

PHISHDECT & MITIGATOR: SDN BASED PHISHING ATTACK DETECTION

Jibi Mariam Biju¹, Anju J Prakash²

¹Mtech, CSE Department, Sree Buddha College of Engineering, Kerala, India

²Assistant Professor, CSE Department, Sree Buddha College of Engineering, Kerala, India

Abstract - Phishing is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users threat actors find a chance to gain access to critical information systems. It has become a widespread problem across every industry because this type of scam is extremely easy to pull off. Phishing can be done through the use of e-mail communication with an embedded hyperlink. the detection and mitigation of phishing attack was a grand challenge due to its complexity. Therefore, PhishDect and Mitigator, a new detection and mitigation approach using Software-Defined Networking (SDN) to identify adverse phishing behaviors is proposed. In order to classify phishing attack signatures, convolutional neural network (CNN) is used. Along with this, to cluster the different phishing attacks, K-means clustering algorithm is used.

Key Words: Cyber-attack, Software defined network, Deep packet inspection, Convolutional neural network

1. INTRODUCTION

Technology has made human life more straightforward as it brings everything to our finger tips. The invention of computers and mobile phones brought us higher attainment with time and they played a vital role in accomplishing our day to day task with ease both in professional as well as in personal lives. They are not only a simple means of gaining information and communication but also a means for data storing and data warehousing i.e we can store much important information on them. They include credit card details, passwords, user credential information etc. Most of these data are stored in plain text and can be easily obtained. Cyber criminals aim at getting the information, gaining access to information between a client and a server by spreading malware and thereby gaining unauthorized access which is known as cyber-attack. There are different types of cyber-attacks and there is a need to be aware of such attacks in order to protect ourselves from attackers.

Most common type of cyber attacks include Denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks, Man-in-the-middle (MitM) attack, Phishing and spear phishing attacks, Drive-by attack, Password attack, SQL injection attack, Cross-site scripting (XSS) attack, Malware attack. The paper "PHISHDECT & MITIGATOR: SDN BASED PHISHING ATTACK DETECTION AND MITIGATION" gives a description about the phishing attack and it can be detected and prevented.

1.1 OBJECTIVE

Phishing is one of the most harmful social engineering techniques in which end users can access critical information systems. It can be done through an e-mail communication with an embedded hyperlink. Due to the complexity of the phishing attack, its detection is a great threat. Current methods are often too time-consuming for detection and mitigation time to be used in the real world. Hence "PHISHDECT & MITIGATOR: SDN BASED PHISHING ATTACK DETECTION AND MITIGATION" is proposed.

- New Deep Packet Inspection (DPI)[1] technique and then use Software-Defined Networking (SDN)[2] to identify phishing activities via email and web-based communication.
- Based on the programmability of SDN, Store and Forward (SF) mode and the Forward and Inspect (FI) mode is developed to direct network traffic by using Convolutional Neural Network (CNN) model to classify phishing attacks.
- In order to classify phishing attack signatures, convolutional neural network (CNN) is used.
- To classify the different phishing attacks, K-means clustering algorithm is used.

1.2 SCOPE

Phishing is a social engineering attack aimed at exploiting the system user's weakness found in system processes. For example, a system may be sufficiently secure against password theft, however unaware end users may leak their passwords if they are asked by an attacker to update their passwords via a given Hypertext Transfer Protocol (HTTP) link, which ultimately intimidate the entire system security. In addition, technical vulnerabilities like Domain Name System (DNS) cache poisoning can be used by attackers to build much more convincing socially engineered messages. This makes phishing attacks a layered

problem, and effective mitigation on the technical and human layers would require addressing issues. Because phishing attacks aim to exploit the weaknesses found in humans, they are difficult to mitigate.

Phishing has become a widespread problem in every industry because it is extremely easy to remove this type of scam. Anyone having internet access and grasp of language can do this. Every year, countless firms find the hard way to invest more time and money for their employees in phishing protection training. The 3 main purpose of attackers behind phishing attack are financial gain, identity hiding, Fame and notoriety. Phishing attack detection and mitigation are important so as to save money, protecting customers as well as company reputation, to keep corporate secrets safe, to avoid black mailing and identity theft, to eliminate brand abuse etc.

2. METHODOLOGY

2.1 EXISTING SYSTEM

Phishing attack is a very common approach to social engineering aimed at an organization and end users. It has become one of today's most harmful attacks. Numerous studies have been carried out to detect and mitigate phishing attacks. Inline inspection techniques like IPS or proxy service based on IDS and static string-matching techniques based on BRO [3] are the traditional methods used. The existing system is PhishLimiter which is a new solution to thwart phishing attacks. PhishLimiter has the ability to handle network traffic dynamics for containing phishing attacks and can provide a better traffic management. An ANN model using a PLS system was developed to classify phishing signatures. It is a system for SDN traffic flow engineering by introducing PLS and the OVS switching score for SF and FI modes that can use the programmability of SDN to deal with the dynamics of phishing attacks in the real world. The inspection approach of two SF and FI modes within PhishLimiter detects and mitigates phishing attacks before reaching end users if the flow has been determined untrustworthy.

The phishing signature is classified using the ANN (Artificial Neural Network) model. It is a computational model based on the neural biological network structure and functions. ANNs are considered to be nonlinear statistical data modeling tools where modeling or patterns are found for the complex relationships between inputs and outputs. ANNs have three interconnecting layers. The input neurons constitute the first layer and these neurons sent data to the second layer which then sends to the output neurons to the third layer in turn.

2.1.1 DISADVANTAGES OF THE EXISTING SYSTEM

- Dependence on hardware: Artificial neural networks require parallel processing processors according to their structure. The realization of the equipment is therefore dependent.
- Network unexplained behavior: This is ANN's most important issue. When a probing solution is produced by ANN, it does not give any indication as to why and how. This reduces network confidence.
- Determination of proper network structure: There is no specific rule for artificial neural network structure. Through experience and trial and error, appropriate network structure is achieved.
- Difficulty showing the problem to the network: Numerical information can be used by ANN. Problems must be translated into numerical values before entering ANN. The display mechanism to be determine will directly affect the network's performance which depends on the ability of the user.
- The duration of the network is unknown: Reducing the network to a certain value of the sample error means the training has to be completed. This value does not give us the best results.

2.2 PROPOSED SYSTEM

The title of the proposed system is " PHISHDECT & MITIGATOR: SDN BASED PHISHING ATTACK DETECTION AND MITIGATION " to identify and mitigate adverse traffic communication. SDN is an emerging technology that separates its control plane from its data plane for better traffic management and aims to overcome limitations of legacy networks. The idea of DPI utilizes a two-type inspection approach of SF and FI where we consider the two SF/FI path as slow or fast lanes, respectively. Store and Forward lane: In SF approach, the information is kept, inspected and classified for analysis when a packet enters a switch. If a packet's analysis has not resulted in malicious intent, the information will be forwarded to the destination concerned. Forward and Inspect approach: In this approach the packet is suddenly forwarded to the concerned destination and only a copy of the information is temporarily stored for inspection. Then the packet is evaluated using a series

of classification and feature extraction techniques during the inspection process and developed using CNN to identify the intended malicious packet.

The proposed system helps to classify the phishing signatures using CNN (Convolutional Neural Network). CNN is a type of artificial neural network that is specifically designed to process pixel data for image recognition and processing. CNN uses a system much like a multilayer perceptron that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers and normalization layers. Removing limitations and enhancing image processing efficiency results in a system that is far more efficient, easier to train for image processing and natural language processing.

3. SYSTEM DESIGN

There are 6 main modules in "PHISHDECT & MITIGATOR: SDN BASED PHISHING ATTACK DETECTION AND MITIGATION".

1. Configuration
2. Packet Capture
3. Feature Extraction
4. Classification
5. Clustering
6. Evaluation

3.1. Configuration module

In this module, two main configurations are set up.

- Threshold configuration
- OVS configuration

Threshold configuration: The threshold may be defined as a minimum or maximum value set for an attribute, characteristic, or parameter that serves as a benchmark for comparison or guidance, and any infringement of which may require a complete review of the situation or a system redesign. It is set for both packet and openvSwitch. Although network communication can traverse through either the SF or FI lane, the deciding factor is PhishLimiter Score (s). Moreover, each SDN flow has a s value based upon their source of either an IP address or interface. Let s be a PLS value and s_{j-1} be the PLS value for packet $j - 1$ where $j \geq 1$. Let s_{OVS} be a threshold value for PhishLimiter in determining whether F should be placed in SF or FI, where s_{OVS} is predefined by the OVS. If $s_{j-1} > s_{OVS}$, then F will be placed in SF mode otherwise F is placed in FI mode.

OVS configuration: In order to get the OVS value, network simulator is needed. As the simulator is hardcoded, less programmable and has no datamining functionalities. OVS value configuration is done as an application. It is assumed that all the packets having certain pls value having OVS value in a specific range are passed through a particular openvSwitch. In this module, switched, macid, ovsto and ovsfrom are set and can be updated.

3.2 PACKET CAPTURE

Incoming packets from a network is captured and parsed in this module. The features like packet length, source IP, destination IP, TTL and protocol can be obtained. Packet Capture is a networking term used to intercept a data packet that crosses a particular point in a data network. Once a packet is captured in real time, it is stored so that it can be analyzed and either archived or discarded. Packets are captured and examined to help diagnose and resolve network problems such as identifying security threats, troubleshooting undesirable network behaviors, identifying network congestion, identifying data / packet loss, forensic network analysis. It is possible to capture whole packets or specific portions of a packet. Typically, packet capture data analysis requires significant technical skills and is often done with tools like Wireshark. Wireshark is a packet analyzer that is free and open source. Wireshark captures packets and enable to examine the contents. As the number of features obtained is not sufficient to recognize it as malignant or benign, an external data set contain 30 packet features are used.

3.3 FEATURE EXTRACTION

A plaintext decrypted from cryptomodule is processed using url regular expression. The types of phishing attack [4] features that are extracted can be classified into 3.

- URL-based features: Features 1 through 13 contain the phishing characteristics of URL.
- HTML-based features: Features 14 to 23 is used to detect the anomaly of HTML and JavaScript code.
- Domain-based features: Features 24 to 30 identifies the domain information from the URL.

Feature selection can be done with the help of weka classifier. Weka is a collection of algorithms for data mining tasks in machine learning. You can either apply the algorithms directly to a dataset or call them from your own Java code. Weka includes tools for pre-processing, classification, regression, clustering, rules of association and visualization of data. It is also suitable for the development of new machine learning schemes.

3.4 CLASSIFICATION

As the number of features obtained is not sufficient to recognize it as malignant or benign, an external data set contain 30 packet features are used. This uploading of the dataset management is done in the classification module. Then these packets are classified using machine learning based classifiers. ANN was used for previous classification purpose. The proposed system uses CNN for packet prediction. CNN is a type of artificial neural network that is specifically designed to process pixel data for image recognition and processing. CNN uses a system much like a multilayer perceptron that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers and normalization layers. . Removing limitations and enhancing image processing efficiency results in a system that is far more efficient, easier to train for image processing and natural language processing.

3.5 CLASSIFICATION

In this module, a clustering algorithm is applied to group phishing activities so that the most occurred phishing can be traced out. The clustering algorithm used is K-means. Clustering is one of the most common techniques used in exploring data analysis to gain an intuition of the data structure. The task of identifying subgroups in the data can be defined as such that data points are very similar in the same subgroup (cluster) while data points are very different in different clusters. Cluster analysis can be performed on the basis of features that attempt to find sample subgroups based on features or samples that attempt to find sample-based subgroups of features.

3.6 EVALUATION

- CNN evaluation based on (a) mode (b) time complexity.
- Feature Analysis.
- Time Complexity based on number of records.
- Clustering.

4. RESULT AND ANALYSIS

4.1 RESULT

This section discusses the experimental results of finding the accuracy of predicting whether a packet is benign or fraud. The system that uses the operating system for windows 10 and windows platforms here is c#.net. And the database created is a SQL server. Using PhishDect, it is possible to thwart a threat varying on both inspection modes of SF and FI. CNN algorithm is used for the classifying the predicted result which is more efficient than the ANN machine learning algorithm. Moreover, for clustering the predicted result K-means algorithm is used.

4.2 ANALYSIS

This system provides accurate result to the users. CNN algorithm used instead of ANN for the prediction of the result found to be more efficient. Here the graph for feature analysis, clustering, time complexity based on the number of records and CNN evaluation based on the mode and their complexity is evaluated in this section.

4.2.1 CNN Evaluation

CNN Evaluation can be done based on the mode and their time complexity. The precision, recall and accuracy are calculated for each mode i.e, SF and FI. Based on this mode of network traffic, its performance measure is evaluated and is shown in figure 4.1.

As the number of records in SF mode is more as compared to the number of records in the FI mode, the time required for the calculation of precision, recall and accuracy is more in SF mode. The graph representing the time complexity against the mode of network traffic is also shown in the figure 4.1.

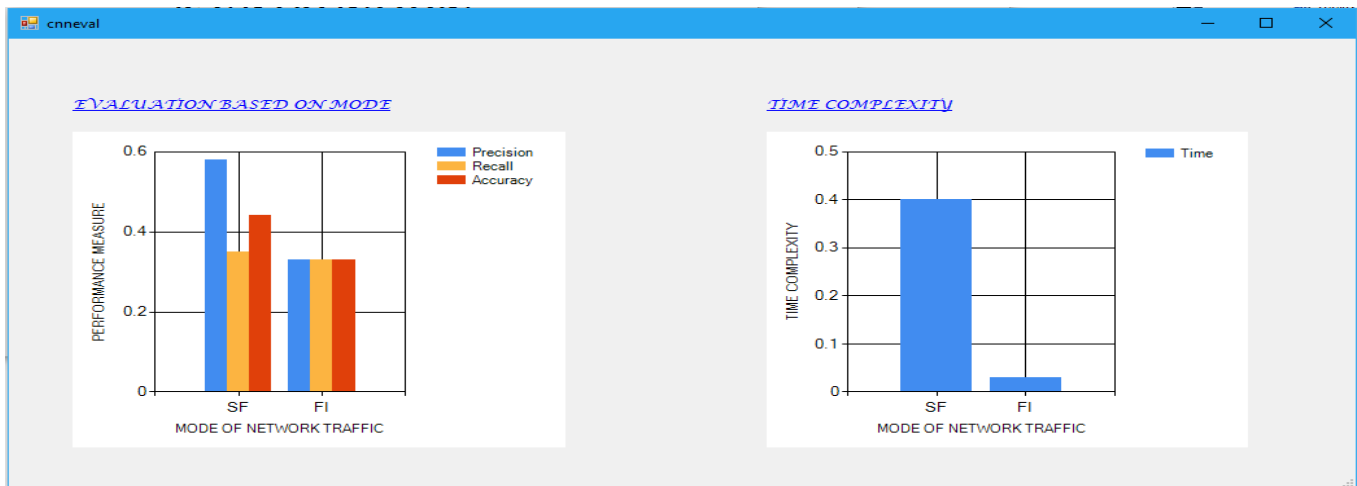


Fig 4.1 Evaluation based on mode and time complexity

4.2.2 Feature Analysis

The features that are more suitable for predicting whether the packet is benign or fraud can be done with the help of feature selection. The selected features and their info_gain in each mode is also represented graphically in the figure 4.2.

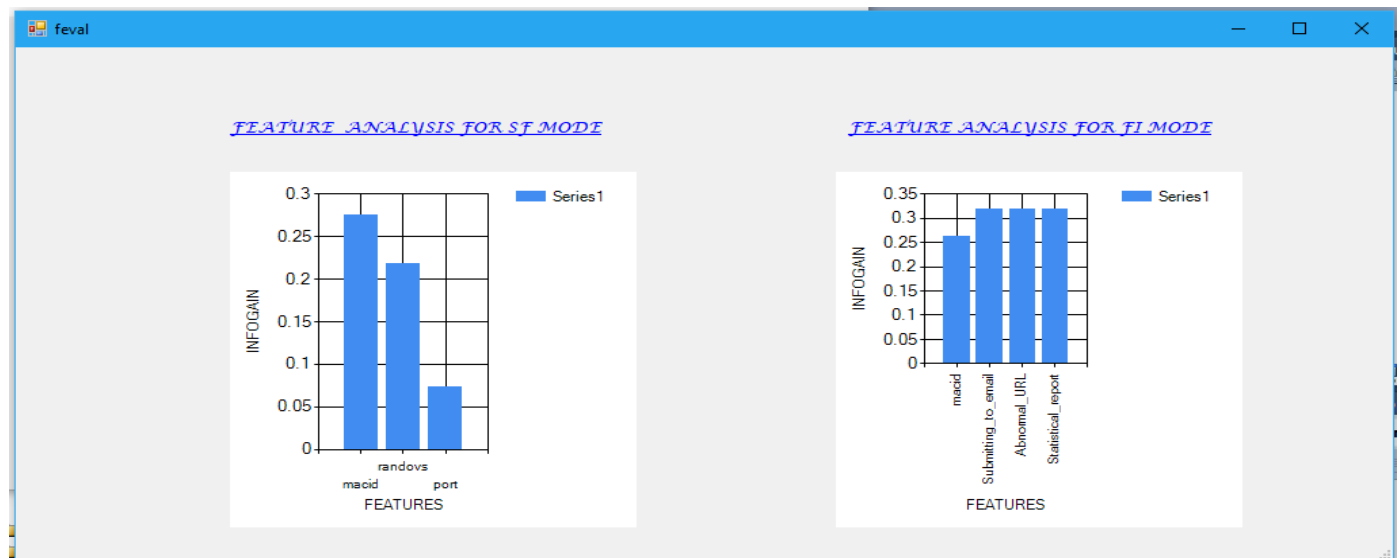


Fig 4.2 Feature analysis based on the mode

4.2.3 Time complexity based on the number of records

As the number of records increases, the time required for feature selection also increases. The number of records in the SF and FI mode are different and its time complexity is calculated for different number of records and is shown in figure 4.3.

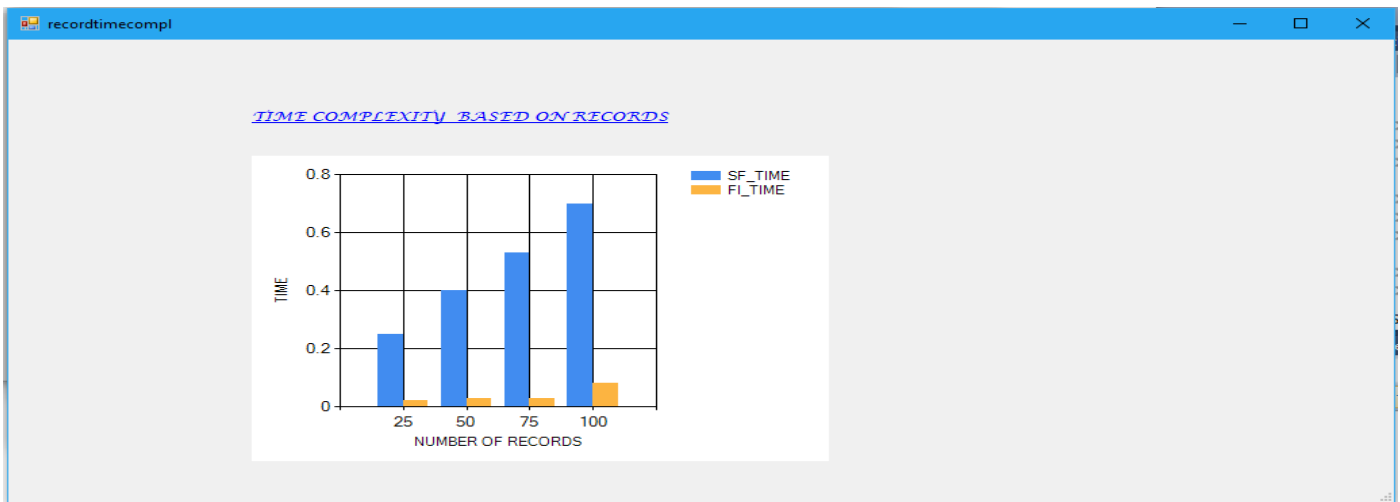


Fig 4.3 Time complexity based on the number of records.

4.2.4 Clustering Analysis

The time required for clustering also varies as the number of records increases. The time complexity in clustering the different number of records is calculated in the figure 4.4.

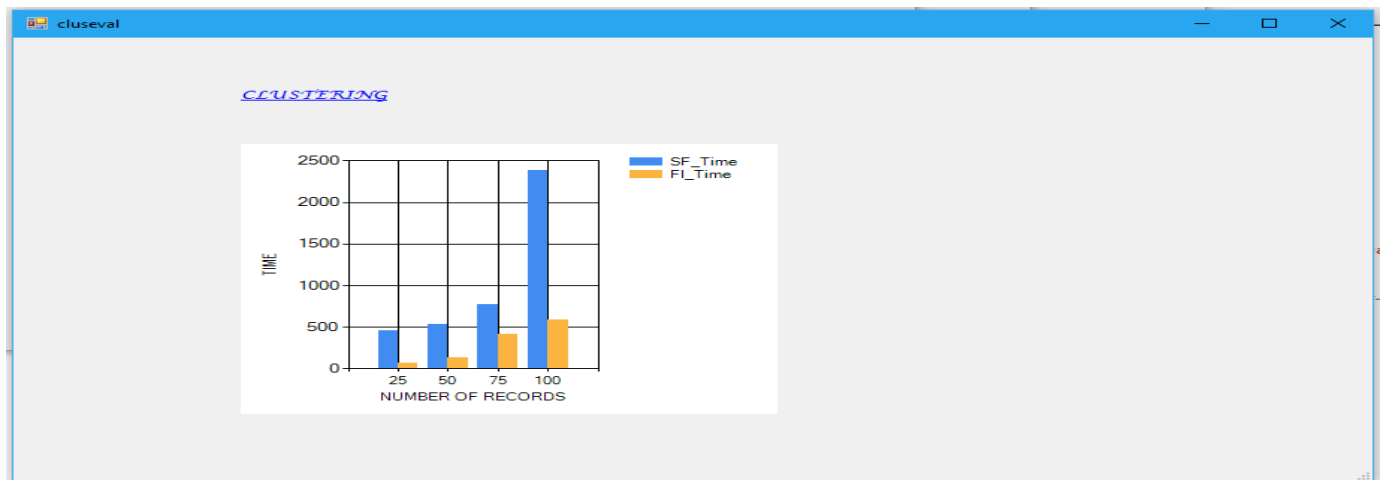


Fig 4.4: Time Complexity of Clustering

3. CONCLUSION



Phishing is the simplest kind of cyberattack and, at the same time, the most dangerous and effective. It is a crime of deceiving people into sharing sensitive information like passwords and credit card numbers. The user may receive a malspam and tempts the user to click the link that may take them to an illegitimate website. The prime target of phishing attack is social engineering sites that aims users to get sensitive information. Therefore, PhishDect and Mitigator is proposed as a solution that helps in the detection of phishing attacks. With the help of SDN that prefers two modes, it is possible to reduce the network traffic and provides better management. A CNN model is developed to classify the phishing attack and shows better result even though there are missing layer in the neural network. In order to cluster the phishing records, K-means clustering algorithm is used.

REFERENCES

[1] A. Bremler-Barr, Y. Harchol, D. Hay, and Y. Koral, "Deep packet inspection as a service," in Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies. ACM, 2014, pp. 271–282.

- [2] Tommy Chin, Member, IEEE, Kaiqi Xiong, Senior Member, IEEE, and Chengbin Hu, "PhishLimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking" IEEE Access.
- [3] V. Paxson, "BRO: a system for detecting network intruders in real-time," Computer Networks, 1999.
- [4] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, "Phishing Websites Features".

BIOGRAPHIES

	Jibi Mariam Biju, she is currently pursuing M.tech in Computer Science and Engineering in Sree Buddha College of Engineering, Elavumthitta. Her research areas include the field of data mining, cryptography and security.
	Anju J Prakash is working as Asst.Professor in computer science and engineering in Sree Buddha College of engineering, meanwhile pursuing her PhD in the field of image processing or data mining from Noorul Islam Centre for higher education.