ZKG INTERNATIONAL

ISSN: 2366-1313

Predicting Cyber Hacking Breaches using Machine learning Algorithm

[1]BOKKA CHANDRA SEKHAR, [2]M. NARESH

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract: *Analysing cyber incident data units is important to deepening our knowledge of potential situational evolution. This is an incredibly new research topic, and there is still a lot of research to be done. In this paper, we report on a similar 12-year (2005–2017) data analysis of breach incident data from game hacking involving malware attacks. In contrast to the results suggested in the literature, we show that each inter-arrival hacking breach event and breach size should be modeled by stochastic techniques rather than because they are automated. Show correlation. We then recommend stochastically specific models to accommodate interarrival incidence and gap size. We also show that these models can predict future events and default sizes. We performed qualitative and quantitative fashion analysis on the data set to gain deeper insight into the evolution of hacking breach incidents. We draw several cyber security insights, including that the threat of cyber-attacks is getting worse in frequency but not in the extent of their damage.*

Keywords: *Analysis cyber incidents, stochastic process, prediction of hacking*

I. INTRODUCTION

As the data grows exponentially, the risk to the records will also increase, so it is very important to relax the statistics. Cyber threats are becoming more serious as data is the most valuable commodity in the 21st century. Various problems of cybercrime are protected by cyber security. Today, cybercrime is a significant problem, but it differs widely from traditional crimes such as burglary, hacking and theft [1].

Cybercrimes do not imply the complete presence of a hacker, unlike other types of (criminal) crime. Cybercriminals these days have started targeting and exploiting the infrastructure that enables online payments and e-commerce. These attacks have become more severe day by day and the threat will only increase. Now the only strategy to counter attacks and protect yourself from a breach is to teach the device how to deal with attacks that counter them. Versions can be tested with specific forms of attacks, attacks on the digital community, and physical attacks. Breach occurs both locally and remotely, so the version can learn how the attacks will happen and has enough time to build a new layer of protection. If the attack is physical, the version uses facial data and voice along with the user's biometrics to authenticate with legal personnel. When the model detects a match, it takes preemptive action to reduce the statistics [2].

The attacks could have been greatly reduced if there was no threat given that there was nothing to target. But because software regularly moves quickly across markets, is designed by men, and people make mistakes, there will always be vulnerabilities that allow users to attack and corrupt devices. Any device with just a small vulnerability can be attacked.

Keeping a community safe from attacks is almost impossible because there will usually be loopholes that attackers can exploit. However, we can teach the device how to protect itself from an attack or keep itself clean before it actually starts. All an attacker wants to hijack any device is a one-touch vulnerability [3].

Hackers can also have a motive to attack and exploit security. This reason can be considered protest, economic gain, theft of facts, competition or priority to identify security flaws in the community. An advantage of security is exploiting vulnerabilities that can already be recognized. For example, flaws in SQL, Report Exchange Protocol, Hypertext Exchange Protocol (HTTP), Telnet, PHP, SSH, etc. The most popular hacking methods currently used by hackers are listed below.

Checking for known flaws on internet-connected systems is done using a vulnerability scanner. Applications called port scanners are frequently used by attackers to verify the access points that are accessible to attack the machine

II. LITERATURE SURVEY

Hammouchi et. Al [4] proposed a STRisk predictive system where they expand the scope of the prediction task by bringing into play the social media dimension. They study over 3800 US organizations including both victim and non-victim organizations. For each organization, they design a profile composed of a variety of externally measured technical indicators and social factors. In addition, to account for unreported incidents, they consider the non-victim sample to be noisy and propose a noise correction approach to correct mislabeled organizations. They then build several machine learning models to predict whether an organization is exposed to experience a hacking breach. By exploiting both technical and social features, they achieve an Area Under Curve (AUC) score exceeding 98%, which is 12% higher than the AUC achieved using only technical features. Furthermore, our feature importance analysis reveals that open ports and expired certificates are the best technical predictors, while spreadability and agreeability are the best social predictors.

Mandal et. Al [5] aimed at considering the different aspects of social events, responses and their relations to further improve the classification of the social sentiment. The proposed method covers not only the response due to major social events but also predicting and generating alert for situations of significant social importance. The approach has made use of Twitter datasets and performed aspectbased sentiment analysis on the obtained text data. It is shown to outperform the state-of-the-art methods.

Poyraz et. al [6] investigates various factors that can affect the monetary impact of data breaches on companies. This paper introduces a model for the total cost of a mega data breach based on a data set created from multiple sources that categorises stolen data for U.S. residents as personally identifiable information (PII) and sensitive personally identifiable information (SPII). They use a rigorous stepwise regression analysis that includes polynomial and factorial multilevel effects of the independent variables. There are three significant findings. First, our model finds a significant relation between total data breach cost and revenue, the total amount of PII and SPII, and class action lawsuits. Second, the categorisation of personal information as sensitive and non-sensitive explains the cost better than previous work. Finally, all of the independent variables demonstrate multilevel factorial interactions.

Guru Akhil et. al [7] reported a measurable examination of a break occurrence datasets relating to 11 years (2005–2018) of digital hacking exercises are incorporate breach assaults. They show that, as opposed to the discoveries revealed in the writing, both the hacking break going to happen in the middle, appearance times and the penetrate size need to be shown by stochastic cycles, rather than by disseminations since they show auto associations. At that point, they propose specific stochastic cycle models to independently fit the between entry time and the break size. They moreover appear that the between 21 appearance times and the break sizes can be anticipated by these models. They conduct subjective and quantitative pattern reviews on the dataset in arrange to pick up advance insights into the progress of hacking break episodes. They draw a lot of knowledge from network protection bits, counting that the risk of digital hacks is certainly deteriorating as distant as their repeat is concerned, but not as to the degree of their damage.

Fang et. al [8] initiated the study of modeling and predicting risk in enterprise-level data breaches. This problem is challenging because of the sparsity of breaches experienced by individual enterprises over time, which immediately disqualifies standard statistical models because there are not enough data to train such models. As a first step towards tackling the problem, they propose an innovative statistical framework to leverage the dependence between multiple time series. In order to validate the framework, they apply it to a dataset of enterprise-level breach incidents. Experimental results show its effectiveness in modeling and predicting enterprise-level breach incidents.

Kure et. al [9] aims for an effective cybersecurity risk management (CSRM) practice using assets criticality, predication of risk types and evaluating the effectiveness of existing controls. They follow a number of techniques for the proposed unified approach including fuzzy set theory for the asset criticality, machine learning classifiers for the risk predication and comprehensive assessment model (CAM) for evaluating the effectiveness of the existing controls. The proposed approach considers relevant CSRM concepts such as asset, threat actor, attack pattern, tactic, technique and procedure (TTP), and controls and maps these concepts with the VERIS community dataset (VCDB) features for the risk predication. The experimental results reveal that using the fuzzy set theory in assessing assets criticality supports stakeholder for an effective risk management practice. Furthermore, the results have demonstrated the machine learning classifiers exemplary performance to predict different risk types including denial of service, cyber espionage and crimeware. An accurate prediction of risk can help organisations to determine the suitable controls in proactive manner to manage the risk.

III. PROPOSED 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 interarrival 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 interarrival 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

SYSTEM ARCHITECTURE

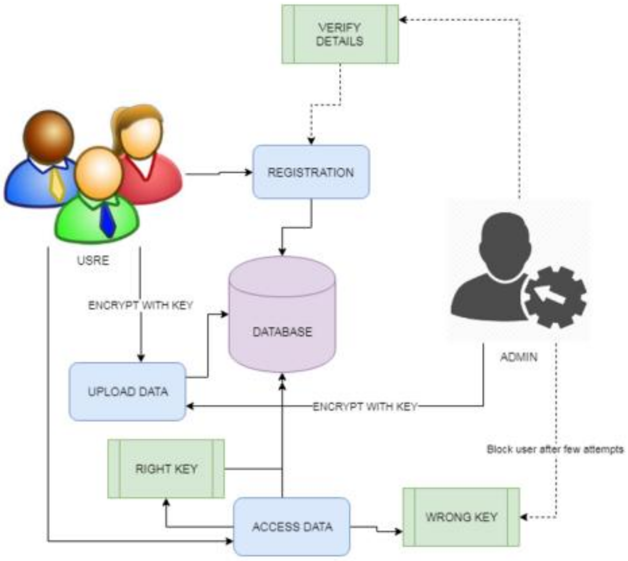


Fig.1 Proposed system architecture

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.

Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite dimensional space, which can be used for classification, regression, or other tasks like outliers' detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

IV. RESULTS

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 accessing 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 analysed through this pictorial representation in order to better understand of the data details.

EDA results

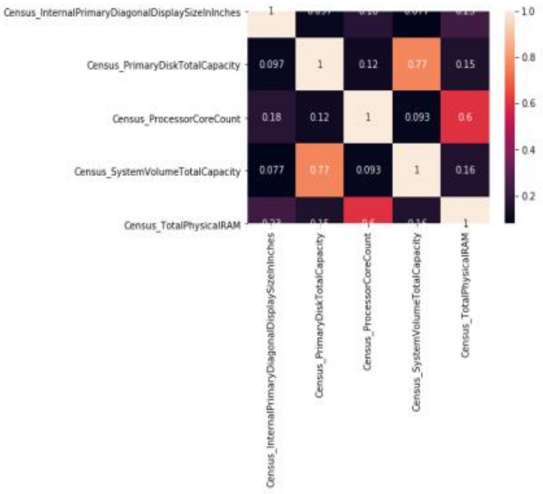
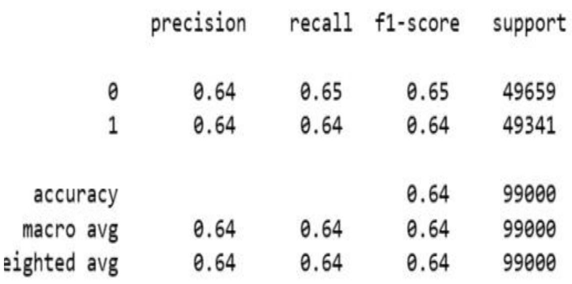


Fig.2 Confusion matrix



```
              precision    recall  f1-score   support

           0       0.64      0.65      0.65     49659
           1       0.64      0.64      0.64     49341

    accuracy                           0.64     99000
   macro avg       0.64      0.64      0.64     99000
eighted avg       0.64      0.64      0.64     99000
```

Fig.3 Performance analysis

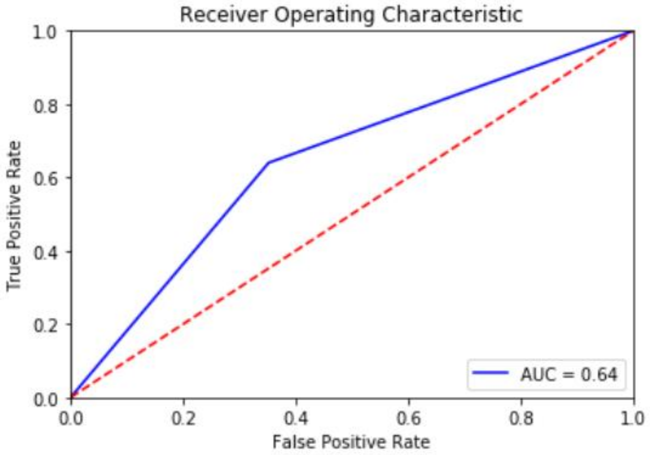


Fig.4 Performance metrics

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 work 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. 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] P. R. Clearinghouse. "Privacy Rights Clearinghouse's Chronology of Data Breaches". Accessed: Nov. 2017.

[2] ITR Center. "Data Breaches Increase 40 Percent in 2016, Finds New Report From Identity Theft Resource Center and Cyber Scout".

[3] C. R. Center. Cybersecurity Incidents. Accessed: Nov. 2017. [Online]. Available: https://www.opm.gov/cybersecurity/cybersecurity-incidents

[4] Prasadu Peddi (2019), Data Pull out and facts unearthing in biological Databases, International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.

[5] NetDiligence. The 2016 Cyber Claims Study. Accessed: Nov. 2017.

[6] H. Hammouchi, N. Nejjari, , "STRisk: A SocioTechnical Approach to Assess Hacking Breaches Risk,".

[7] Mandal, S, (2020). "Exploiting Aspect-Classified Sentiments for Cyber Crime Analysis and Hack Prediction" .

[8] Poyraz, O.I., Canan, M., McShane, M. et al. "Cyber assets at risk: monetary impact of U.S. personally identifiable information mega data breaches". Geneva Pap Risk Insur Issues Pract 45, 616–638 (2020).]

[9] Guru Akhil, C., Kumar, A.K. (2022). "Cyber Hacking Breaches for Demonstrating and Forecasting".

[10] Prasadu Peddi (2016), Comparative study on cloud optimized resource and prediction using machine learning algorithm, ISSN: 2455-6300, volume 1, issue 3, pp: 88-94.

[11] Afreen Bari, Dr. Prasadu Peddi. (2021). Review and Analysis Load Balancing Machine Learning Approach for Cloud Computing Environment.Annals of the Romanian Society for Cell Biology,25(2), 1189–1195.