Threat Prediction Using Honeypot and Machine Learning

Vishal Mehta[1], Pushpendra Bahadur[2],
[1]Mphasis-An Hp Company, Magarpatta Cyber City Pune,India
[2]Bhushan Steel Limited, Sahibabad, Ghaziabad (U.P.)-201010
E-mail: sirius.mehta@gmail.com

Manik Kapoor[3] , Dr. Preeti Singh[3] and Dr. Subhadra Rajpoot[3]
[3]Amity University, Greater Noida (U.P.)-201308

***Abstract*—Data is an abstraction which encapsulates information .In today's era businesses are data driven which gives insight to predict the destiny of the business by making predictions but another side of the coin is data also helps in placing the present health of the business under our radar and looking back in our past and answer some important questions: what exactly went wrong in the past?. In this paper we try to look into the architecture of frameworks which can predict threat using Honeypot as the source of data and various machine learning algorithms to make precise prediction using OSSEC as Host Intrusion Detection System [HIDS], SNORT for Network Intrusion Detection System [NIDS] and Honeyd an open source Honeypot.**

IndexTerms—Host Intrusion Detection System (HIDS), Network Intrusion Detection System (NIDS), Low Interaction Honeypots (LIH), High Interaction Honeypots (HIH)

I. INTRODUCTION

For any organization data is an asset in this paper we try to review the present techniques and technologies architecture used for threat prediction. In today's era anything and everything we do on any digital entity (say) mobile, laptop etc. gets logged, these logs are really important source of information to us. They can answer our basic questions we are interested in as: Who was responsible to break into the system? When they break into the system? What was damaged? Till what extent they were successful to penetrate into the system? Is it a compromise?
Was the compromise successful? And so on.

In order to protect the integrity of the data we need to self-answer some questions: Who are our enemies? What are their skill levels? What are the tools and technologies they are using in order to exploit our systems and gain unauthorized access to the data? Is the exploit is made just for fun or intruder has serious goal to spread the threat in the network being silent and fast. Hence in order to protect data we need some alerting systems which can alarm the administrators in time when somebody tries to break into the system and here SNORT [NIDS] and OSSEC [HIDS] come into the picture.

OSSEC is a host based intrusion detection system which helps us in alerting and maintain the integrity of data. OSSEC can be installed on single machine or in a client-server environment in the production. SNORT is network based intrusion detection system which tries to sniff the network packets and analyze the payload in order to identify malicious packet in the stream of packets by doing deep packet inspection.

With the timeline the intruders are becoming more powerful by creating new exploiting tools and intrusion techniques to quickly come up with the new solutions for the exploits we use honeypots and honeynets. In order to learn about the intruders: Who are the intruders? When the break into the system? Which are the new tools they use in order to exploit system vulnerability? Honeypots are information resources whose value is to get compromise. There are different types of honeypots:
Low Interaction Honeypots and High Interaction Honeypots.
Interaction is how much amount of activity we allow to the intruder, if the amount of activity allowed is low we say it's a low interaction honeypot and if the amount of activity which can be performed is high we say it's a high interaction honeypot.

Low Interaction Honeypots are safe, can be used inside production, and if one gets compromised the administrator can immediately remove it from the network for further analysis of the malware and the production system still remains live. But the only cons of using Low Interaction Honeypots is they allow very less interaction with them hence the data collected from these honeypots are very less because they only collect that data when the transactions are specifically done with the honeypot and since any interaction done with the honeypot is considered as malicious one can use this collected information from the honeypots and use in the production to protect their data from new threats and vulnerabilities.Low Interaction Honeypots are capable to emulate various operating systems and services

978-1-4799-8433-6/15/$31.00©2015 IEEE

High Interaction Honeypots are more powerful as compared. They are also termed as research honeypots because their sole purpose is to do research about the hackers, learning their new tools and techniques they try to break into the system and even can record every keystroke. The risk attached with High Interaction Honeypots is also high because the intruders can gain access to real operating systems and services. The data collected is large.

II. Architecture Of Threat Prediction Framework

In any company environment most of the malicious network traffic gets hidden by the normal or the baseline network traffic

In order to make company data secure a research system is deployed to learn the tools and techniques used by the hackers and after analyzing the pattern of attack being used by the hacker one can create signatures for those patterns and deploy those patterns in the production and also the same patterns found in the data can be used by various prediction algorithms in order to predict the attacks. The company network is protected by the firewall to filter network packets which are malicious in nature but all the network traffic is not filtered by the firewalls.

Honeypots are the information resource systems which are meant to compromise but these honeypot systems do not get the malicious network traffic by its own we need to route that malicious network traffic in the honeypot systems. There are various techniques in order to route the malicious traffic in the honeypot systems, the whole network address space is divided into two sub address spaces, one address space is used by the production and the other sub address space is by the honeypots so any transactions taking place with honeypots is considered malicious.

The honeypot system can be a dedicated computer with real operating system, network stack, open ports with running services, these kind of honeypots are High Interaction Honeypots with which hackers can do a lot on the parallel there are low interaction honeypots with virtual operating system and emulated services. These kind of honeypot systems are Low Interaction and allow very less activity. Low Interaction honeypots are widely used because of their low risk and generally an open source honeypot is used called Honeyd which can virtualize multiple operating systems and services.

In order to understand the malicious we use SNORT which has capability to sniff packets from the network and help in analyzing these network packets. The network packets captured by honeypot: Honeyd which come for analysis are TCP/IP packets. The applications hand over their data to network stack and the data is appended with the IP-HEADER and
TCP-HEADER as we can see in the Fig1.

IP-HEADER	TCP-HEADER	APPLICATIO N DATA

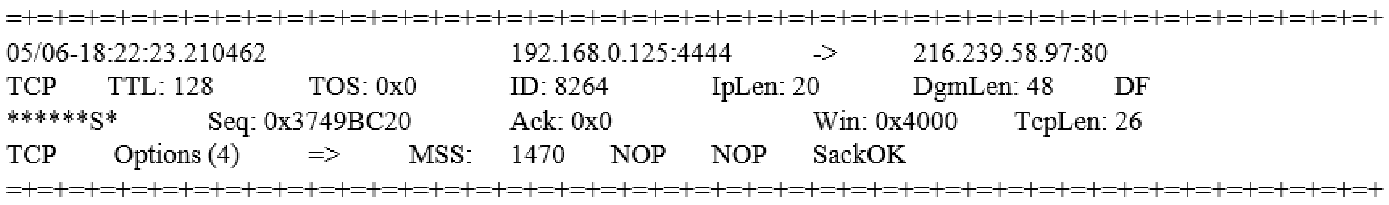
Fig1.

Snort sniffs the traffic from the Ethernet interface and displays the headers of the packets on the console. As we can see in the Fig2. The basic structure of the sniffed network packet, the TCP header is followed by IP header. The packet structure consist of:

- 05/06-18:22:23.210462 : month/day-time the packet was captured
- 192.168.0.125:4444 -> 216.239.58.97:80 : sourceip : port -> destinationip : port
- TCP : Tells which is the transport protocol which handed over the packet to IP
- TTL: Time To Live is the integer which gets assigned by the operating system to the packet either windows or UNIX and decrements after reaching every hop.
- TOS : Type Of Service
- ID : IP Identification number
- IpLen : IP header length
- DgmLen: Total datagram Length
- DF : Do Not Forget
- ******S* : Tells which are the flags are turned ON in the TCP header
- Seq : The initial sequence number set by the client
- Ack: Acknowledgement number
- Win: Window size
- TcpLen : Length of the TCP Header
- Options (4) =>MSS:1470 NOP NOP SackOK: Values in the Optional field in the TCP header

By seeing the structure of our packet we can understand what exactly the hacker was trying to do in our honeypot system. As we can see in the packet structure we have source and destination IP address and port numbers to analyze what are the ip addresses and port numbers by the help of which we can further analyze what has happened between the attacker and the honeypot? What was the pattern of the activity? What are the port numbers which the attacker tried? To which service he wanted to exploit? And once we get all this information one can reconstruct the whole session between the attacker and the honeypot to understand, what are the attacks the attacker tried, which are the rootkits he tried to download and install on the honeypot? What are the commands he tried? Is the attacker trying to connect on the IRC channel? Is the attacker trying to stage the honeypot system to attack other machines on the network?Did the attacker communicated with others on the IRC channel?

The answers to these questions will help us to reconstruct the rootkit which the attacker tried to download on the honeypot and installed it in order to compromise the system. Reconstructing the rootkit will help us understand what the tools are and malwares attacker tried to use were. Is there any new tool, the information of which help in developing the new tool of defense.



```
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
05/06-18:22:23.210462                    192.168.0.125:4444      ->      216.239.58.97:80
TCP     TTL: 128        TOS: 0x0         ID: 8264         IpLen: 20      DgmLen: 48      DF
******S*        Seq: 0x3749BC20          Ack: 0x0                 Win: 0x4000     TcpLen: 26
TCP     Options (4)     =>      MSS:     1470     NOP     NOP     SackOK
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
```

Fig2.

As we can see in Fig5. All these packets sniffed from the Ethernet interface is collected onto the centralized database in the tcpdump file format in the form of log. These logs are files which are the collection of events happening in the network. The logs are being consistently monitored using OSSEC Host Intrusion Detection System so that no hacker trying to get into the machines plays with the integrity of these log files by deleting their session information from the logs so that no administrator can trace the attack.As OSSEC is the HIDS it can be deployed as the client server or the agent-server-architecture. All the events from these agents is collected and forwarded to the centralized storage via remote syslog protocol on port 1514. OSSEC has four processes running on the server side:

- Remoted
- Analysysd
- Maild
- Execd

Remoted is responsible to collect all the incoming events on the port 1514 and forward these events to Analysisd for the further analysis. The Analsysd takes these events logs as an input, the log is sent for Predecoding phase where the static information such as date, time, program name etc. regarding the as we can see in the Fig3. Once the static information gets decoded the log is further send to decode the dynamic information. Once all the information is parsed from the log the decoded values are fed to the rule tree where these values are compared with the predefined or custom signatures and accordingly generate the alert. We can configure OSSEC to email if the alert of high priority gets generated or take an action by executing a script. The email gets generated by OSSEC Maild process and the script gets executed by Execd process. If the tries to break into the network and further is able to get into the system, tries to Brute Force the system login then immediately OSSEC generates a high priority alert.

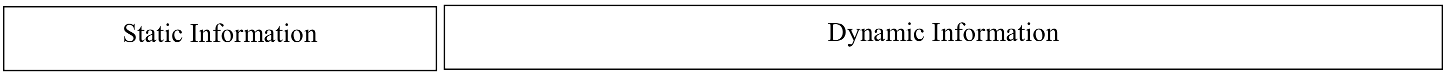
There are various features supported by OSSEC as File Integrity Check. OSSEC is able to monitor the integrity of the files by comparing their checksums from the baseline files. If the attacker tries to modify the size of the file then OSSEC can alarm on that because it keeps a check on all the attributes of the file.

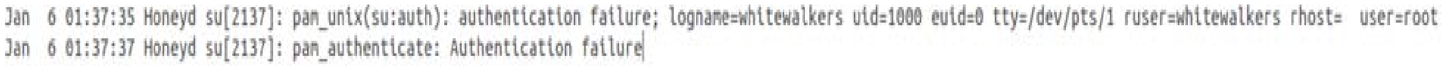
All this information is send to the analyst workstation in order to monitor. Threat can predicted using the information collected from the honeypot systems applying various machine learning algorithms such as neural networks, Bayesian Classification by modelling the data as in Fig4.

$$P(Class\,|Compromise) = P(Compromise\,|\,Class)P(Class)/\,P(Compromise)$$

Fig4.

Threat can have multiple classes as attack has occurred due to internal privilege escalation or it's the outside attacker. Hence we are trying to predict class of the threat given the machine is already compromised. The $P(Cla$ in the above figure is very important part of the Bayesian theorem as it includes the patterns of attacks occurred in the past. There are many machine learning algorithms which try to predict the threat before it happens but challenge how to face the new upcoming attacks with higher skilled attackers. How can we automate the tools developed from the learning of honeypots to provide better defense systems?





```
Jan  6 01:37:35 Honeyd su[2137]: pam_unix(su:auth): authentication failure; logname=whitewalkers uid=1000 euid=0 tty=/dev/pts/1 ruser=whitewalkers rhost=  user=root
Jan  6 01:37:37 Honeyd su[2137]: pam_authenticate: Authentication failure
```

Fig3.

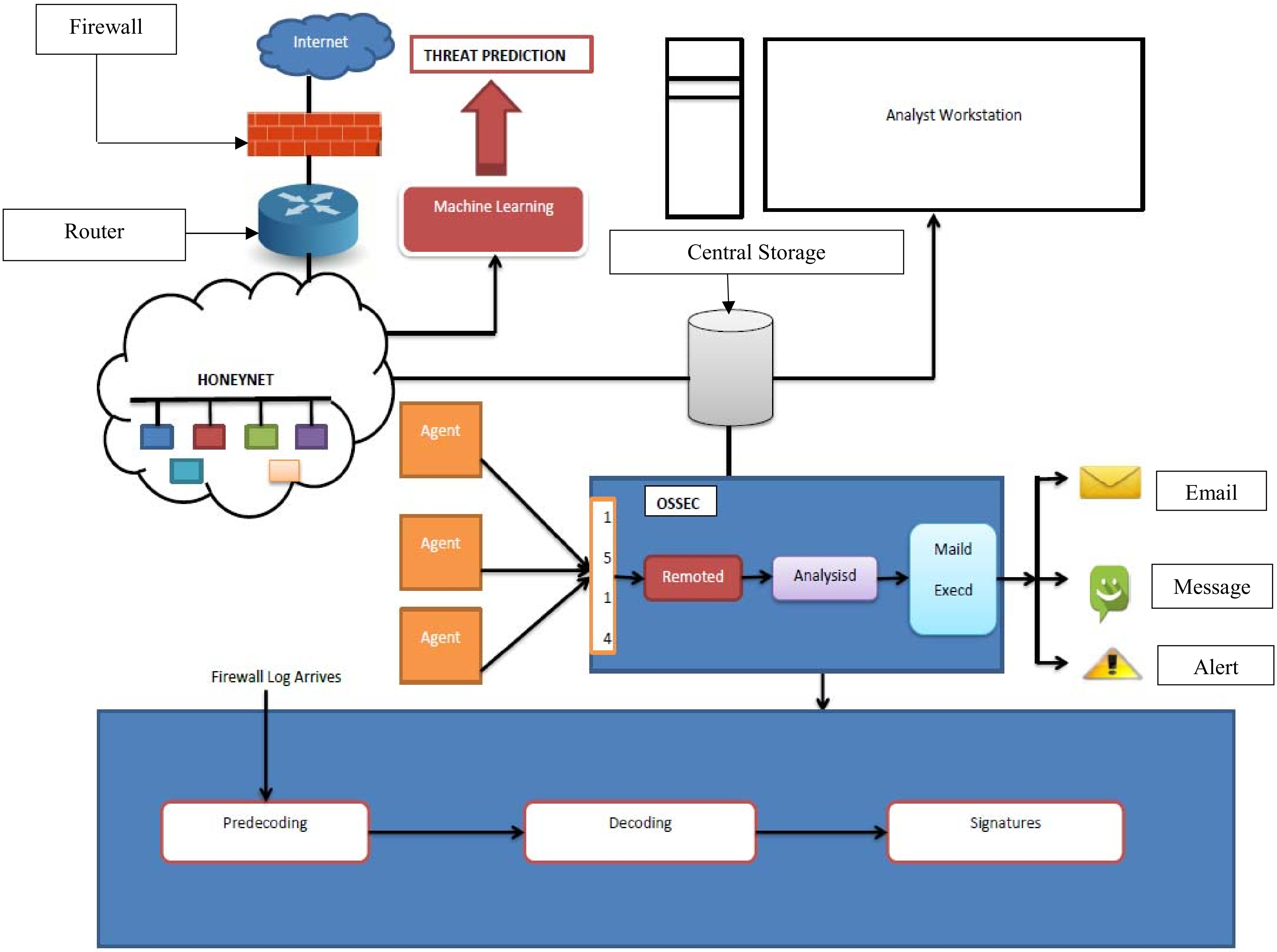


Fig5.

ACKNOWLEDGMENT

The authors would like to thanks Mr. Shailesh Kulkarni and Ms. Ritika Wason for their valuable suggestions.The authors would also like to thanks Mr. Venkat Suresh for the valuable comments and discussions which helped us in improving the quality of the paper.

REFERENCES

[1] An Intelligent Intrusion Prevention System for Cloud Computing (SIPSCC) Alqahtani, S.M. ; Al Balushi, M. ; John, R. Computational Science and Computational Intelligence (CSCI), 2014 International Conference on Volume:2 *.(references)*

[2] Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on Date 2-4 April 2010.

[3] A Stochastic Game Theoretic Approach to Attack Prediction and Optimal Active Defense Strategy Decision Wei Jiang ; Zhi-Hong Tian ; Hong-Li Zhang ; Xin-fang Song Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference.

[4] Enabling attack behavior prediction in ubiquitous environments Anagnostopoulos, T. ;Anagnostopoulos, C. ; Hadjiefthymiades, S. Pervasive Services, 2005. ICPS '05. Proceedings. International Conference.

[5] RAPn: Network Attack Prediction Using Ranking Access Petri Net Traore, M.D. ; Hai Jin ; DeqingZou ; WeizhongQiang ; Guofu Xiang Chinagrid Conference (ChinaGrid), 2011 Sixth Annual.

[6] Attack Scenario Prediction Methodology Fayyad, S. ; Meinel, C. Information Technology: New Generations (ITNG), 2013 Tenth International Conference

[7] A Data Mining Approach to Generating Network Attack Graph for Intrusion Prediction Zhi-tang Li ;Jie Lei ; Li Wang ; Dong Li Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference

[8] A Prediction Model of DoS Attack's Distribution Discrete Probability WentaoZhao ;Jianping Yin ; Jun Long Web-Age Information Management, 2008. WAIM '08. The Ninth International Conference

[9] Cyber Attacks Prediction Model Based on Bayesian Network Jinyu Wu ; Lihua Yin ; YunchuanGuo Parallel and Distributed Systems (ICPADS), 2012 IEEE 18th International Conference

[10] Detection of Syn Flooding Attacks using Linear Prediction Analysis Divakaran, D.M. ; Murthy, H.A. ; Gonsalves, T.A. Networks, 2006. ICON '06. 14th IEEE International Conference on Volume:1

[11] ARM-CPD: Detecting SYN flooding attack by traffic prediction Sun Qibo ; Wang Shangguang ; Yan Danfeng ; Yang FangchunBroadband Network & Multimedia Technology, 2009. IC-BNMT '09. 2nd IEEE International Conference

[12] Prediction of DoS attack sequences Reshamwala, A. ; Mahajan, S. Communication, Information & Computing Technology (ICCICT), 2012 International Conference

[13] Hybrid Framework for Behavioral Prediction of Network Attack Using Honeypot and Dynamic Rule Creation with Different Context for Dynamic Blacklisting Renuka Prasad, B. ; Abraham, A. Communication Software and Networks, 2010. ICCSN '10. Second International Conference

[14] Malicious Modification Attacks by Insiders in Relational Databases: Prediction and Prevention Yaseen, Q. ; Panda, B. Social Computing (SocialCom), 2010 IEEE Second International Conference

[15] An outlook on network honeypot Hongxia Li ; Junming Chen ; Xin Jin Computer Science and Service System (CSSS), 2011 International Conference

[16] Prediction of past unsolved terrorist attacks Ozgul, F. ;Erdem, Z. ; Bowerman, C. Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference.

[17] Model and evaluation of a new Honeynet Liu Yongli ; Zhu Jie ; Wu Shufang ; Wang Zixian Robotics and Applications (ISRA), 2012 IEEE Symposium.