, it will not be able to capture malware. Another problem comes from the shellcode. Shellcode manager should consider about shellcode and understand it.

In this paper, we have presented an introduction of machine learning and honeypot as solutions for cyber security. We also presented an efficient algorithm which returns two important information, one for profile creation and the other for classifying this profile. In fact, the specific solution based on honeypot and the combination of machine learning algorithms forms a solid modeling and predictive system for suspicious profile recognition and classification. Hence, it represents an integrated efficient system for cyber security to deal with future and 0-day attacks. Our next work will be devoted to implement the smart agent in a real environment in order to evaluate and test its performances.

REFERENCES:

1. GDataGData, *MalwareNumbers*, 017, http://www.gdatasoftware.com.

2. P. Owezarski, "Unsupervised classification and characterization of honeypot attacks," in *Proceedings of 10th International Conference on Network and Service Management (CNSM) and Workshop*, pp. 10–18, Rio de Janeiro, Brazil, November 2014.View at: Publisher Site | Google Scholar

3. S. Dowling, M. Schukat, and E. Barrett, "Improving adaptive honeypot functionality with efficient reinforcement learning parameters for automated malware," *Journal of Cyber Security Technology*, vol. 2, no. 2, pp. 75–91, 2018.View at: Google Scholar

4. I. M. M. Matin and B. Rahardjo, "Malware detection using honeypot and machine learning," in *Proceedings of 2019 7th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–4, Bandung Institute of Technology, Bandung, Indonesia, November 2019.View at: Google Scholar

5. L. Spitzner, *Honeypots: Tracking Hackers*, Addison-Wesley, Clemson, SC, USA, 2003.

6. T. Luo, Z. Xu, X. Jin, Y. Jia, and X. Ouyang, "Iotcandyjar: towards an intelligent-interaction honeypot for iot devices," in *Proceedings of the Black Hat*, Las Vegas, NV, USA, 2017.View at: Google Scholar