Journal of Engineering Sciences
Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

NUMEROUS TOPOGRAPHIES BASED QUASI SUPERVISED CLUSTERING DDOS DETECTION METHOD

N. SWAPNA[1], Dr. BHALUDRA RAVEENDRANADH SINGH[2], S. SHRAVANI[3], V. SRILEKHA[4], M. PRAVALIKA[5], M. JYOTHSNA[6]
UG SCHOLAR[1,3,4,5&6], PROFESSOR& HEAD[2]
DEPARTMENT OF CSE, AVN INSTITUTE OF ENGINEERING AND TECHNOLOGY, KOHEDA ROAD, IBRAHIMPATNAM(M), R.R.DIST-501510, HYDERABAD

ABSTRACT

DDoS attack stream from different agent host converged at victim host will become very large, which will lead to system halt or network congestion. Therefore, it is necessary to propose an effective method to detect the DDoS attack behavior from the massive data stream. In order to solve the problem that large numbers of labeled data are not provided in supervised learning method, and the relatively low detection accuracy and convergence speed of unsupervised -means algorithm, this paper presents a semisupervised clustering detection method using multiple features. In this detection method, we firstly select three features according to the characteristics of DDoS attacks to form detection feature vector. Then, Multiple-Features-Based Constrained--Means (MF-CKM) algorithm is proposed based on semi supervised clustering. Finally, using MIT Laboratory Scenario (DDoS) 1.0 data set, we verify that the proposed method can improve the convergence speed and accuracy of the algorithm under the condition of using a small amount of labeled data sets.

INTRODUCTION: A denial-of-service attack (DoS attack) is a cyber-attack where the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet. Denial of service is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled [1]. A distributed denial-of-service (DDoS) is a cyber-attack where the perpetrator uses more than one unique IP address, often thousands of them. The scale of DDoS attacks has continued to rise over recent years, by 2016 exceeding a terabit per second [2]. DDoS attacks are distributed that an attacker initiated this attack by manipulating distributed Internet agent host of different locations at the same time. When the attack stream from different agent host converged, the stream at victim host will become very large and will soon become system halted or network congestion [3].DDoS attacks can be performed by a large group of cooperating people, a small group of people, or a single person that controls one or more sufficiently powerful botnets. All types of motivations can lead to an organized attack: political and social issues are among the top motivations, but any public or private institution or company can be a victim because small groups or individual criminals usually have specific targets which are chosen based on revenge, competition, or simply the desire to cause damage [4]. Therefore, it is necessary to propose an effective method to detect the DDoS attack in the massive data stream.In this paper, we propose a novel method for DDoS attack detection, which is based on multiple-features-based semi-supervised clustering algorithm, and the provided method uses only small amount of labeled data and relatively large amount of unlabeled data to detect DDoS attack behavior. Compared with previous detection solutions based on supervised learning and unsupervised learning methods, our proposed algorithm has the following advantages:

- Compared with supervised learning detection algorithms, our method requires fewer labeled data sets to training detection models.
- Compared with unsupervised learning detection algorithms, our method has higher detection accuracy and can improve the convergence speed of the model

(reduce the time complexity of the algorithm).

RELATED WORK: There are two classes of DDoS detection techniques: misuse detection and anomaly detection [5]. The misuse detection techniques try to detect attack by comparing the current activity of destination network to a database of known attack signatures. These techniques cannot detect unknown attacks, while the anomaly detection techniques try to detect attack by comparing the current activity of destination network to an established normal activity represented as a profile.In recent years, machine learning algorithms are often used in anomaly detection. Machine learning methods mainly include supervised learning and unsupervised learning [6]. Classification and regression problems belong to the category of supervised learning. Commonly used classification algorithms include decision tree classification [7], naive Bayes classification algorithm [8, 9], Support Vector Machine (SVM) classifier [10], Neural Network method [11], and -nearest neighbor (kNN) [12]. The problem of association and clustering belongs to the category of unsupervised learning, and -means algorithm is the most commonly used clustering algorithm [13].Liao and Vemuri [14] used -nearest neighbor classifier (KNNC) to categorize process into normal or intrusive class. The KNNC calculates the similarity between the new process and each training process instance and basically assumes that the processes belonging to the same class will cluster together in the vector space. It is excellent in attack detection, but the detector is computationally expensive for real-time implementation when the number of processes simultaneously increases. Support Vector Machine (SVM) is a technique based on machine learning, where data is classified by determining a group of support vectors and characteristics to be quantified are described. As proposed by [15, 16], the Hybrid Intrusion Detection System (HIDS), based on machine learning and specifically the SVM technique, improves the detection rate. More recent study as [11] presents a better classification using an Artificial Neural Network (ANN) to flag detection engine known and unknown attacks from genuine traffic. Ramos et al. [12] use -NN classifier method and cosine formula based algorithm to detect the DDoS attack, but this method needs some time to train the original packets. Xiao et al. [17] present a detection approach based on CKNN (-nearest neighbors traffic classification with correlation analysis) to detect DDoS attacks. The approach exploits correlation information of training data to improve the classification accuracy and reduce the overhead caused by the density of training data. Öke et al. [18] used multiple Bayesian classifiers to detect DDoS attacks. However, naive Bayes are based on a very strong independence assumption, which is not always satisfied. Amor et al. [19] compared the performance of naive Bayes with C4.5 decision tree and find the good performance of Bayes with respect to existing best results performed on KDD'99.Clustering algorithm mainly includes two categories: hierarchical and partitioning [20]. Partitioning method is inappropriate for our case because the number of clusters should be predetermined in partitioning. Therefore, the paper adopts a hierarchical method. This method is often used to classify plants and animals and is expected to be adequate for classifying the phases of the DDoS attack by the use of their features. In clustering, the learning algorithm finds similarities among instances to build the clusters (i.e., group of instances). Instances that belong to the same cluster are assumed to have similar characteristics or properties and then are assembled into the same class. -means algorithm belongs to partitioning one, which has been successfully used to detect anomalies [13] and DDoS [21], using clustering methodologies to formulate the normal patterns, since one of the advantages of clustering methods over statistical methods is that they do not rely on any prior known data distribution. But machine learning based techniques require a lengthy learning period and hence currently these methods cannot operate in real-time [22].Many advantages and disadvantages related to the above machine learning algorithms with anomaly detection have been reported by many researchers [23, 24]. Supervised learning has the advantage to achieve better accuracy to classify similar examples. But, one shortcoming of supervised learning is the need for large scale labeled instances. This raises ambiguity about the performance of supervised learning, since it requires a sufficient amount of labeled data to train the classifier [25]. Unsupervised learning techniques deal with the learning tasks with unlabeled or untagged data, and clustering is the most popular unsupervised learning technique [26]. They have the advantage

of detecting new examples better than supervised learning techniques and are considered to be more robust in IDSs. However, the disadvantage of unsupervised learning is the manual assignment of cluster numbers, which results in relatively low accuracy in predictions. In case of unsupervised learning, large amount of uncertainty is associated with modeling the data set. In addition, for the typical unsupervised learning algorithm -means, the selection of value and the initial clustering centers have great influence on the clustering accuracy and the convergence speed of the algorithm.In order to integrate the advantages of supervised and unsupervised learning methods, and considering the actual application scenes which have small amount of labeled data and relatively large amount of unlabeled data, this paper provides a semisupervised clustering method to detect DDoS attacks.

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 United States
- 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

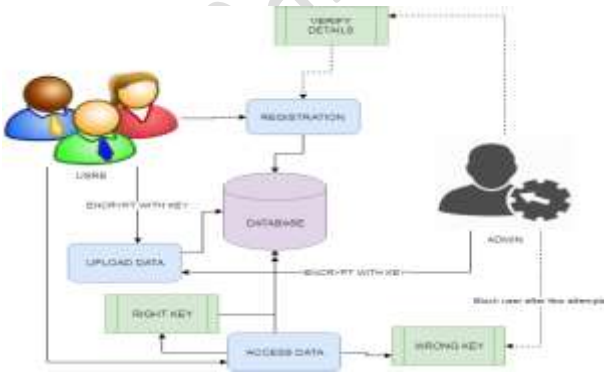
In this paper, we make the following three contributions. First, we show that both the hacking breach incident interarrival times (reflecting incident frequency) and breach sizes should be modeled by stochastic processes, rather than by distributions. We find that a particular point process can adequately 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 than distributions, 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 include malware attacks. We show that, in 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

SYSTEM ARCHITECTURE:



MODULES:

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.

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.

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.

DATA ANALYSIS

Data analyses are done with the help of graph. The collected data are applied to graph in order to get the best analysis 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.

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 empiricalevaluation of information metrics for low-rate and high-rate ddosattack detection. Pattern Recogn Lett 51:1–7

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

3. Chang RKC (2002) Defending against flooding-based distributeddenial-of-service attacks: a tutorial. IEEE Commun Mag40(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 largestof 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, Kinoshita 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 caseof summarization of ir results. In: Proceedings of the 19thinternational conference on computational linguistics, vol 1.Association for Computational Linguistics, pp 1–7

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

12.Pravin Kshirsagar, Dr. Sudhir Akojwar, "A hybridized neural network and optimization algorithms for prediction & classification of neurological disorders", Int. Journal of Biomedical Engineering and Technology, Vol. 28, No. 4, 2018.

13. Pravin Kshirsagar, Nagaraj Balakrishnan & Arpit Deepak Yadav (2020) Modelling of optimised neural network for classification and prediction of benchmark datasets, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, DOI:10.1080/21681163.2019.1711457,2020

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

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