Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 13, Issue 1, January 2022: 95-103

Semi Supervised Machine Learning Forddos Detection

Mr. M Shiva Rama Krishna, PG Student, Department of Computer Science and Engineering, Holy Mary Institute of Technology & Science, Hyderabad, India.
Email id: shiva57.m@gmail.com
Mrs.A. Eenaja , Assistant Professor, Department of Computer Science and Engineering, Holy Mary Institute of Technology & Science, Hyderabad, India.

ABSTRACT

Analyzing cyber incident data sets is an essential way to have a better understanding of how the threat environment is evolving. This is a very new study area, thus there are still a lot more studies to be done. We present a statistical study of a breach incidence data set spanning 12 years (2005–2017) of cyber hacking operations involving malware assaults in this research. In contrast to previous research, we show that due of autocorrelations, Then, we offer specific stochastic process models that suit the inter-arrival timings and breach sizes, respectively.forecast breach sizes and inter-arrival periods. We do both qualitative and quantitative trend studies on the data set in order to have a better understanding of the progression of cyber breach occurrences. We derive a number of cybersecurity conclusions,

Keywords :semi-Supervised, Clustering, Random forest

INTRODUCTION

Machine Learning is the study of modelling computers to learn and solve problems accurately in the same way that humans do. Machine learning is primarily concerned with presenting algorithms that can be trained to accomplish a task. Python is a fantastic programming language for machine learning. Python has practically all machine learning algorithms built-in, making it simple to access such algorithms for a specific purpose by installing the relevant libraries. This makes the code smaller, easier to understand, and improves the quality of the outputs. The network traffic data is first read in this method. The data's average entropy is calculated, and three groups are generated using the clustering technique. The information gain ratio is derived using the average entropy of data. By joining the clusters with the greatest gain ratio value, an anomalous cluster is generated. To effectively identify the data and detect the DDoS assault, the produced anomalous cluster is subjected to a random forest method. The NSL-KDD network traffic dataset is utilised to evaluate the performance of the suggested methodology. The NSL-KDD dataset comprises many attack data kinds. It contains 42 distinct characteristics that may be categorised into several groups. There are 125973 records in the training set and 22554 records in the testing set in this dataset.

There are three types of Machine Learning-based DDoS detection methods currently available. To construct the detection model, supervise ML methods that employ produced labelled network traffic datasets. The supervised methods have two significant challenges. To begin with, creating labelled network traffic datasets is time and computational intensive. The supervised machine learning techniques are unable to forecast new legitimate and attack activities without a constant update of

their detection models. Second, the inclusion of a high volume of irrelevant normal data in the incoming network traffic makes supervised ML classifiers noisy and lowers their performance. Unsupervised methods, unlike the first, do not require a labelled dataset to build the detection model. The difference between DDoS and regular traffic is determined by examining their underlying distribution characteristics. The primary disadvantage of unsupervised methods is the high rate of false positives. The distance between points becomes meaningless in high-dimensional network traffic statistics and tends to homogenise. The ‘curse of dimensionality,' as it is called, hinders unsupervised methods from correctly detecting assaults [9]. The flexibility to work on labelled and unlabeled datasets allows semi-supervised ML methods to benefit from both supervised and unsupervised approaches. Additionally, combining supervised and unsupervised methods improves accuracy while lowering false positive rates. Semi-supervised methods, on the other hand, are constrained by the disadvantages of both approaches. As a result, semi-supervised methods need a complex implementation of its components to solve the challenges.

SYSTEM STUDY

EXISTING SYSTEM:

The current study is prompted by a number of unanswered concerns, including: Are data breaches caused by cyber-attacks rising, decreasing, or stabilising? A principled response to this issue will provide us with a clear picture of the current state of cyber dangers. Previous research has not provided an answer to this question. Researchers have recently begun simulating data breach instances. Between the years 2000 and 2008, Maillart and Sornette investigated the statistical features of personal identity losses in the United States. They discovered that the number of data breaches skyrocketed from 2000 to July 2006, but then levelled off. Wheatley et al. looked examined a dataset that was compiled from organisational breach instances that occurred between 2000 and 2015. They discovered that the frequency of major breach occurrences (those involving more than 50,000 records) affecting US enterprises remains unaffected by time, but that the frequency of major breach episodes affecting non-US enterprises is on the rise.

PROPOSED SYSTEM:

The following three contributions are made in this study. First, we show that stochastic processes, rather than distributions, should be used to characterise both hacking breach incident interarrival durations (which represent event frequency) and breach magnitude. We show that the inter-arrival periods and breach sizes may be predicted using stochastic process models. To our knowledge, this is the first work to indicate that stochastic processes should be utilised to describe these cyber threat elements rather than distributions. We also demonstrate that when estimating inter-arrival durations and breach widths, the dependency must be taken into account; otherwise, the prediction results would be inaccurate. To our knowledge, this is the first paper that demonstrates the existence of this dependency and the consequences of disregarding it. Third, we analyse the cyber hacking breach instances in both qualitative and quantitative ways. We discover that while the situation is deteriorating in terms of incident inter-arrival time as hacking breach incidents become more common, We hope that the findings of this study will spur more research into other risk mitigation strategies. Insurance firms, government organisations, and regulators can benefit from such information

RELATED WORKS

In SDN architecture, the logically centralised control-plane is a two-edged sword. On the one hand, its global network perspective and dynamic updating of forwarding rules make DDoS assaults easier to identify and respond to. However, it creates a single point of failure in the network, making it vulnerable to DDoS attacks. We will offer a selection of prior research on DDoS attack detection and mitigation using SDN in this part. AVANT-GUARD [4] is a framework that detects TCP SYN flood assaults using two modules. Connection migration is the first module, which is placed as an extension in the data plane and acts as a proxy for ingress SYN packets to prevent saturation attacks from reaching the control plane. The actuating triggers module, which is inserted in the controller, is the second module. If malicious network traffic is detected, the actuating triggers module initiates an event that instructs the controller to establish a flow rule in the switch to reduce reaction time. The most noticeable side effect of this technique is the performance cost, since it relies on connection migration, which necessitates classification of each flow. To implement AVANT-GUARD capabilities, all switches in the data plane must be upgraded. As an extension of AVANT-Guard, the authors of [5] offered a solution named LineSwitch. For all TCP connections from a specific IP source, LineSwitch uses the SYN proxy approach in data plane switches. It will apply probabilistic black-listing and prohibit all TCP packets from this IP source if it detects the SYN flood assault. This approach may successfully prevent SYN flood attacks against SDN networks at a low cost; however, when using the LineSwitch mechanism, it is necessary to update data plane switches, comparable to AVANT-GUARD.

METHODOLOGY

The methodology diagram for the suggested approach is shown in Figure 1. It consists of a number of interconnected components that work together to implement the system. The technique proposed here includes four basic steps: traffic network entropy estimate, coclustering, computation of information gain ratio, and network traffic categorization. Data on Network Traffic: NSL-KDD[18] network traffic data is employed in the proposed work. The data set is known as NSL KDD, and it also presents a solution to the KDD 99 data set's fundamental issues. Though there is a newer version of KDD for the data set, it is still suffering from some issues, and thus for the existing real network it may not be a perfect representative because public data sets are lacking in IDSs based networks. It can also be applied based on the belief that some effective as well as benchmark data sets are applicable so that investigators can be assisted in such a way that the

Figure 1: Diagram of the Methodology

Entropy Calculation: In the NSL KDD dataset, two file size distribution (FSD) characteristics, such as source and destination bytes, are employed to estimate average entropy using the entropycalculation method. This permits data to be reduced in its highdimensionality. Network Traffic Co-clustering: In this stage, the traffic data is separated into three primary groups using a co clustering technique such as the spectral co clustering technique. When compared to other clustering methods, the spectral coclustering approach is deemed simple and delivers higher accuracy. The purpose of network traffic splitting is to reduce the quantity of data that is expected to be categorised by excluding typical data.

NETWORK TRAFFIC DATASETS

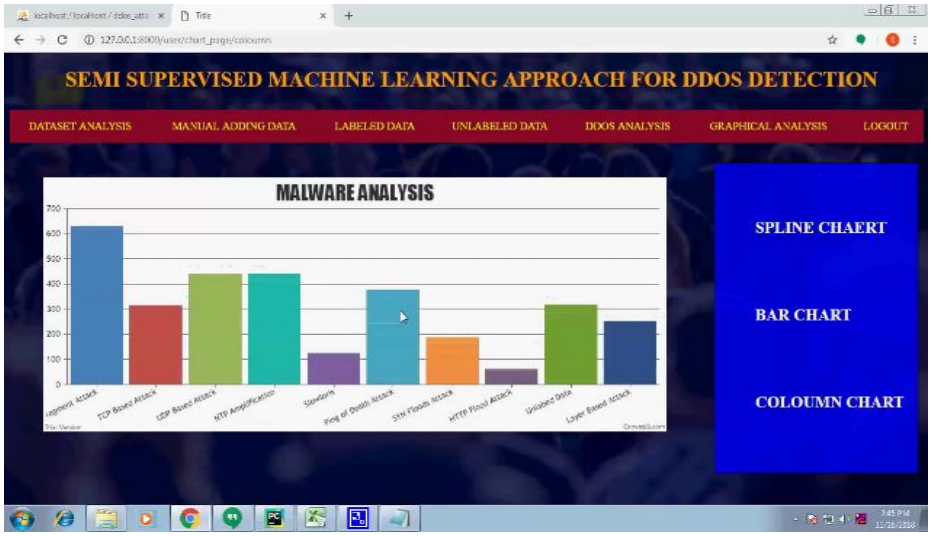
There are four forms of assaults in the NSL-KDD dataset: DoS, Probe, R2L, and U2R. It includes 41 features separated into three categories: basic, traffic, and content features. Both the training and testing sets of this dataset comprise a total of 148,517 entries. This dataset was chosen for three reasons. For starters, it is frequently utilised in the literature for IDS benchmarking. Second, it contains a significant amount of attack traffic, with DoS assaults accounting for 34.62 percent of the dataset. It also solves some of the issues that plagued its predecessors, KDD Cup'99 and DARPA'98 [12], such as redundancy and duplication of records. The UNB ISCX IDS 2012 dataset includes tagged network traces, as well as complete packet payloads in pcap format and pertinent profiles. The data is organised into profiles, which include thorough descriptions of intrusions as well as abstract distribution models for apps, protocols, and lower-level network elements. During the week, a testbed network of 21 linked workstations is utilised to construct seven sections of the dataset, one

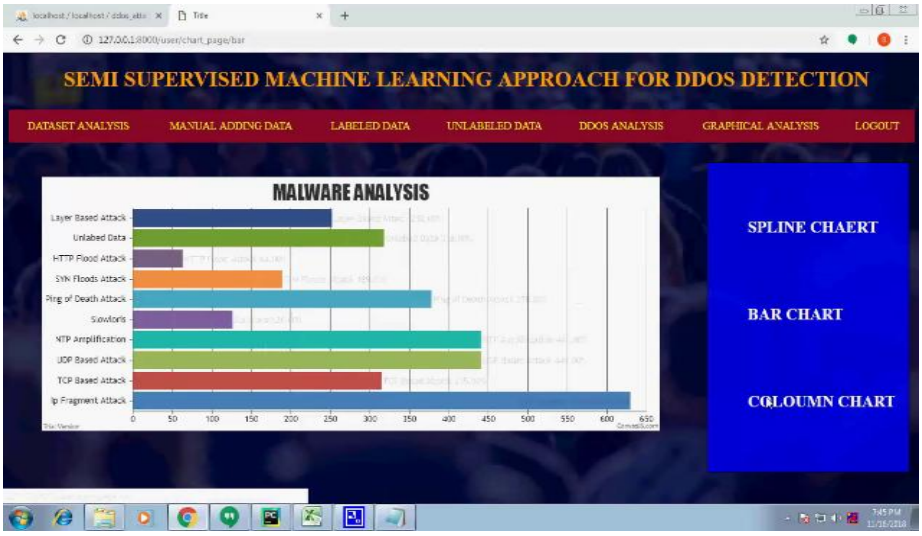
of which is dedicated only to DDoS assaults. This is the primary rationale for utilising our dataset in this study. The ISCX-Dataset-15-June dataset comprises 19 features and a total of 196,032 DDoS and regular traffic records. We can see that DDoS traffic accounts for 19.11 percent of the sample, making it an essential benchmark for DDoS detection systems.

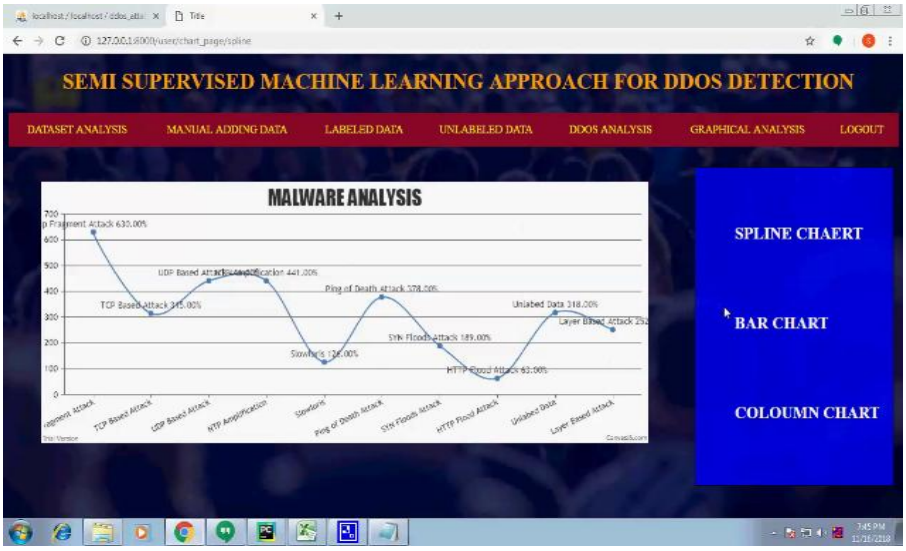
Using a setup of 60% and 40%, the datasets above are divided into train and test subsets. The Extra-Trees ensemble classifiers are fitted using the train subsets, and the suggested method is tested using the test subsets. The train subsets are normalised before fitting the classifiers using the M inMax technique described in Section 4.3.2. We also observe that additional attack traffic is kept out of the datasets because this research is solely concerned with detecting DDoS attacks.

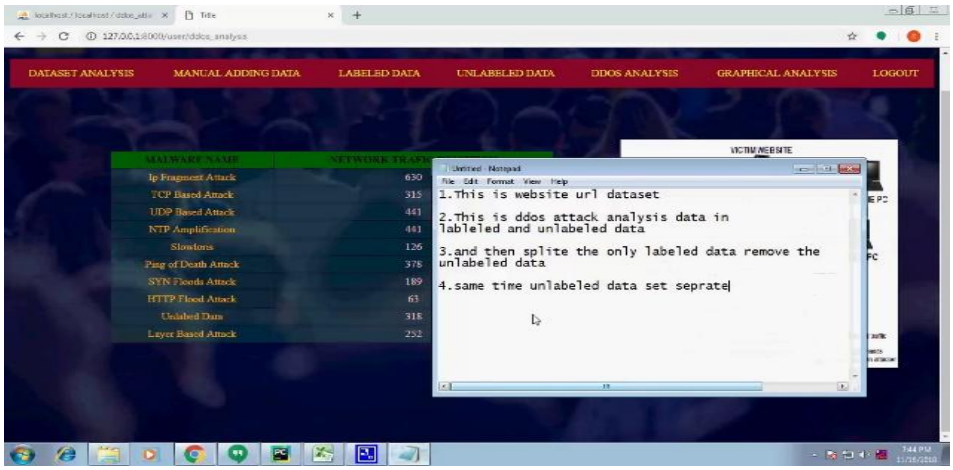
RESULTS

In this section we give the obtained results of the experiments. The obtained results illustrate the contribution of each component of the proposed approach and the entire approach. To validate the results we compared them with the state-of-the-art DDoS detection approaches.

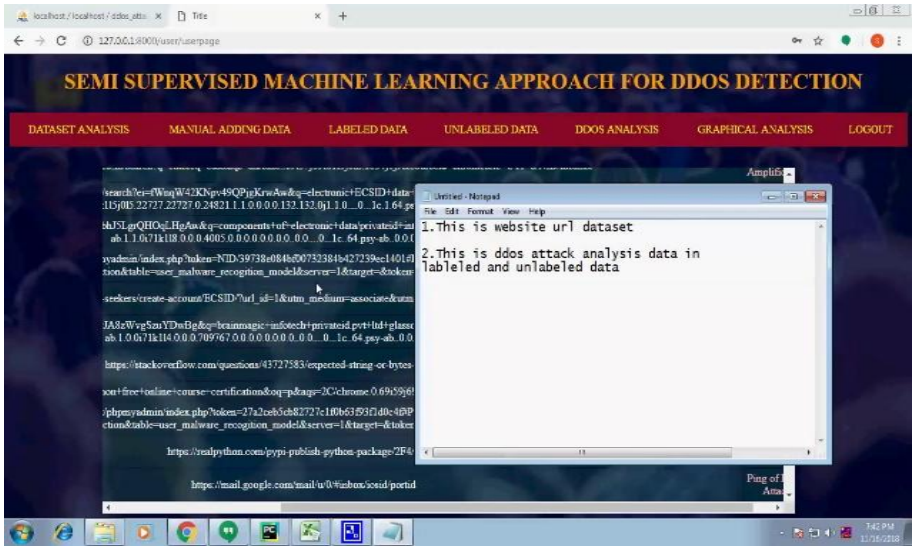


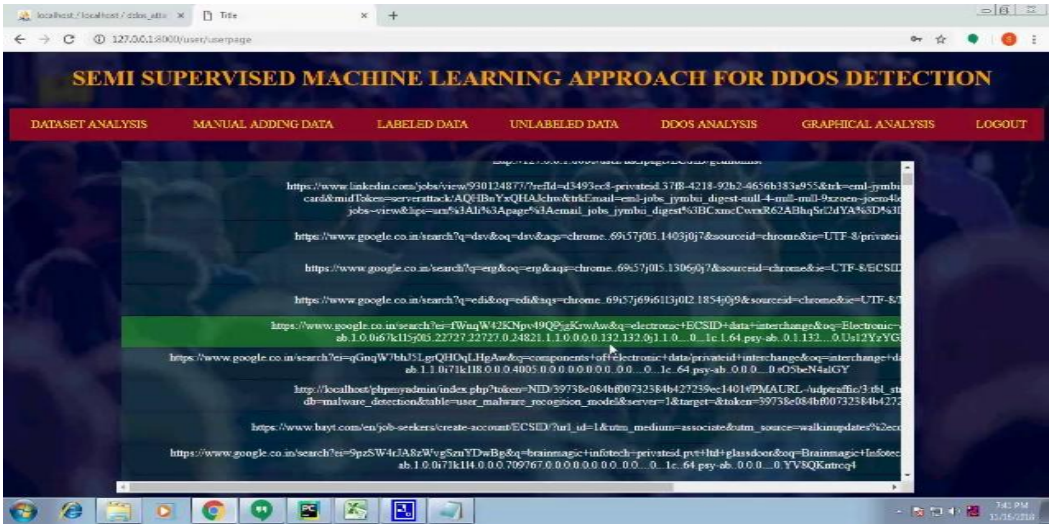



SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION
DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE ANALYSIS
SPLINE CHAERT
BAR CHART
COLOUMN CHART


DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE NAME
NETWORK TRAFFIC POSSITION
Ip Fragment Attack 630
TCP Based Attack 315
UDP Based Attack 441
NTP Amplification 441
Slowloris 126
Ping of Death Attack 378
SYN Floods Attack 189
HTTP Flood Attack 63
Unlabed Data 318
Layer Based Attack 252
1.This is website url dataset
2.This is ddos attack analysis data in labeled and unlabeled data
3.and then splite the only labeled data remove the unlabeled data
4.same time unlabeled data set seprate


SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION
DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE NAME
NETWORK TRAFFIC POSSITION
Ip Fragment Attack 630
TCP Based Attack 315
UDP Based Attack 441
NTP Amplification 441
Slowloris 126
Ping of Death Attack 378
SYN Floods Attack 189
HTTP Flood Attack 63
Unlabed Data 318
Layer Based Attack 252
VICTIM WEBSITE
ZOMBIE PC
ZOMBIE PC
ZOMBIE PC
ZOMBIE PC
ZOMBIE PC
ZOMBIE PC
Internet
ATTACKER'S PC





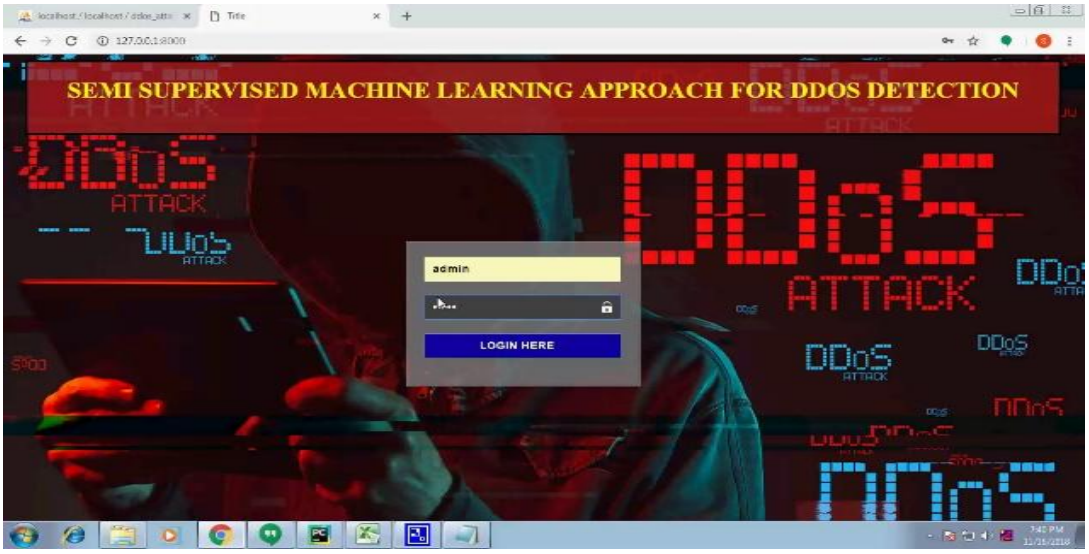
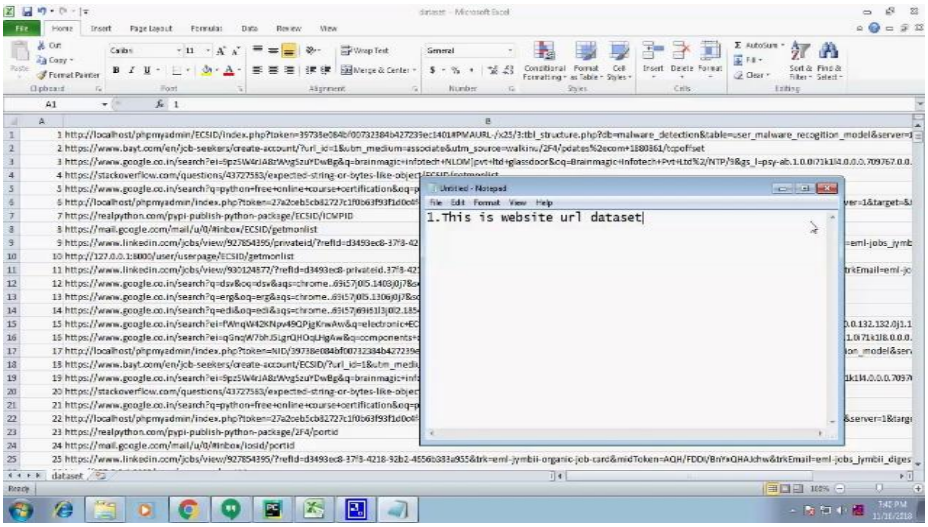
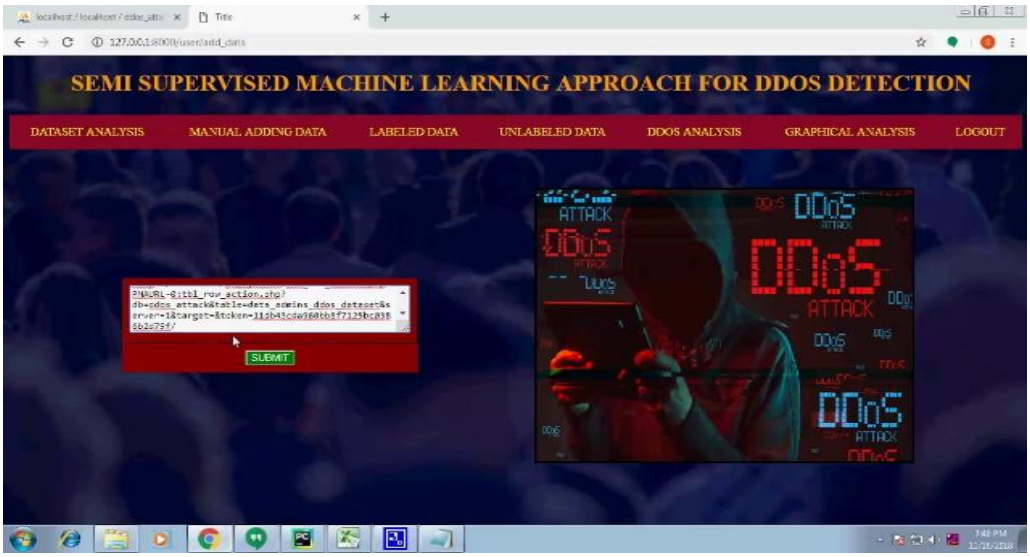


FIG ADMIN LOGIN





CONCLUSION

We looked at a hacker breach dataset from the standpoints of event inter-arrival time and breach magnitude, and found that both should be described using stochastic processes instead of distributions. The fitting and prediction accuracies of the statistical models proposed in this study are good. We suggest, in particular, employing a copula-based methodology to forecast the combined chance of an incident of a specified magnitude of breach size occurring in the future. Statistical tests reveal that the approaches described in this study are superior to those in the literature, since the latter overlooked both temporal correlations and the connection between incident inter-arrival periods and breach sizes. To get further understanding, we did qualitative and quantitative assessments. We reached a number of cybersecurity conclusions, including the factThe methods given in this study can be used or changed to evaluate comparable datasets.

REFERENCES

[1] P. R. Clearinghouse. *Privacy Rights Clearinghouse's Chronologyof Data Breaches*. Accessed: Nov. 2017. [Online]. Available:https://www.privacyrights.org/data-breaches

[2] ITR Center. *Data Breaches Increase 40 Percent in 2016, FindsNew Report From Identity Theft Resource Center and CyberScout*.Accessed: Nov. 2017. [Online]. Available: 2016databreaches.html
[3] C. R. Center. *Cybersecurity Incidents*. Accessed: Nov. 2017. [Online].Available: https://www.opm.gov/cybersecurity/cybersecurity-incidents
[4] *IBM Security*. Accessed: Nov. 2017. [Online]. Available:https://www.ibm.com/security/data-breach/index.html
[5] NetDiligence. *The 2016 Cyber Claims Study*. Accessed: Nov. 2017.
[Online].Available: https://netdiligence.com/wp-ontent/uploads/2016/0/P02_NetDiligence-2016-Cyber-Claims-Study-ONLINE.pdf
[6] M. Eling and W. Schnell, "What do we know about cyber risk and cyberrisk insurance?" *J. Risk Finance*, vol. 17, no. 5, pp. 474–491, 2016.
[7] T. Maillart and D. Sornette, "Heavy-tailed distribution of cyber-risks,"
Eur. Phys. J. B, vol. 75, no. 3, pp. 357–364, 2010.
[8] R. B. Security. *Datalossdb*. Accessed: Nov. 2017. [Online]. Available: V https://blog.datalossdb.org
[9] B. Edwards, S. Hofmeyr, and S. Forrest, "Hype and heavy tails: A closerlook at data breaches," *J. Cybersecur.*, vol. 2, no. 1, pp. 3–14, 2016.
[10] S. Wheatley, T. Maillart, and D. Sornette, "The extreme risk of personal
data breaches and the erosion of privacy," *Eur. Phys. J. B*, vol. 89, no. 1,p. 7, 2016.
[11] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling ExtremalEvents: For Insurance and Finance*, vol. 33. Berlin, Germany:Springer-Verlag, 2013.
[12] R. Böhme and G. Kataria, "Models and measures for correlation incyber-insurance," in *Proc. Workshop Econ. Inf. Secur. (WEIS)*, 2006,pp. 1–26.
[13] H. Herath and T. Herath, "Copula-based actuarial model for pricingcyber-insurance policies," *Insurance Markets Companies: Anal. ActuarialComput.*, vol. 2, no. 1, pp. 7–20, 2011.
[14] A. Mukhopadhyay, S. Chatterjee, D. Saha, A. Mahanti, andS. K. Sadhukhan, "Cyber-risk decision models: To insure it or not?"*Decision Support Syst.*, vol. 56, pp. 11–26, Dec. 2013.
[15] M. Xu and L. Hua. (2017). *Cybersecurity Insurance: Modelingand Pricing*. [Online]. Available: https://www.soa.org/research-reports/2017/cybersecurity-insurance
[16] M. Xu, L. Hua, and S. Xu, "A vine copula model for predicting theeffectiveness of cyber defense early-warning," *Technometrics*, vol. 59,no. 4, pp. 508–520, 2017.
[17] C. Peng, M. Xu, S. Xu, and T. Hu, "Modeling multivariate cybersecurityrisks," *J. Appl. Stat.*, pp. 1–23, 2018.
[18] M. Eling and N. Loperfido, "Data breaches: Goodness of fit, pricing,and risk measurement," *Insurance, Math. Econ.*, vol. 75, pp. 126–136,Jul. 2017.
[19] K. K. Bagchi and G. Udo, "An analysis of the growth of computer and
Internet security breaches," *Commun. Assoc. Inf. Syst.*, vol. 12, no. 1,p. 46, 2003.
[20] E. Condon, A. He, and M. Cukier, "Analysis of computer securityincident data using time series models," in *Proc. 19th Int. Symp. Softw.Rel. Eng. (ISSRE)*, Nov. 2008, pp. 77–86.