RESEARCH ARTICLE OPEN ACCESS

Semi-supervised Intelligence for the The determination of DDOS attacks

KANDHIBANDA PAVAN,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
PONUGOTI AKSHARA,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
DUGUNOOR HEMANTH KUMAR,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
KARLA SAIKIRAN,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)

T.GLORY,Asst.Prof.CSIT,Sri Indu College Of Engineering & Technology(A)

Abstract Analyzing cyber incident data sets is animportant method for deepening our understanding of the evolution of the threat situation. This is a relatively new research topic, and many studies remain to be done. In this paper, we report a statistical analysis of a breach incident data set corresponding to 12 years (2005–2017) of cyber hacking activities that include malware attacks. We show that, in contrast to the findings reported in the literature, both hacking breach incident inter-arrival times and breach sizes should be modeled by stochastic processes, rather than by distributions because they exhibit autocorrelations. Then, we propose particular stochastic process models to, respectively, fit the inter-arrival times and the breach sizes. We also show that these models can predict the inter-arrival times and the breach sizes. In order to get deeper insights into the evolution of hacking breach incidents, we conduct both qualitative and quantitative trend analyses on the data set. We draw a set of cybersecurity insights, including that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage.

I. INTRODUCTION

Despite the important evolution of the information security technologies in recent years, the DDoS attack remains a major threat of Internet. The attack aims mainly to deprive legitimate users from Internet resources. The impact of the attack relies on the speed and the amount of the network traffic sent to the victim. Generally, there exist two categorie to flood directly the victim host with a large number of network packets. Whereas, in the Reflectionbased DDoS attack the attacker uses the zombie hosts to take control over a set of compromised hosts called Reflectors. The latter are used to forward a massive amount of attack traffic to the victim host. Recently, destructive DDoS attacks have brought down more than 70 vital services of Internet including Github,Twitter, Amazon, Paypal, etc [5, 6]. Attackers have taken advantages of Cloud Computing and Internet of Things technologies to generate a huge amount of attack traffic; more than 665 Gb/s [5, 6]. Analyzing this amount of network traffic at once is inefficient, computationally costly and often leads the intrusion detection systems to fall. Data mining techniques have been used to develop sophisticated intrusion detection systems for the last two decades. Artificial Intelligence, Machine Learning (ML), Pattern Recognition, Statistics, InformationTheory are the most used data mining techniques for intrusion detection [7]. Application process of data mining techniques in general and ML techniques more specifically requires five typical steps selection, preprocessing, transformation, mining, and interpretation [8]. Despite that preprocessing and transformation steps may be trivial for intrusion detection applications, selection, mining and interpretation steps are crucial for selecting relevant data, filtering noisy data and detecting intrusions [7]. These three crucial steps are the most challenging of the existing data mining based intrusion detection approaches. The existing Machine Learning based DDoS detection approaches can be divided

Supervised ML approaches that use generated labeled network traffic datasets to build the detection model. Two major issues are facing the supervised approaches. First, the generation of labeled network traffic datasets is costly in terms of computation and time. Without a continuous update of their detection models, the supervised machine learning approaches are unable to predict the new legitimate and attack behaviors. Second, the the presence of large amount of irrelevant normal data in the incoming network traffic is noisy and reduces the performances of supervised ML classifiers. Unlike the first category, in the unsupervised approaches no labeled dataset is needed to built the detection model. The DDoS and the normal traffics are distinguished based on the analysis of their underlying distribution characteristics. However, the main drawback of the unsupervised approaches is the high false positive rates. In the high dimensional network traffic data the distance between points becomes meaningless and tends to homogenize. This problem, known as 'the curse of dimensionality', prevents unsupervised approaches to accurately detect attacks [9]. The semi-supervised ML approaches are taking advantages of both supervised and unsupervised approaches by the ability to work on labeled and unlabeled datasets. Also, the combination of supervised and unsupervised approaches allows to increase accuracy and decreases the false positive rates. However, semi-supervised approaches are also challenged by the drawbacks of both approaches. Hence, the semi-supervised approaches require a sophisticated implementation of its components in order to overcome the drawbacks of supervised and unsupervised approaches. In this paper we present an online sequential semisupervised ML approach for DDoS detection. A time based sliding window algorithm is used to estimate the entropy of the network header features of the incoming network traffic. When the entropy exceeds its normal range, the unsupervised co-clustering algorithm splits the incoming network traffic into three clusters. Then, an information gain ratio [10] is computed based on the average entropy of the network header features between the network traffic subset of the current time window and each one of the obtained clusters. The network traffic data clusters that produce high information gain ratio are considered as anomalous and they are selected for preprocessing and classification using an ensemble classifiers based on the Extra-Trees algorithm [11]. Our approach constitutes of two main parts unsupervised and supervised. The unsupervised part includes entropy estimation, co-clustering and information gain ratio. The supervised part is the Extra-Trees ensemble classifiers. The unsupervised part of our approach allows to reduce the irrelevant and noisy normal traffic data, hence reducing false positive rates and increasing accuracy of the supervised part. Whereas, the supervised part is used to reduce the false positive rates of the unsupervised part and to accurately classify the DDoS traffic. To better evaluate the performance of the proposed approach three public network traffic datasets are used in the experiment, namely the NSL-KDD [12], the UNB ISCX IDS 2012 dataset [13] and the UNSW-NB15 [14, 15]. The experimental results are satisfactory when compared with the state-of-the-art DDoS detection methods. The main contributions of this paper can be summarized as follows: • Presenting an unsupervised and time sliding window algorithm for detecting anomalous traffic data based on co-clustering, entropy estimation and information gain ratio. This algorithm allows to reduce drastically the amount of network traffic to preprocess and to classify, resulting in a significant improvement of the performance of the proposed approach. • Adopting a supervised ensemble ML classifiers based on the Extra-Trees algorithm to accurately classify the anomalous traffic and to reduce the false positive rates. • Combining both previous algorithms in a sophisticated semi-supervised approach for DDoS detection. This allows to achieve good DDoS detection performance compared to the state-of-the-art DDoS detection methods. • The unsupervised part of our approach allows to reduce the irrelevant and noisy normal traffic data, hence reducing false positive rates and increasing accuracy of the supervised part. Whereas, the supervised part allows to reduce the false positive rates of the unsupervised part and to accurately classify the DDoS traffic.

II. LITERATURE REVIEW

Several approaches have been proposed for detecting DDoS attack. Information theory and machine learning are the most common techniques used in the literature. This section summarizes some of the recent works in DDoS detection.

Akilandeswari V. et al. [16] have used a Probabilistic Neural Network to discriminate flash crowd events from DDoS attacks. The method achieves high DDoS detection accuracy with lowe false positives rates.

Similarly, Ali S.B. et al. [17] have proposed an innovative ensemble of Sugeno type adaptive neuro-fuzzy classifiers for DDoS detection using an effective boosting technique named Marliboost. The proposed technique was tested on the NSL-KDD dataset and have achieved good performance.

Mohiuddin A. and Abdun Naser M. [18] have proposed an unsupervised approach for DDoS detection based on the co-clustering algorithm. The authors have extended the co-clustering algorithm to handle categorical attributes. The approach was tested on the KDD cup 99 dataset and achieved good performance.

Alan S. et al. [19] have proposed a DDoS Detection Mechanism based on ANN (DDMA). The authors used three different topologies of the MLP for detecting three types of DDoS attacks based on the background protocol used to perform each attack namely TCP, UDP and ICMP. The mechanism detects accurately known and unknown, zero day, DDoS attacks.

Similarly, Boro D. et al. [20] have presented a defense system referred to as DyProSD that combines both the merits of feature-based and statistical approach to handle DDoS flooding attack. The statistical module marks the suspicious traffic and forwards to an ensemble of classifiers for ascertaining the traffic as malicious or normal.

Recently, Van Loi C. [21] proposed a novel oneclass learning approach for network anomaly detection based on combining auto-encoders and density estimation. Authors have tested their method on the NSL-KDD dataset, and obtained satisfactory results.

Mohamed I. et al. [22] have proposed a supervised DoS detection method based on a feed-forward neural network.

This method consists of three major steps:

(1) Collection of the incoming network traffic,

(2) selection of relevant features for DoS detection using an unsupervised Correlation-based Feature Selection (CFS) method,

(3) classification of the incoming network traffic into DoS traffic or normal traffic. The approach achieves good performances on the UNSW-NB15 and NSLKDD datasets.

Mustapha B. et al. [23] have presented a two-stage classifier based on RepTree algorithm and protocols subset for network intrusion detection system. The first phase of their approach consists of dividing the incoming network traffic into three type of protocols TCP, UDP or Other. Then classifying it into normal or anomaly traffic. In the second stage a multi-class algorithm classify the anomaly detected in the first phase to identify the attacks class in order to choose the appropriate intervention. Two public datasets are used for experiments in this paper namely the UNSW-NB15 and the NSL-KDD.

The performances of network intrusion detection approaches, in general, rely on the distribution characteristics of the underlaying network traffic data used for assessment. The DDoS detection approaches in the literature are under two main categories unsupervised approaches and supervised approaches. Depending on the benchmark datasets used, unsupervised approaches often suffer from high false positive rate and supervised approach cannot handle large amount of network traffic data and their performances are often limited by noisy and irrelevant network data. Therefore, the need of combining both, supervised and unsupervised approaches arises to overcome DDoS detection issues.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM:

The present study is motivated by several questions that have not been investigated until now, such as: Are data breaches caused by cyber-attacks increasing, decreasing, or stabilizing? A principled answer to this question will give us a clear insight into the overall situation of cyber threats. This question was not answered by previous studies. Specifically, the dataset analyzed in [7] only covered the time span from 2000 to 2008 and does not necessarily contain the breach incidents that are caused by cyber-attacks; the dataset analyzed in [9] is more recent, but contains two kinds of incidents: negligent breaches (i.e., incidents caused by lost, discarded, stolen devices and other reasons) and malicious breaching. Since negligent breaches represent more human errors than cyber-attacks, we do not consider them in the present study. Because the malicious breaches studied in [9] contain four sub-categories: hacking (including malware), insider, payment card fraud, and unknown, this study will focus on the hacking sub-category (called hacking breach dataset thereafter), while noting that the other three sub-categories are interesting on their own and should be analyzed separately. Recently, researchers started modeling data breach incidents. Maillart and Sornette studied the statistical properties of the personal identity losses in the United States between year 2000 and 2008. They found that the number of breach incidents dramatically increases from 2000 to July 2006 but remains stable thereafter. Edwards et al. analyzed a dataset containing 2,253 breach incidents that span over a decade (2005 to 2015). They found that neither the size nor the frequency of data breaches has increased over the years. Wheatley et al., analyzed a dataset that is combined from corresponds to organizational breach incidents between year 2000 and 2015. They found that the frequency of large breach incidents (i.e., the ones that breach more than 50,000 records) occurring to US firms is independent of time, but the frequency of large breach incidents occurring to non-US firms exhibits an increasing trend.

DISADVANTAGES:

- Analyzing cyber incident data sets is an important method for deepening our understanding of the evolution of the threat situation
- Modeling data breach incidents. Maillart and Sornette the statistical properties of the personal identity losses in the UnitedStates
- The monetary price incurred by data breaches is also substantial. Reports that in the global average cost for each lost or stolen record containing sensitive or confidential information

PROPSOED SYSTEM

In this paper, we make the following three contributions. First, we show that both the hacking breach incident interarrival times (reflecting incident frequency) and breach sizes should be modeled by stochastic processes, rather than by distributions. We find that a particular point process can adequately describe the evolution of the hacking breach incidents inter-arrival times and that a particular ARMA-GARCH model can adequately describe the evolution of the hacking breach sizes, where ARMA is acronym for “AutoRegressive and Moving Average” and GARCH is acronym for “Generalized AutoRegressive Conditional Heteroskedasticity.”We show that these stochastic process models can predict the inter-arrival times and the breach sizes. To the best of our knowledge, this is the first paper showing that stochastic processes, rather thandistributions, should be used to model these cyber threat factors. Second, we discover a positive dependence between the incidents inter- arrival times and the breach sizes, and show that this dependence can be adequately described by a particular copula. We also show that when predicting inter-arrival times and breach sizes, it is necessary to consider the dependence; otherwise, the prediction results are not accurate. To the best of our knowledge, this is the first work showing the existence of this dependence and the consequence of ignoring it. Third, we conduct both qualitative and quantitative trend analyses of the cyber hacking breach incidents. We find that the situation is indeed getting worse in terms of the incidents inter-arrival time because hacking breach incidents become more and more frequent, but the situation is stabilizing in terms of the incident breach size, indicating that the damage of individual hacking breach incidents will not get much worse. We hope the present study will inspire more investigations, which can offer deep insights into alternate risk mitigation approaches. Such insights are useful to insurance companies, government agencies, and regulators because they need to deeply understand the nature of data breach risks.

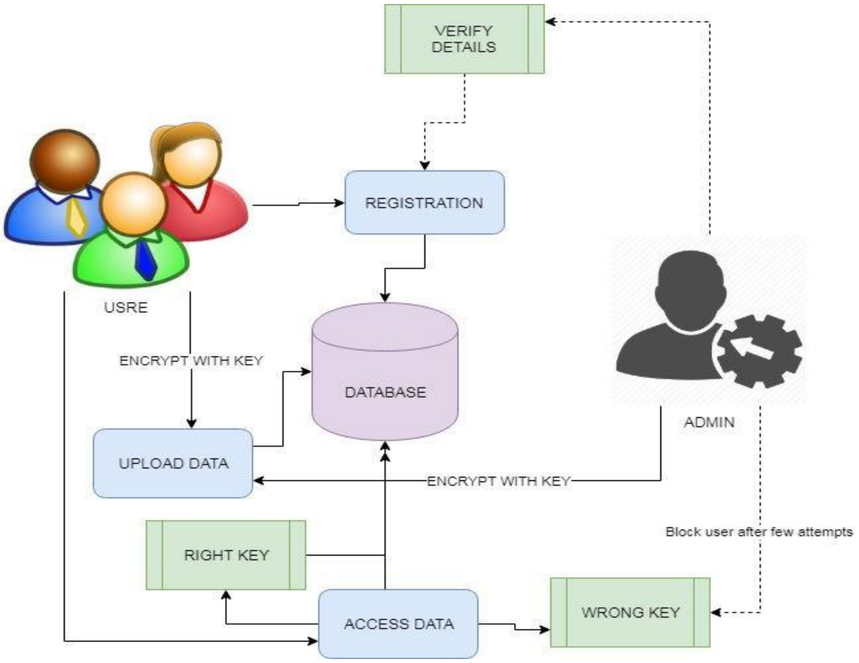
ADVANTAGES:

- Cyber hacking activities that includemalware attacks. We show that, in

SYSTEM ARCHITECTURE:

contrast to the findings reported in the literature

- Incident inter-arrival times and breach sizes should be modeled by stochastic processes
- we propose particular stochastic process models to, respectively, fit the inter-arrival times and the breach sizes



IV. IMPLEMENTATION

MODULES:

1. UPLOAD DATA

The data resource to database can be uploaded by both administrator and authorized user. The data can be uploaded with key in order to maintain the secrecy of the data that is not released without knowledge of user. The users are authorized based on their details that are shared to admin and admin can authorize each user. Only Authorized users are allowed to access the system and upload or request for files.

2. ACCESS DETAILS

The access of data from the database can be given by administrators. Uploaded data are managed by admin and admin is the only person to provide the rights to process the accessing details and approve or unapproved users based on their details.

3. USER PERMISSIONS

The data from any resources are allowed to access the data with only permission from administrator. Prior to access data, users are allowed by admin to share their data and verify the details which are provided by user. If user is access the data with wrong attempts then, users are blocked accordingly. If user is requested to unblock them, based on the requests and previous activities admin is unblock users.

4. DATA ANALYSIS

Data analyses are done with the help of graph. The collected data areapplied to graph in order to get the best analy sis and prediction of dataset and given data policies. The dataset can be analyzed through this pictorial representation in order to better understand of the data details.

V. CONCLUSION

We analyzed a hacking breach dataset from the points of view of the incidents inter-arrival time and the breach size, and showed that they both should be modeled by stochastic processes rather than distributions. The statistical models developed in this paper show satisfactory fitting and prediction accuracies. In particular, we propose using a copula-based approach to predict the joint probability that an incident with a certain magnitude of breach size will occur during a future period of time. Statistical tests show that the methodologies proposed in this paper are better than those which are presented in the literature, because the latter ignored both the temporal correlations and the dependence between the incidents inter-arrival times and the breach sizes. We conducted qualitative and quantitative analyses to draw further insights. We drew a set of cybersecurity insights, including that the threat of cyber hacking breach incidents is indeed getting worse in terms of their frequency, but not the magnitude of their damage. The methodology presented in this paper can be adopted or adapted to analyze datasets of a similar nature

REFERENCES

1. Bhuyan MH, Bhattacharyya DK, Kalita JK (2015) An empirical evaluation of information metrics for low-rate and high-rate ddos attack detection. Pattern Recogn Lett 51:1–7

2. Lin S-C, Tseng S-S (2004) Constructing detection knowledge for ddos intrusion tolerance. Exp Syst Appl 27(3):379–390

3. Chang RKC (2002) Defending against flooding-based distributed denial-of-service attacks: a tutorial. IEEE Commun Mag 40(10):42–51

4. Yu S (2014) Distributed denial of service attack and defense. Springer, Berlin

5. Wikipedia (2016) 2016 dyn cyberattack. https://en.wikipedia.org/ wiki/2016 Dyn cyberattack. (Online; accessed 10 Apr 2017)

6. theguardian (2016) Ddos attack that disrupted internet was largest of its kind in history, experts say. https://www.theguardian.com/ technology/2016/oct/26/ddos-attack-dyn-mirai-botnet. (Online; accessed 10 Apr 2017)

7. Kalegele K, Sasai K, Takahashi H, Kitagata G, Kinoshi
ta T (2015) Four decades of data mining in network and systems management. IEEE Trans Knowl Data Eng 27(10):2700–2716

8. Han J, Pei J, Kamber M (2006) What is data mining. Data mining: concepts and techniques. Morgan Kaufinann

9. Berkhin P (2006) A survey of clustering data mining techniques. In: Grouping multidimensional data. Springer, pp 25–71

10. Mori T (2002) Information gain ratio as term weight: the case of summarization of ir results. In: Proceedings of the 19th international conference on computational linguistics, vol 1. Association for Computational Linguistics, pp1–7

11. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42

12. Tavallaee M, Bagheri E, Lu W, Ghorbani A-A (2009) A detailed analysis of the kdd cup 99 data set. In: Proceedings of the second IEEE symposium on computational intelligence for security and defence applications 2009

13. Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA (2012) Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput Secur 31:357–374

14. Moustafa N, Slay J (2015) Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: Military communications and information systems conference (MilCIS), 2015. IEEE, pp 1–6

15. Moustafa N, Slay J (2016) The evaluation of network anomaly detection systems: statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set. Inf Secur J: Glob Perspect 25:18–31