International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 65-71, 2020, ISSN:-2319–8656

A Hybrid Data Loss Detection and Prevention Framework Using Snort Signature Based Detection System and Knowledge Based Anomaly Detection System

Henry Gekone Ondieki[1]
Jomo Kenyatta University of Agriculture and Technology
Masters Student
Nairobi, Kenya

Dr. Kennedy Ogada[2]
Jomo Kenyatta University of Agriculture and Technology
School of Computing.
Nairobi, Kenya

Prof. Wilson Cheruiyot[3]
Jomo Kenyatta University of Agriculture and Technology
Associate Professor of Computer Science.
Nairobi, Kenya

Abstract: In the modern day and era, data has become part and parcel of daily life and business. The concerns for data security have as a result of this, emerged as a major concern when seeking to prevent data leakages and data losses. The need to prevent unauthorized access to data has become a major factor affecting the survival of organizations today due to the consequences that could arise when data falls into the wrong hands. For instance, the level of credibility and trust-worthiness of various companies would be put into question wen sensitive data becomes accessed by unauthorized people. The existing traditional data security methods have not been enough preventative mechanisms to prevent loss and leakage of sensitive data. This calls for the development of a new and improved data security architecture creating the new data leakage prevention systems (IDSs/IPSs). Burgeoning research has seen new innovations and increased funding towards improvement of data security architecture. This study makes a contribution trough use of DLPs to propose hybrid data loss detection and prevention system. Signature based solutions provide accurate identification of the attacker and thus suitable for prevention, they cannot be used when unknown attacker or the attacker who uses different path attacks the system, also anomaly-based solutions can detect the unknown attacker but the false positive results are high thus limiting their allocation on systems. Due to this, in this thesis we propose a hybrid system which combines both the signature based and anomaly-based solutions which enables the detection and prevention of data loss.

Keywords: Hybrid Data Loss Detection, Snort Signature Based Detection System, Knowledge Based Anomaly Detection System

1. INTRODUCTION

In the current digital world, large amounts of data are being processed at every single moment due to the combination of various aspects of technology that include cell phones, internet and cloud computing which are part of business and daily life [1]. The amount of data that is processed on a daily basis is 2.5 quintillion bytes and it is on a steady growth on a steady growth because of an increasing data demand [2]. A major challenge facing the data industry is data loss and the need to protect information which is a key responsibility of data companies. This has seen the invention of methods to protect data from leakage. This led to the creation of Data Leakage Prevention systems (DLPs) which are meant to protect against data leakage and reduce the damage caused by data leakage [3].

DLP solutions ensure they carry out duties of prevention and to detect any attempts of unauthorized access to obtain sensitive data. Potential data breaches are prevented using the DLP in a way that is timely, which reduces the possible effects it could have created [4]. Signature-based IDS refers to the detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. The terminology is generated by anti- virus software, which refers to these detected patterns as signatures. Even though signature-based IDS can easily detect known attacks, it is impossible to detect new attacks, for which no pattern is available. This technique automatically contains the signature to detect an intruder [5].

The anomaly-based intrusion detection system refers to a method used in the detection of computer and network intrusions, and possible misuse through having a system that monitors activity and categorizing it to be either anomalous or normal. The categorization follows certain rules instead of signatures or patterns, and ascertains activity that is malicious and which do not appear to be normal operations of the system. Conversely, systems that are signature based are able to only detect types of attacks which already have previous signatures [5]. Knowledge based detection method is one of the Anomaly based IDS. In Knowledge based detection, knowledge is gathered on the attacks on data, and this knowledge is then used to detect any attacks or system vulnerabilities. When a system lacks knowledge about a particular attack, then it is not capable of identifying an attack. This means that the model requires a significant amount of knowledge on several attacks [6].

2. PROBLEM STATEMENT

Over the recent years, the challenge of ensuring data security has been acknowledged, and several methods have been developed to address the problem. Data leakage, corruption and loss are the main goals of attackers, and thus most of the methods that have been proposed aim at addressing this problem. Examined literature shows concerted efforts of data loss prevention systems.

A generic model of malicious behavior, distinguishing motives, actions and associated observables is in existence[7].

The study developed several prototypes that provided early warning of malicious activity including use of network traffic profiling, honey tokens and knowledge-based algorithms for data fusion and structure analysis. [8] proposed a hybrid intrusion detection system that detected and prevented malicious activity on a network. The method proved effective on evaluation for preventing data loss on a network and in dealing with considered attacks on data. [9] proposed a hybrid intrusion detection system in which a normal profile at activity intervals is first detected and then subsequently used to detect any anomalies in behavior nodes. Through considering anomaly types that detected, intrusion is then detected using predefined expert knowledge rules.

Existing data loss prevention methods continue to reveal loopholes that have been used by attackers who end up gaining unauthorized access to data. There is a need to develop and improve on existing data loss prevention methods. More secure data loss prevention methods will simply improve security architecture for data which will ensure accuracy, completeness, veracity and reliability of data. This study develops a hybrid data security method to protect against data loss by combining anomaly-based intrusion detection systems and signature-based intrusion detection system.

3. MAIN RESEARCH OBJECTIVE

The main objective of this study is to design and propose a hybrid data loss prevention system using a Snort Signature Based IDS and an Anomaly Knowledge Based Detection IDS.

3.1 Specific objectives

The specific objectives of this research are:

i. To study how the existing data loss prevention systems, work.
ii. To design a hybrid data loss prevention system using Snort Based Signature IDS and anomaly Knowledge based detection IDS.
iii. To evaluate the designed hybrid data loss prevention system.

The main aim of the hybrid data loss prevention system is to provide better security that prevents data loss in addition to providing a more secure hybrid data security system. An intrusion detection system (IDS) normally analyzes data from a network and detects any actions that are malicious and behaviors that may compromise the security of data. When activities that are malicious are detected, an alarm is raised. A hybrid data security combines the knowledge based detection anomaly and Snort signature based IDS to provide a more secure data security method which is the primary objective of this study.

3.2 Research questions

i) What are the existing data loss prevention methods and their shortcomings?
ii) How to design a hybrid data loss prevention system using anomaly Knowledge based detection model and Snort signature-based IDS?
iii) How will this developed hybrid system be evaluated?

4. LITERATURE REVIEW

4.1 Snort Based IDS

Snort is an open and free system for prevention of intrusions created by Martin Roesch in 1998.

4.1.1 Components of Snort

Snort is categorized into different components that all work in collaboration with each other to detect specific attacks and for the generation of output from the detection system a format that is pre-specified [10].

The packet decoder selects packets from different network interfaces and prepares them to be pre-processed or sent to the detection engine. Pre-processors are components or plug-ins that are to be used with Snort for modifying or arranging packets before performance of any operation by the detection engine in order to establish if any packets have been exploited by an intruder. The detection engine is responsible for detecting any intrusion that may exist in a packet based on rules of Snort [10].

4.1.2 Anomaly Based IDS

Anomaly detection methods are very important in intrusion detection systems because an intrusion activity is different from the normal activity of the system. Anomaly based IDS uses a reference of a pattern of normal system activity that has been learned or baseline in order to establish active intrusion attempts. Any behavior that fall outside the accepted model of behavior or pre-defined pattern creates an anomaly [6.]

4.1.3 Knowledge Based Detection

Knowledge based detection obtains knowledge about attacks and this knowledge is then used to detect any system vulnerabilities or attacks. They are useful in recognizing transitions that occur when an intruder penetrates the security system. Expert systems contain sets of facts, rules, and inferences methods. Each event that occurs in the system is translated into corresponding rues and facts and inference methods are able to generate conclusions from the existing rules and facts. Signature analysis contains similar knowledge acquisition approach as in an expert system, but the way of knowledge acquired is different. The exact evidence of every attack is available in the audit trail and this information is consolidated as sematic description of attack [6]

4.2 Data Leakage Prevention (DLP)

Data Leakage Prevention (DLP) refers to one of the specialized philosophies and arrangements that are used to protect information that is delicate from being accessed by unauthorized people who are either inside or outside the organization [11.] DLP is a method used to conceal the secrecy of information being gotten to by unapproved client [12]. DLP arrangements address information leakages through three different categories by using specific types of technology arrangements [13; 14; 15].

4.3 Data security threats and vulnerability

To begin with, issues of data security and protection have grown as a result of volume, speed, and assortments. For instance, large scale infrastructure used for cloud computing, rich sources of data and configurations, nature of information spilling, and large volume between cloud relocation [16]. Further, usage of these large-scale infrastructures for cloud computing are spread around the globe with different types of software have seen an increase in system attacks and thus traditional systems for security are inadequate. Further, better and more improved technologies that are able to quickly respond to the growing demands of streaming data across several centers of data [16].

4.4 Data loss prevention Techniques

Techniques for data loss prevention can be grouped into non-sensitive data and sensitive data and thus techniques or detective purposes are categorized into two main areas i.e. content-based analysis techniques and context –based techniques which are discussed under [17;18].

4.4.1 Context analysis technique:

This technique works through considering metadata (format, source, size, timing, and destination) that is often linked to the actual confidential data without emphasizing on how sensitive the data is. The Dkey contextual features such as size, source, timing, and destination would be examined. The system then compares these features with certain patterns of transactions or pre-defined policies. This technique is sometimes combined with content-based analysis techniques for it to be effective.

4.4.2 Data fingerprinting:

This is the most widely recognized strategy which is utilized to distinguish information spillage. In many DLPSs, an entire record can be hashed utilizing ordinary hash capacities, for example, MD5 and SHA1, where the hash values of each single delicate archive are stored in nearby machines or databases. These DLPs are able to have 100% accuracy in identifying whether a record changed by using any and all available means. Records that are secret are prone to being changed; DLPs may be inadequate because hash esteem cannot defend itself from change. This then means conventional fingerprinting methods are infective to prevent significant changes to the data. This can be solved using more advanced fingerprinting methods for instance Rabin's randomized fingerprinting and fuzzy fingerprinting.

4.4.3 Regular expression:

Most DLPs use this famous technique. These DLPs are made of set of terms or characters that are utilized to frame location designs. These examples will be utilized to match and think about arrangement of information strings numerically. This procedure is for the most part utilized as a part of web crawlers and content preparing to approve, extricate and supplant information. In any case, as far as data security, consistent articulation is utilized generally in information examination for vindictive codes or secret information.

4.4.4 Statistical examination:

This strategy can encourage certain devices, for example, machine learning order and data recovery term weighting. Most part of them depends on the terms and n-grams recurrence inside arrangement of archives. The disadvantage of consistent articulation and information fingerprinting were tackled by N-gram measurable examination procedure. A term basically implies a word, while a n-gram may be bits or a word E.g. unigram (one character), bigram (two characters) and trigram (three characters).

4.5 Preventive method

Policy and Access Rights: Prevention of possible data leaks using strict access controls. Some organizations have policies that restrict use of CDs and USB.

Virtualization and Isolation: these are used to protect data that is sensitive from through h ensuring the creation of virtual environments when accessing data that is sensitive. Access that is allowed will be the only one permitted.

Cryptographic Approaches: This is a method used to hide data that is sensitive from being accessed by users who are not authorized by using algorithms and cryptographic tools. [19] used Attribute Based Encryption (ABE) algorithm a prevention method for data leakage which allowed sensitive data to be preserved. This is a preventative method. The system worked by keeping sensitive data locked and only allow users that are authorized to access it. The idea is that reliance on detective approaches can result in data leakage. Encryption prevents such data losses from occurring.

Quantifying and Limiting: Security administrators use this method to pretend to be system attackers and block all possible loopholes that lead to data that is sensitive by making attacks on their own systems. This approach is applicable for both detection and prevention methods.

4.6 Detective Method:

Data Identification: Refers to the way in which data that is sensitive is detected depending on the previous knowledge of content and some techniques such as data fingerprints among other types of matching.

Social and Behavioral: analysis of Patterns and behaviors on Social network can enable detection of irregularity and raise alarms to enable security administrators take action.

Data Mining / Text Clustering: Data mining areas have capabilities to perform advanced undertakings, for example, inconsistency location, bunching and order by removing information designs from vast datasets. Identification of Information mining by using machine realizing that has algorithms to establish patterns that are complex and enable better decision making. Clustering of texts is related to retrieval of information that is essential in DLPs.

5.0 Methodology

In this study experimental research design is used. Publicly available datasets were used in the literature for testing Intrusion Detection Systems. Such datasets served as a benchmark for the various parameters of an IDS like false positives, false negatives and detection rates. This also served the purpose of analyzing such parameters in relation to other existing IDS systems. Parameters like true and false positives rates, true and false negatives rates are the most important parameters of any IDS. These parameters are a measure of the effectiveness of the detection mechanism of an IDS. For the given dataset, the results of these parameters were analyzed for different thresholds and r-values. ROC curve analysis was used to test the effectiveness of the IDS systems.

6.0 Findings and Discussions

The developed system was tested against the dataset and parameter values like false positives, false negatives and detection rates were calculated. ROC curves were also plotted for the obtained results. Testing was repeated with different values of n, r, and anomaly threshold values and the corresponding rates were calculated.

6.1 Processing data

Some pre-processing had to be done on the dataset before testing it on the hybrid data loss detection and prevention system. Since the negative selection module accepts only self-traffic as an input for training, a Python script was written to retrieve only the normal traffic from the labeled training data.

Also, since the self-data has to have strings that are of the same length, the records were padded with the character '0' so that all the records were the length of the largest string in the le. Similarly, a different Python script was used for calculating false positive, false negative, and detection rates from the logs of the developed hybrid data loss detection and prevention system.

6.2 Detection and false alarm rates

False positive and detection rates were calculated by running the test data dataset. These rates were obtained for different n, r and threshold values. Since the length of strings in the test dataset was 152 (including the added padding), the n values tested were between 70 and 120. Higher n values only resulted in poorer detection rates and insufficient heap memory. As stated earlier, r values were tested between 40% and 80% of the given n value. Anomaly threshold was varied between 10% and 100%. Highest detection rate obtained was 81.56% for n=100, r=40 and threshold = 20%. But the corresponding false positive rate was also high at 39.52%.

By varying the above mentioned parameters, the false alarm and detection rates were calculated. Table 4.1 shows the different average rates for given n values. It can be seen than an n value of 100 yields the best average detection rate of 64.90%. But this also resulted in a moderate average false positive rate of 31.37%. As n value was increased above 100, the average detection rate started to decline. This may be attributed to the increased noise being included for generation of detectors. Also with higher n values, the system ran out of heap space and crashed abruptly. This can be attributed to huge storage overhead associated with generation of large detectors. The percentage of test cases crashed for given n values can be seen in the Figure 4.1.

6.3 ROC Curves

ROC curves are plots of detection rates against false positive rates. Such curves indicate the operating region of a system and also lets the network administrator decide on the operating region that is suitable for a given network environment. As mentioned earlier, the points (0,0) and (1,1) occur at the worst operating conditions of any hybrid data loss detection and prevention system. Figure 4.2 shows the ROC curve for the tested dataset with these two points. This curve was plotted using about 565 tested cases, obtained by varying n, r and threshold values.

Table 4.1: Comparison of average rates for given n values

	Average False	Average	Average False
n-value			
	Positive rate (%)	Detection Rate (%)	Negative Rate (%)
70	28.81126149	61.32805454	38.67195
75	31.37546036	59.14728688	40.85271
80	32.26197881	58.73763987	41.26236
85	32.56920446	58.97770202	41.0223
90	32.59815859	59.43509852	40.5649
95	32.25811202	59.92991625	40.07008
100	31.37725954	64.9015537	35.09845

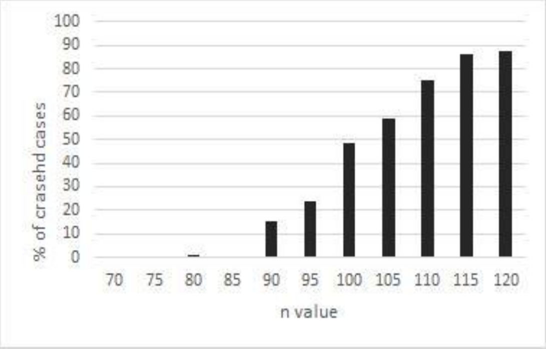

Figure 4.1.: Percentage of crashed test cases Vs. n value

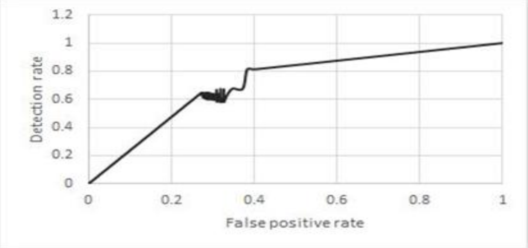

Figure 4.2.: ROC curve including points (0,0) and (1,1)

Figure 4.3 shows the actual ROC curve without the points (0,0) and (1,1). This curve is generally convex and does not have abrupt drops in slope. The region from false positive rate of 0.27 to slightly after 0.32 can be considered as ideal operating region for this system. With the increase in false positives beyond 0.27, there is no abrupt drop in detection rate, indicating a wide operating region with decent false positive and detection rate. There is a slight concave region in the curve where false positive rate is 0.37. This is a sign of poor detection mechanism in that region, but since it is outside the ideal operating region of the system where the false positive rate is already higher, this is not a huge compromise on the effectiveness.

4.4 Effect of r-value on detection and false positive rates

The r values tested ranged from 40% to 80% of the value of n (calculated from (r/n)*100). The false positive and detection rates against various r value percentages can be seen in the Figure 4.4. Individual detection rate (at 81.56%) and average detection rate (at 62.68%) were the maximum when the r value was 40% of n value. But the average false positive rate was also slightly higher at 30.50%. It is to be noted that these are values are specific to the dataset and may not always be true for another dataset. Ideal r and n values for a different dataset can be deduced by similar testing.

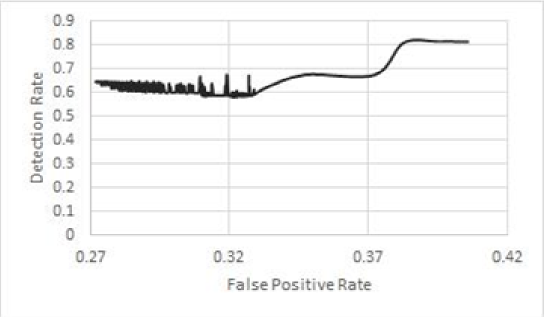

Figure 4.3.: ROC curve

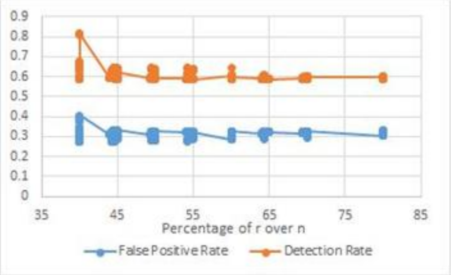

Figure 4.4.: False positive and detection rates against percentage of r values over n

4.5 Comparison of effectiveness with other IDS systems

Figure 4.5 shows the comparison of false positives and detection for different r values and constant n and threshold values. This graph shows the direct effect of r value on the detection rate. The n value was set at 75 and threshold value was set at 70%. It could be observed that as the r value increases, detection rate decreases and the false positive rate increases. As already seen in Figure 4.1, higher n values resulted in heap spa running out and the program crashing. The insufficient heap space issue further increased with higher r values. It can be concluded that for the dataset, the effectiveness and performance of the system decreases beyond an r percentage of 40%.

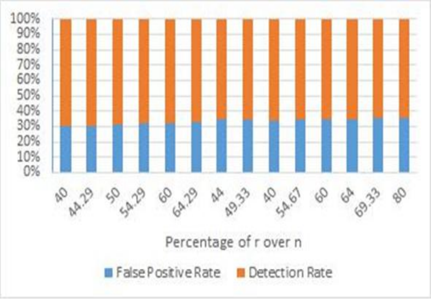

Figure 4.5.: False positive and detection rates against percentage of r values over n (constant n and threshold values)

The Figure 4.6 shows the ROC curves of the developed hybrid data loss detection and prevention system along with the other prominent systems like Snort (Alder et al., 2007), Bro (Paxson, 1999), Hybrid Intrusion Detection System (HIDS) proposed in Hwang, Cai, Chen, and Qin (2007). ROC curves for these three IDS systems were already compared in Hwang et al. (2007). HIDS is a combination of anomaly-based and signature-based detection mechanisms and would serve as a proper comparison. HIDS having a combination of anomaly based and signature based detection mechanisms, performs better overall compared the other three IDS systems. The operating range of HIDS has a detection rate around 30% better than Snort and around 38% better than Bro. In comparison with the developed IDS, HIDS has almost an identical detection rate range in the operating region before a false positive rate of 0.32

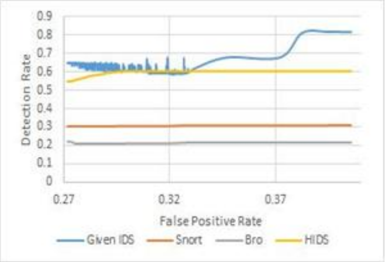

Figure 4.6.: ROC curves - comparison against other IDS systems.

4.6 Effectiveness against zero-day attacks

It is to be noted that the developed IDS uses only self-traffic for training and not non-self-traffic. Thus, in the view of the IDS, all the attacks are unknown (like zero day attacks). This makes the detection rates and false positive rates mentioned above equally applicable to zero days' attacks as it is to any other attack. Even so, the dataset has only 568 (out of the total of 22,544 records) of the non-self-instances in the test dataset carried over from the training dataset. That is, only 2.43% of the entries in the test dataset is comprised of previously known attacks.

5.0 SYSTEM ANALYSIS AND DEVELOPMENT

To overcome the shortcomings of existing intrusion detection systems, a multi-layer model is provided (Figure 5.1) which consists of three processing layers: 1) Packet Analysis; 2) Intrusion Detection; and 3) Security Information and Event Management (SIEM).

5.1 Packet Analysis

Being responsible for all the preprocessing tasks required for the intrusion detection, Packet analysis layer contains two important modules, namely flow analyzer and traffic classification.

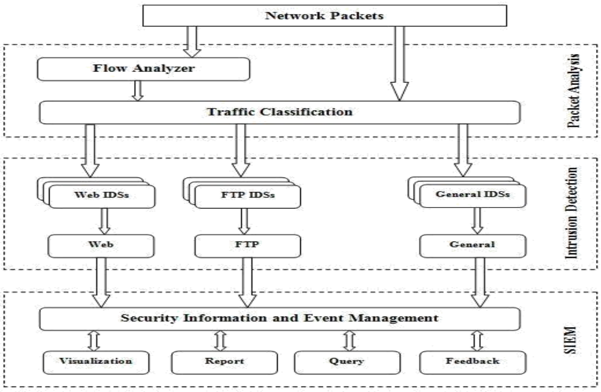

Figure 5.1: The proposed framework

5.1.2 Flow Analyzer

In order to keep up with the high speed of gigabit links, current network monitoring and management systems use network flow data (e.g., netflow, sflow, ipfix) as their information sources. Network flows are a group of net-work packets belonging to the same connection. Apparently, these packets have a lot of information in common (e.g., source IP, source port, destination IP, destination port, protocol, application), which can be stored only once using the flow concept. Furthermore, in order to deal with a huge amount of payload information, usually only the first few bytes of each flow (e.g., 512 bytes), which is more informative for the analysis, will be stored by the monitoring devices.

5.1.3 Traffic Classification

As network applications are getting more diverse and complex, the idea of using special-purpose intrusion detection systems in the application layer be-comes more popular. Focusing on a small subset of network applications will have the advantage of designing more specific signatures, which results in a better detection rate and a lower false positive rate. In addition, as illustrated in Figure 5.1, application-based IDSs can be applied in parallel which is of high importance in dealing with large networks with millions of packets per second.

5.2 Intrusion Detection

As indicated in Figure 5.1, the proposed intrusion detection module consists of several application-based intrusion detection systems components. Sharing a similar architecture and detection mechanism, each component is specifically designed for a special type of application such as Web, FTP, Mail, etc. There is also a component designed for applications with no specialized intrusion detection system and applications that are not detected with the traffic classifier.

5.2.1 Security Information and Event Management

Although a lot of efforts have been done to decrease the number of false alarms generated by intrusion detection systems, we believe that having an IDS with no false alarm is almost impossible due to the dynamic nature of computer networks. However, we can minimize these false alarms by gathering and processing different types of information from various sources such as intrusion detection systems, anti-viruses, operating systems logs, application level logs, among others.

5.3 Traffic Classification Module

Accurate classification of network traffic has received a lot of attention due to its important role in many subjects such as network planning, QoS provisioning, class of service mapping, to name a few. Traditionally, traffic classification relied to a large extent on the association of a particular port with a specific protocol. Such a port number based traffic classification approach has been proved to be ineffective due to: 1) the constant emergence of new peer-to-peer networking applications that IANA does not define the corresponding port numbers; 2) the dynamic port number assignment for some applications (e.g. FTP); and 3) the encapsulation of different services into a same application (e.g., chat or steaming can be encapsulated into the same HTTP protocol). To overcome this issue, there have been recently significant contributions towards traffic classification. The most currently successful approach is to inspect the content of payloads and look for the deterministic character strings for modeling the applications. For most applications, their initial protocol handshake steps are usually different and thus can be used for classification.

5.3.1 Weighted Unigram Model

N-grams are a language-independent means of gauging topical similarity in text documents. Traditionally, the n-grams technique refers to passing a sliding window of *n* characters over a text document and counting the occurrence of each n-gram. This method is widely employed in many language analysis tasks as well as network security. Applying the same idea on network packets, one can consider unigram (1-gram) of a network packet as a sequence of ASCII characters ranging from 0 to 255. This way similar packets can be identified using the frequencies of distinct ASCII.

5.3.2 Problem Formulation

In this section, we formally describe how the network application discovery problem can be performed through the combination of genetic algorithms and decision trees. Essentially, we formulate the network application discovery problem as a classification problem, i.e., given the values for a specific set of features extracted from the network flows, we identify the possible application that has generated this payload using a statistical machine learning technique (decision tree).

5.4 Intrusion Detection Module

Traditionally, intrusion detection techniques are classified into two categories: misuse (signature-based) detection and anomaly detection. Misuse detection is based on the assumption that a large number of cyber-attacks leave a set of signatures in the stream of network packets or in audit trails, and thus attacks are detectable if these signatures can be identified by analyzing the audit trails or network traffic behavior. However, misuse detection is strictly limited to the known attacks and detecting new attacks is one of the biggest challenges faced by misuse detection.

5.4.1 Anomaly-based Detector

As the first step to have an effective anomaly detector, we should extract robust network features that have the potential to discriminate anomalous behavior from normal network activities. Since most current network intrusion detection systems use network flow data (e.g. netflow, sflow, ipfix) as their information sources, we focus on features generated based on these flows.

5.4.2 Signature-based Detector

As our first signature-based detector we chose Snort because of its popularity and availability to researchers. However, our proposed hybrid detection scheme is completely independent from Snort, and any other signature-based detector can be used instead. As mentioned earlier, our anomaly-based detector works on flows. However, Snort is designed to work on packets.

6.0 CONCLUSION AND RECOMMENDATION

A system is proposed which has an adaptive hybrid data loss detection and prevention intrusion detection system to overcome the main shortcomings of the existing IDSs.

With regard to the hybrid intrusion detection, we have identified two main is-sues that highly affects the performance of the system. First, anomaly-based methods cannot achieve an outstanding performance without a comprehensive labeled and up-to-date training set with all different attack types, which is very costly and time-consuming to create if not impossible. Second, efficient and effective fusion of several detection technologies becomes a big challenge for building an operational hybrid intrusion detection system. To solve the first issue, we have proposed applying the idea of adaptive learning. To meet this goal, we have defined learning time intervals, e.g. 1 day, at the end of which the anomaly-based detector will be trained by the two most recent training sets. These training sets are the flows that are labeled by the hybrid detector in the previous intervals.

7. REFERENCES

[1] Mahajan, P., Gaba, G., & Chauhan, N. S. (2016). Big Data Security. IITM Journal of Management and IT, 7(1), 89-94.

[2] Harish Kumar, M. & Menakadevi, T. (2017), A Review on Big Data Analytics in the field of Agriculture, International Journal of Latest Transactions in Engineering and Science, vol. 1, issue 4, pp. 0001-0010. Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004), Design Science in Information Systems Research, MIS Quarterly, vol. 28, no. 1, pp. 75-105.

[3] Sagiroglu, S., and Sinanc, D. 2013. “Big data: A review,” in Proceeding of the International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47 (doi: 10.1109/CTS.2013.6567202).

[4] Fang, Z., & Li, P. (2014). The mechanism of “big data” impact on consumer behavior. American Journal of Industrial and Business Management, 4(1), 45-50.

[5] Tiwari, M., Kumar, R., Bharti, A. & Kishan, J. (2017). Intrusion Detection System. International Journal of Technical Research and Applications, 5(2):2320-8163. Retrieved from https://www.ijtra.com/view/intrusion-detection-system.pdf?paper=intrusion-detection-system.pdf

[6] Jose, S., Malathi, D., Reddy, B., Jayaseeli, D. (2018). A Survey on Anomaly Based Host Intrusion Detection System. Journal of Physics: Conf. Series 1000 (2018) 012049 doi :10.1088/1742-6596/1000/1/012049

[7] Maybury, M.T., Chase, P., Cheikes, B.A., Brackney, D., Matzner, S., Wood, B.J., Longstaff, T., Hetherington, T., Marin, J., Spitzner, L., Copeland, J.S., Lewandowski, S.M., & Haile, J. (2005). Analysis and Detection of Malicious Insiders.

[8] Granjal, J., & Pedroso, A. (2018). An intrusion detection and prevention framework for internet-integrated CoAP WSN. Security and Communication Networks, vol. 2018, Article ID 1753897, 14 pages, 2018.

[9] Desnitsky, Vasily & Kotenko, Igor & Nogin, S.. (2015). Detection of anomalies in data for monitoring of security components in the Internet of Things. 189-192. 10.1109/SCM.2015.7190452.

[10] Funke Olanrewaju, R., Ul Islam Khan, B., Rahman Najeeb, A., Afiza Ku Zahir, K., & Hussain, S. (2018). Snort-Based Smart and Swift Intrusion Detection System. Indian Journal Of Science And Technology, 11(4). doi:10.17485/ijst/2018/v11i4/120917

[11] Kale, A. V., Bajpayee, V. & Dubey, S. P. (2015), Analysis of Data Leakage Prevention Solutions, International Journal For Engineering Applications And Technology (IJFEAT), vol. 1, issue, 12, pp. 54- 57.

[12] Jain, M & Lenka, S. K. (2016), A Review on Data Leakage Prevention using Image Steganography, International Journal of Computer Science Engineering (IJCSE), vol. 5, no. 02, pp. 56-59.

[13] Tahboub, Radwan & Saleh, Yousef. (2015). Precaution Model for Data Leakage Prevention/Loss (DLP) Systems.

[14] S. W. Ahmad and G. R. Bamnote, "Data Leakage Detection and Data Prevention using Algorithm," International Journal of Computer Science and Application, vol. 6, pp. 394-399, 2013.

[15] Peneti, S. & Rani, B. P. (2015a), Data Leakage Detection and Prevention Methods: Survey. Discovery, vol. 43, no. 198, pp. 95-100.

[16] Shirudkar, K. & Motwani, D. (2015), Big-Data Security. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, issue 3, pp. 1100-1109.

[17] Alneyadi, S., Sithirasenan, E. & Muthukkumarasamy, V. (2016), A survey on data leakage prevention systems, Journal of Network and Computer Applications, vol. 62, issue C, pp. 137-152.

[18] Alneyadi, S., Sithirasenan, E. and Muthukkumarasamy, V. (2015), Detecting Data Semantic: A Data Leakage Prevention Approach, In the Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, August 20 - 22, IEEE Computer Society Washington DC, USA, vol. 1, pp. 910-917.

[19] Margathavalli, P., Manjula, R., Pramila, V., Priya, R. & Abirami, P. (2016), Preserving Sensitive Data by Data Leakage Prevention Using Attribute Based Encryption Algorithm, International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), vol. 21, issue 3, pp. 705-711.