Intrusion Detection in Meta Search Engine Through Heuristics: Honeypot

Priyanka[1], Prof. Ela Kumar[2]
priyankasingh8568@gmail.com[1], ela_kumar@rediffmail.com[2]
Department of Computer Science,Indira Gandhi Delhi Technical University for Women, New Delhi

ABSTRACT

A honeypot is a supplemented active defense system for Network Security. It traps attacks made by third party, record the information about the activities and tools of the hacking process, and prevent attacks withdrawer the comprised system. To prevent, detect and react to intrusions without disturbing the existing system is the serve problem for the Network Security and Web Application. Hence, in this project an Honeypot is made for search engine to make its firewall stronger so that it can prevent and protect the internal network data. IDS work well on detecting and alerting attacks of known signature. Most IDS can't detect unknown intrusion attacks. The problem of detecting unknown attacks solved by anomaly detection. In this paper, Honeypot is python based technique and implemented on local host to integrate with real time system so that it can have advantage of both techniques. Implementing the Honeypot is made for virtual Environment and analysis of collected thread intelligence. Network Malware or an intruder on a local network will typically strive to find open ports. Here goal is to open selected ports that are directed to access Honeypot not to Internal Network. The honeypot should direct traffic to any open network or port and never to internal system.

Keywords: *Web Crawling, Indexing, Meta Search Engine, Honeypot, Virtual Environment, Network Security, IDS*

1. INTRODUCTION

Intrusion Detection System is Security Management tool of computer network administrator who monitors system networks to detect inappropriate accesses. An Intrusion is defined as any set of actions that attempt to compromise the **integrity**, **confidentiality** or **availability** of resources.

IDS, today suffer from several shortcomings in the presence of complex and unknown attacks. The problem of unknown attacks with intrusion detection is solved using anomaly detection. An ID enables any organization to notice and discover attacks, to protect their production server from them and give useful techniques. But these tools sometime lack the ability to detect new thread like zero day vulnerabilities based on zero-day attacks. It also not able to collect more information about the intruder's malicious activities, skills and payloads. For Instance: Antivirus or signature based IDS are not capable of detecting these zero day unknown attacks. IDS work well on signatures which are known to the system , but these zero day unknown are not detected nor any alert made by them because they do not have the signature of those new attacks in database.

The best Honeypot definition is given by Lance Spitzner - *"A Honeypot is an information system resource whose values lies in unauthorized or illicit use of that resource"*. Main advantage of using Honeypot is that it allows anyone to analyze how the attackers behave in vulnerable system and what methods they use for exploiting systems vulnerabilities and this provides security researchers valuable information about the skills of the attackers. A malicious network attacker or an intruder will find and try to strive into the number of ports open on a local network to ruin the internal network information.

Web Search engines have become the necessary tools for web users to search useful information. This represented architecture uses an approach of multi-agent to process the queries of users with greater personalization functionality and higher quality results than any other Meta search engine. So to save internal network from attacker ,firewall cannot make the system so well secure in today's technology ,

there are various techniques used by organizations to secure system from intruders or attackers.

1.1 TYPES OF HONEYPOTS

Honeypots are classified into 3 major types. Firstly, it is based on level of interaction that they offer to malicious users, either high and low. Second classification is based on whether it is client-site or server-site of interaction, honeypot is implementing. Thirdly, it is based on utilization of honeypots i.e., whether they are getting used for research or production purpose.

Low/High interaction Honeypots:

Limited response capabilities towards attacker's malicious payload in low Interaction. In low interaction all services provide are not real services but they are simulated. That is why low interaction will not become vulnerable and not even infected by malicious exploit tried against the deployed simulated vulnerability. On the other hand, high interaction honeypots have no limits in term of reaction of attackers' activities. It uses real vulnerability service or vulnerability software versions. In high interaction it provides more information about how an attacker can exploit a system or how a particular malware execute in real-time.

Server/Client site Honeypots:

Server honeypot is traditional honeypot technology, it is based on protecting server from unknown attacks but they have no idea about client-side attacks. Though it is an actual production server, it exposes some software based and platform based vulnerability services and wait for intruder to attack. Whereas Client need to interact with server and to process the response of the server system. Thus for detecting client-side attacks client honeypots are generally used. The idea of client honeypot is to crawl the domain network and interact with malicious server and categorized it on the basis of their malicious nature.

Research/Production Honeypots:

Research honeypots basically gathers information about the actions of intruders. Whereas Production honeypots are actually used to protect an organization. It has Detection capabilities they can provide and the ways for which it can supplement both host and network based intrusion protection.
World Wide interaction: It performs limited activities that are not able to take out every detail of intruder. It is easy to deploy and maintain. There is least risk in this type of honeypots.

2. A POT FULL HONEY: HONEYPOT

Honeypots are valuable tools, but they're mostly used on big networks. Small companies can also gain benefit from a honeypot, but they usually haven't heard about honeypot or don't know how to set them up. One can create their own Honeypot. When a intruder finds the honeypot, he'll likely focus on that instead of diving deeper into the internal network. Essentially, the honeypot protects the internal network through distraction.

Step 1: Set Up Your Honeypot Environment

The Advantage of Linux Environment is that a few free honeypot platform available that take care of much of the configurations.[5]
To set up a virtual machine, so that hacker cannot gain access to entire physical server. Basically, in virtual machines alike main account uses different passwords. And don't store critical or confidential data on clone side of machine. You need to store dummy data & don't join the server to the network domain of the internet.

Step 2: Set Up Logging

What you install on the server is will find for what you log in. Either simple logs login attempt or create a honeypot with proprietary software installed, need to log to any application events.
You should attempt to move logs after the honeypot is breached because it is one of the issue with logging that the hacker can clear logs but it's not a guarantee that they won't be altered. Or you can set alternate logging methods. The hacker will know this OS have

logging events, but he might not have thought of alternative logging.
To preserve logs you can write the events in DVD-RW drive. You also need a Network traffic analyzer and monitoring tools to detect and capture the hacker's package.[5]

Step 3: Configure the Firewall

You should make sure that all traffic routed to honeypot should only go to the honeypot and not to the internal network. If any of the port you accidentally open on the router which directed traffic to the internal network or allowed to , then you have just exposed your internal resources to the hacker.

The goal here is to open the necessary port that will access to the honeypot not to internal network. Be very selective with ports [5]. Your Honeypot should direct the traffic to open network and never to internal network/ system.

Remember make sure that the intruder can't know that he is not on critical server but on a Honeypot. When your Honeypot has all port open, it becomes suspicious.

Step 4: Perform Your Own Testing

There are numerous port scanners and penetration testing tools available. One of the tools is nmap. Nmap is a port scanner tool that shows open ports on a system. Kippo can also list out the number of active ports on a local server. You can also use it to do a quick test to your setup. The intruder will likely perform scan on a common port.
After scanning port it will display several logs, review events and will show the thing that is not properly maintained or log. If you have any intrusion detection, it may block some uncertain activities such as port scan so that your honeypot become more believable and then you can least manipulate security settings in your IDS.

Honeypots are a fun and protective way to secure your network. You don't need to report the hacker when he assesses your honeypot, but you can learn from it. It's a great, safe way to learn the security holes in your product or network environment.[5]

Honeypot is deploying based Intrusion Detection. It is a way to attract hacker to the network system to study their movements and behaviour. It can detect the unauthorized network activities early and easily.
Honeypot can handle too much data. That can be in any readable format file. It may or may not access to Real OS. But can detect and prevent system from hackers.

FALSE POSITIVE: Honeypot side step this problem, so many alerts are generated by IDS which are false because any activity with them are unauthorized.

FALSE NEGATIVES: IDS can easily identify signatures but for them it is difficult to identify unknown attacks or behaviour. Honeypots can easily identify the earlier attacks stand out and can detect unknown activities.

RESOURCES: Even on large network or database Honeypots require minimal resources.

ENCRYPTION: Honeypots provide encryption of data where the public key is open to all but that is for the honeypot only not applicable for internal network.

3. ANOMALY BASED DETECTION

Network Behaviour is based on Anomaly. The predefined behaviour is then either accepted or trigger in the anomaly detection.
The Intrusion Detection System engine is capable at all levels, cut through the various protocols is important phase in network behaviour. The engine must be able to understand its goal and must be able to process the protocols. Through this protocol analysis it generates benefits like increasing the rule set helps in less false positive alarms.
The major drawback is defining its rule set. The efficiency of system depends on all protocols that are implemented and tested. Various protocols that are used by various vendors affect the process of rule defining. It is difficult to define rules based on custom protocols. For detection to occur correctly, the administrator develops the detailed knowledge about the accepted network behaviour. But once the

rules are defined and protocols get built up then anomaly detection system works well.

If the malicious behaviour of the user falls under the accepted behaviour then it goes unnoticed and it does not trigger any out-of-protocol, bandwidth limitation flags.

Major advantage of anomaly based detection is that it can detect the novel attack for which as signature based engine its signature does not exists. If it falls out of network traffic pattern, then system will detects network automated worms. If the network system is infected with any worm it usually start scanning for other system which can be vulnerable at an acceleration on rate filing the network with malicious traffic, this causing the event of a TCP connection abnormality rule.

3.1 User Based on Model Intention

Building the profile of normal behavior and attempting to identify certain pattern or activity deviations from normal profile. It is used to find unknown attacks by using profiling normal behaviors concept. But, momentous false alarm may be occur because it is quite difficult to attain complete normal behaviors.

Intrusion detection can be built upon multiple levels in a real computer network system. It will be choosing the features that characterize the user or the system usage patterns in the best way, such that distinguishing abnormal activities from normal activities is done clearly. Data sources like Unix shell commands, audit events, keystroke, system calls and network packages can be used. Selecting a data source is the first crucial step in building a profiling method for intrusion detection. During the early study on anomaly detection, the main focus was on profiling system or user behaviors from monitored system log or accounting log data.

ATTACK PROFILE

Nowadays there is nothing that cannot be found through some search engine or on the Internet somewhere. As Meta search engine is collaboration of search engines which can provide a precise and more relevant result of searched term from various search engines at a time.

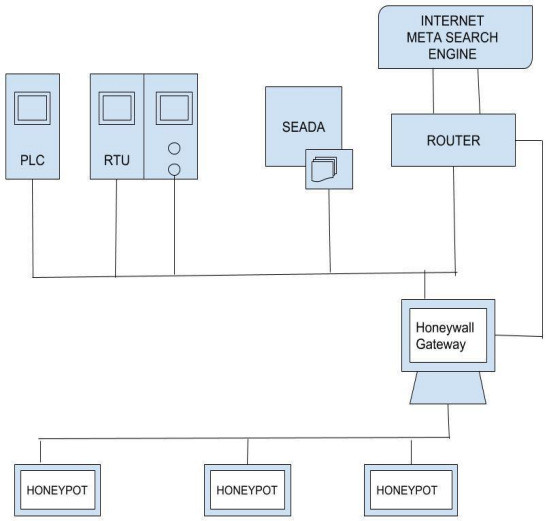

Fig 1. Honeypot Architecture

Honeypot is designed for security purpose, it will detect the intruder who is trying to enter into your internal networks. An intruder is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource. IDS used today suffer from several shortcomings in the presence of complex and unknown attacks. The problems of unknown attacks with IDS are solved using Anomaly Based Detection. Honeypot is a Distributed system helps to identify & reduce the number of spamming sites on the internet by listening the bots. If intruder wants to change the data, it will take the intruder to the Demilitarized zone which is between router and internal network. The honeypot will ask you few questions to enter into the network confidential data. It will consider the data wrong that is being tried by intruder, and take you to the page that website can't be reached or something likes that. And the data that is being entered by the intruder as keys to enter your internal network will get saved into the server's database.

It uses its own bots to crawl the internet to find and add pages to the search index. Dynamic firewall alone does not provide a holistic coverage on detecting novel & emerging pattern of attacks. Somewhere in large network is inevitable that a breach will occur. Honeypot history searched that why does that internal lurking have to go unnoticed, so honeypot can be refocused to tackle. In internal network signal get strong when something going wrong and an even stronger signal is on then it will

recognized within your organization by your important people who have the public keys or someone trying to use that key that maybe stolen and it's been not used now. A Honeypot masquerades as a real server which s open to the public. This section is known as Demilitarized Zone (DMZ). The DMZ is external to the internal network but usually behind a router connected directly to the Internet. DMZ is considered partially insecure, so it's good place for the Honeypot. The field of Honeypot research consists of 2 main pillars:

a) Development of Honeypot system and its effective deployment
b) Analysis of acquired log data in a structured manner.

There is nothing that cannot be found through some search engine or on the Internet somewhere. And meta search engine is collaboration of search engines which can provide a precise and more relevant result of searched term from various search engines at a time.

The meta search engine have 3 phases, they are :
(i) Crawling
(ii)Saving into Database
(iii) Indexing

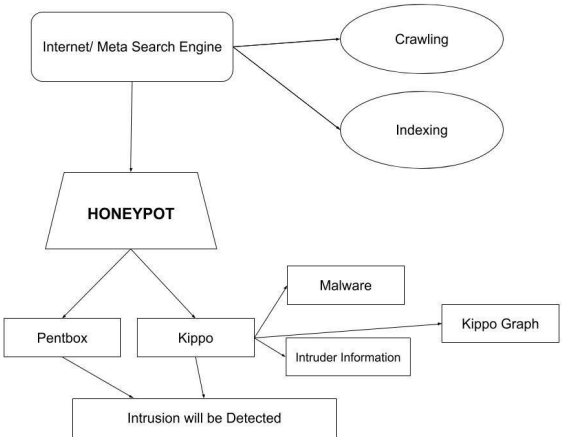

Fig 2: Proposed Architecture

Crawling is the basically scraping which continuously scrape the data in background from various websites. Can specifically save data from the site by mentioning it. Web scraping can be done manually by using software. It is the form of copying, through which data is gathered and copied from web and stored in local database or spreadsheet for further analysis. Crawling is the first stage which can be used to run again and again on different websites. The data that is collected can be large enough that cannot be saved into a spreadsheet then it may require SQL for saving content.

Second stage is saving the data into **Database** the data they gathered after crawling ,it store it in the database or in the form of csv,json or any other format want to save according to need.

Indexing is the last stage in which the queries are searched from large database. Searching through database isn't a easy task, it takes time as the database can be busy. But there are some open source search servers that can make this task easy and time effective that can search in milliseconds.

HONEYPOT INSTALLATION

Installation of Honeypot has been done after downloading the file; you simply have to import the virtual appliance to your virtual machine manager. The Oracle VM Virtual Box is recommended virtualization software; If you want to use HoneyDrive with VMware products or virtual appliance. Easily Importing HoneyDrive to VMware Fusion.

SET UP LOGGING

Pentbox is like a personal Honeypot we need to install pentbox to run our honeypot.

Pentbox is a little piece of software that allows you to open a port on your host and listen for incoming connections (eventually refused) from outside

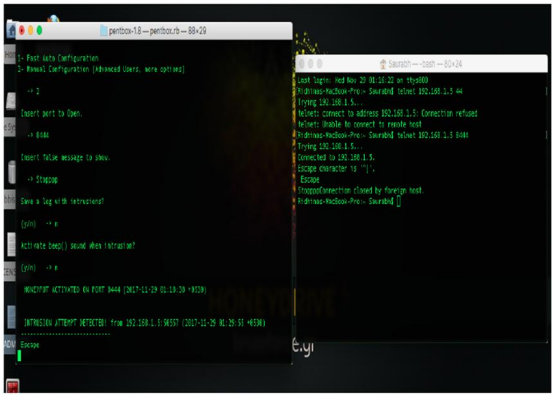

Fig 3 : Intrusion Detection through Pentbox

KIPPO HONEYPOT

Kippo is a medium-interaction SSH honeypot written in Python. **Kippo** is used to log brute force attacks and the entire shell interaction performed by an attacker. It is inspired by Kojoney.

Kippo will perform its own testing by checking the port connections of the system. All the logs which are entered by the intruder or the intruders IP addresses will be displayed on the system. Active internet connection of the server and all those are established will be displayed with the local address and the state of them will be mentioned.

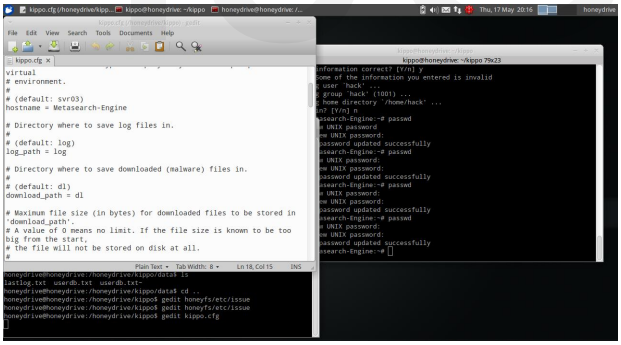

Fig 4: Virtual Environment Running Honeypot and Masquerade Real Server

And if you are trying to get access to database on the native driver port, then it will get access to the local database of the server. Self-testing of the server can be performed by using this SSH Honeypot.

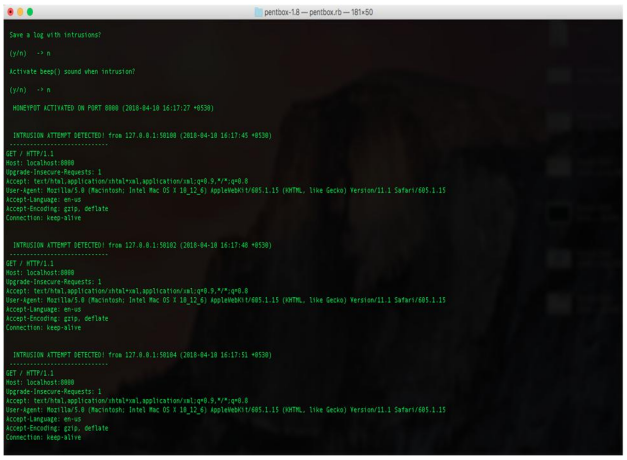

Fig 5 : INTRUSION DETECTION IN HONEYPOT

CONCLUSION AND FUTURE WORK

Now we have the idea of how to install a Honeypot and use one. This Honeypot is a Virtual appliance which helps in detecting the behavior of intruder and freezes the changes made on the server. Intruder can waste its time on the Honeypot server which looks like main server. Honeypot is masquerades as a real server which is used, so that intruder cannot manipulate the actual data.

Intrusion Detection System is something more than security that is provided to the server from intruders. It not only detects the attack of intruder but also prevent the machine from Intruder. Intrusion Detection system through heuristics, can be done either by introducing our own specific heuristics that when mismatch the behavior, it'll be considered as attack from foreign party, or heuristics can be compared with the existing algorithms to check its behavior.

If Mark Zuckerberg can be hacked and so common man even you can. The Facebook founder's hardly used other social media accounts, serving as a reminder that anyone can be hacked, everyone is inclined to hacking. Safeguards you could take include creating passwords more strong and frequently changing them. Yes, all this hurts, and it's not your flaw that the technical industry couldn't deal with the rise in security apertures. But if you do nothing, someone could break into your social media account or use your financial account to puke nasty

or vile messages. That will affect your image and you may lose some relations.

For Future work, Designing of honeypot for Social media sites where people save their confidential and critical content. They use logging credentials which and some IDS techniques which are not enough in today's technical world.

REFERENCES

[1] https://www.researchgate.net/publication/241625938
[2] https://www.researchgate.net/publication/315883340_A_Meta-heuristic_Learning_Approach_for_the_Non-Intrusive_Detection_of_Impersonation_Attacks_in_Social_Networks
[3] https://www.google.com/support/enterprise/static/gsa/docs/admin/70/gsa_doc_set/quick_start/quick_start_crawl.html#1085418
[4] https://www.supinfo.com/articles/single/5458-how-do-search-engines-work
[5] https://www.phase3.net/how-to-create-a-honeypot-to-catch-a-hacker/#
[6] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.5805&rep=rep1&type=pdf
[7] http://www.covert.io/research-papers/security/Heuristics%20for%20Improved%20Enterprise%20Intrusion%20Detection%20(Jim%20Treinen%20PhD%20Thesis).pdf
[8] https://www.usenix.org/legacy/event/detection99/full_papers/klein/klein.pdf
[9] https://link.springer.com/content/pdf/10.1007/0-387-23152-8_42.pdf
[10] http://hci.usask.ca/publications/2002/groupware-HE.pdf
[11] https://web.cs.dal.ca/~zincir/bildiri/ccece04-ajn.pdf
[12] https://www.nngroup.com/articles/ten-usability-heuristics/
[13] http://mat.uab.cat/~alseda/MasterOpt/IntroHO.pdf
[14] Heuristic Algorithm and learning Tech
[15] http://cedric.cnam.fr/~porumbed/papers/theseEn.pdf
[16] https://www.llrx.com/2002/09/features-the-meta-search-engines-a-web-searchers-best-friends/
[17] https://www.researchgate.net/publication/315883340_A_Meta-heuristic_Learning_Approach_for_the_Non-Intrusive_Detection_of_Impersonation_Attacks_in_Social_Networks
[18] https://www.usenix.org/legacy/event/detection99/full_papers/klein/klein.pdf
[19] https://pdfs.semanticscholar.org/e1ee/412f667dc9393c90caee91a982db82e0092f.pdf
[20] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.139.9691&rep=rep1&type=pdf
[21] http://airccse.org/journal/ijcseit/papers/1011ijcseit04.pdf
[22] https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/
[23] https://scrapy.org
[24] https://doc.scrapy.org/en/latest/intro/overview.html
[25] https://github.com/Parsely/python-crawling-slides/blob/master/index.rst
[26] https://www.digitalocean.com/community/tutorials/how-to-crawl-a-web-page-with-scrapy-and-python-3
[27] https://lucidworks.com/2013/06/13/indexing-web-sites-in-solr-with-python/
[28] https://www.google.com/support/enterprise/static/gsa/docs/admin/70/gsa_doc_set/quick_start/quick_start_crawl.html#1085418
[29] https://github.com/django-haystack/pysolr/blob/master/pysolr.py
[30] https://www.deepdotweb.com/2017/08/24/how-to-setup-your-own-honeypot/
[31] https://www.phase3.net/how-to-create-a-honeypot-to-catch-a-hacker/#
[32] https://www.computerworld.com/article/2573345/security0/honeypots--the-sweet-spot-in-network-security.html
[33] https://www.researchgate.net/publication/241625938_PyBot_An_Algorithm_for_Web_Crawling
[34] http://buildsearchengine.blogspot.com
[35] https://arxiv.org/pdf/1608.06249.pdf