JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 07, ISSUE 08, 2020

Modeling and Predicting Cyber Hacking Breaches using Support Vector Machine Algorithm

Kommu Anusha, Mounika Pasam

Department of Computer Science and Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

Abstract

Analyzing cyber incident data sets is an important 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 cyber security 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.

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

1. Introduction

Data breaches are one of the most devastating cyber incidents. The Privacy Rights Clearinghouse [1] reports 7,730 data breaches between 2005 and 2017, accounting for 9,919,228,821 breached records. The Identity Theft Resource Center and Cyber Scout reports 1,093 data breach incidents in 2016, which is 40% higher than the 780 data breach incidents in 2015. The United States Office of Personnel Management (OPM) [2] reports that the personnel information of 4.2 million current and former Federal government employees and the background investigation records of current, former, and prospective federal employees and contractors (including 21.5 million Social Security Numbers) were stolen in 2015. The monetary price incurred by data breaches is also substantial. IBM [3] reports that in year 2016, the global average cost for each lost or stolen record containing sensitive or confidential information was $158. NetDiligence reports that in year 2016, the median number of breached records was 1,339, the median per-record cost was $39.82, the average breach cost was $665,000, and the median breach cost was $60,000 [4]. While technological solutions can harden cyber systems against attacks, data breaches continue to be a big problem. This motivates us to characterize the evolution of data breach incidents. This not only will deep our understanding of data breaches, but also shed light on other approaches for mitigating the damage, such as insurance [5]. Many believe that insurance will be useful, but the development of accurate cyber risk metrics to guide the assignment of insurance rates is beyond the reach of the current understanding of data breaches (e.g., the lack of modeling approaches). Recently, researchers started modeling data breach incidents. 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.

2. Literature survey

Hammouchi et. Al [6] 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 [7] 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 aspect-based sentiment analysis on the obtained text data. It is shown to outperform the state-of-the-art methods.

Poyraz et. al [8] 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 [9] 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 [10] 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 [11] 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.

Subramanian et. al [12] designed a model by using machine learning to defend a website from security breaches. The primary aim of this research work is to create a machine learning model, which trains in Realtime and monitors the website or a system and trains from the state-of-art attacks. The proposed model has created a web application using Django, which takes the data from multiple sources such as Amazon, Flipkart, Snapdeal, and Shop clues, which shows the data that is safe to obtain from the website. Then, the data will be sorted on our page and then it will be made secured and illegal for the external people to access the data from our website and the proposed model will monitor the website 24/7. The model is trained daily and it generates predictions from the several of datasets available and from the previous state-of-the-art attacks. This model will be trained from the existing datasets and the history of attacks and breaches on our website.

3. 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 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.

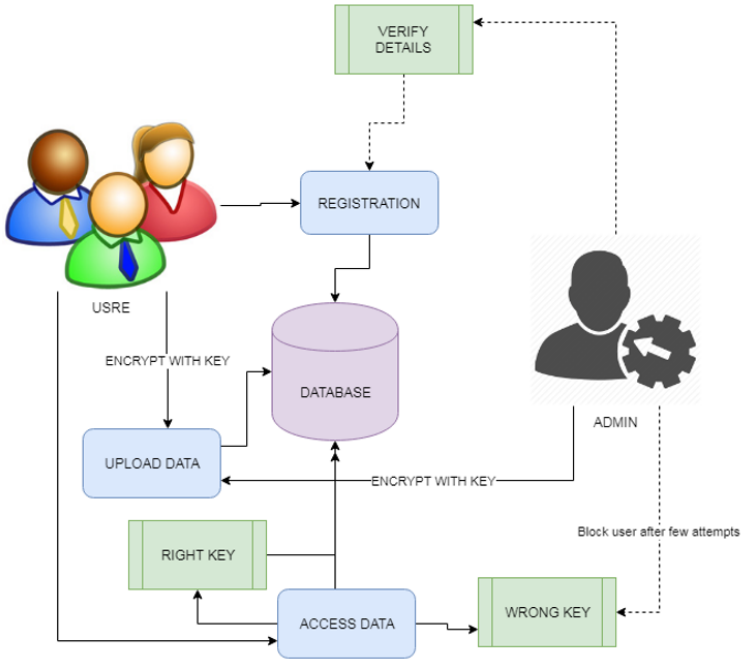


Fig. 1: block diagram of proposed system.

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.

4. 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.

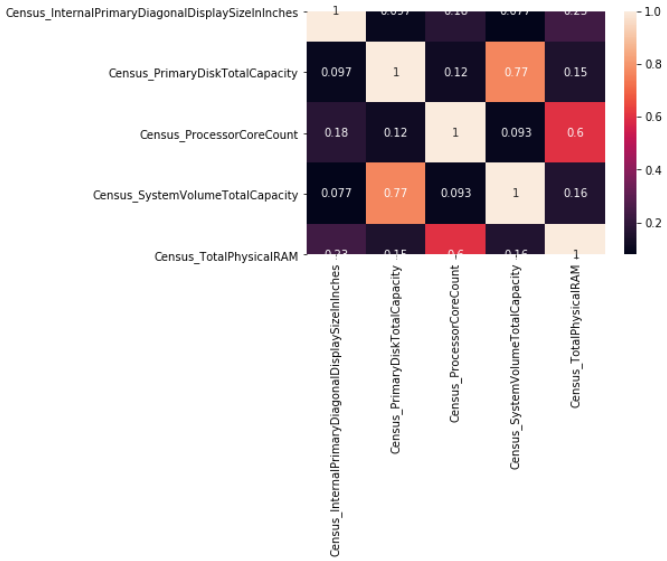
User Permissions

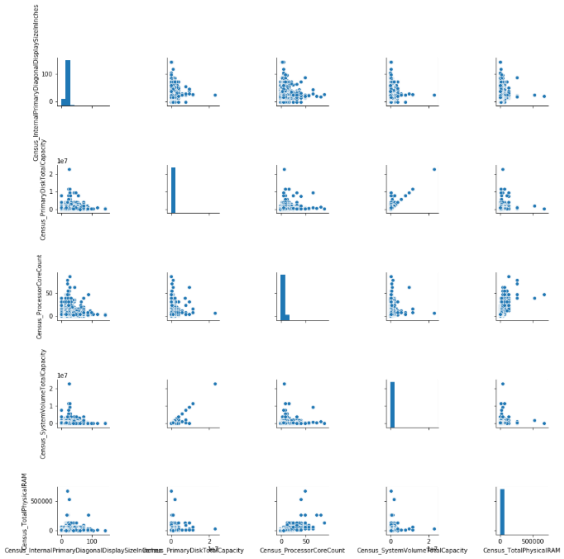
The data from any resources are allowed to access the data with only permission from administrator. Prior to access data, users are allowed by admin to share their data and verify the details which are provided by user. If user is accessing the data with wrong attempts, then, users are blocked accordingly. 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 analyzed through this pictorial representation in order to better understand of the data details.

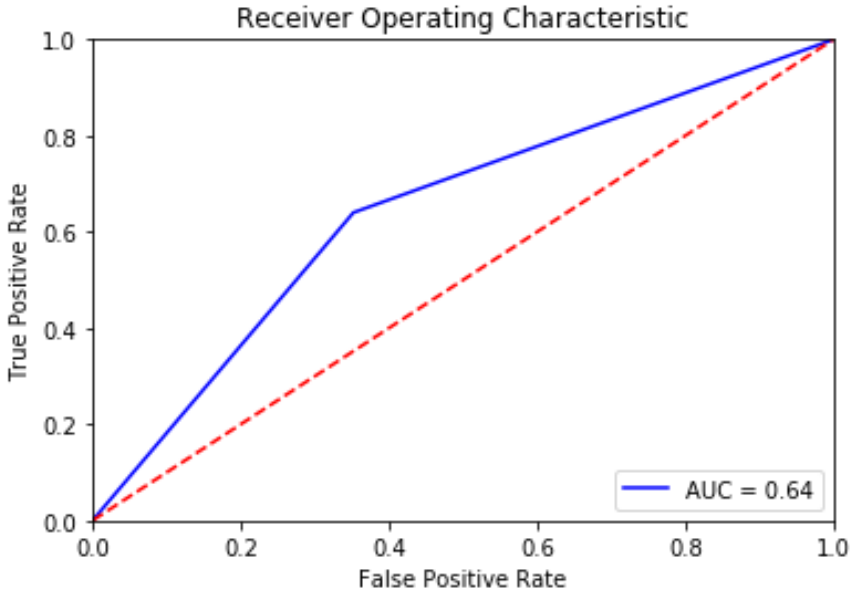
EDA results





Classification report

	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
weighted avg	0.64	0.64	0.64	99000



Feature Importance

	feature	importance	normalized_importance	cumulative_importance
0	index	131055	0.068687	0.068687
1	AvSigVersion	129938	0.068102	0.136789
2	CityIdentifier	118341	0.062024	0.198812
3	Census_InternalPrimaryDiagonalDisplaySizeInInches	108861	0.057055	0.255867
4	Census_SystemVolumeTotalCapacity	106640	0.055891	0.311758
...	...	...	...	...
65	Census_IsAlwaysOnAlwaysConnectedCapable	185	0.000097	0.999942
66	OsVer	111	0.000058	1.000000
67	Census_IsPortableOperatingSystem	0	0.000000	1.000000
68	Census_DeviceFamily	0	0.000000	1.000000
69	SMode	0	0.000000	1.000000

70 rows × 4 columns

5. 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.

Future work

There are many open problems that are left for future research. For example, it is both interesting and challenging to investigate how to predict the extremely large values and how to deal with missing data (i.e., breach incidents that are not reported). It is also worthwhile to estimate the exact occurring times of breach incidents. Finally, more research needs to be conducted towards understanding the predictability of breach incidents (i.e., the upper bound of prediction accuracy).

References

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

[2] ITR Center. Data Breaches Increase 40 Percent in 2016, Finds New Report From Identity Theft Resource Center and CyberScout. Accessed: Nov. 2017. [Online]. Available: http://www.idtheftcenter.org/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-content/uploads/2016/10/P02_NetDiligence-2016-Cyber-Claims-Study-ONLINE.pdf

[6] H. Hammouchi, N. Nejjari, G. Mezzour, M. Ghogho and H. Benbrahim, "STRisk: A Socio-Technical Approach to Assess Hacking Breaches Risk," in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2022.3149208.

[7] Mandal, S., Saha, B., Nag, R. (2020). Exploiting Aspect-Classified Sentiments for Cyber-Crime Analysis and Hack Prediction. In: Kar, N., Saha, A., Deb, S. (eds) Trends in Computational Intelligence, Security and Internet of Things. ICCISIoT 2020. Communications in Computer and Information Science, vol 1358. Springer, Cham. https://doi.org/10.1007/978-3-030-66763-4_18

[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). https://doi.org/10.1057/s41288-020-00185-4

[9] Guru Akhil, T., Pranay Krishna, Y., Gangireddy, C., Kumar, A.K. (2022). Cyber Hacking Breaches for Demonstrating and Forecasting. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_106

[10] Z. Fang, M. Xu, S. Xu and T. Hu, "A Framework for Predicting Data Breach Risk: Leveraging Dependence to Cope With Sparsity," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2186-2201, 2021, doi: 10.1109/TIFS.2021.3051804.

[11] Kure, H.I., Islam, S., Ghazanfar, M. et al. Asset criticality and risk prediction for an effective cybersecurity risk management of cyber-physical system. Neural Comput & Applic 34, 493–514 (2022). https://doi.org/10.1007/s00521-021-06400-0

[12] R. R. Subramanian, R. Avula, P. S. Surya and B. Pranay, "Modeling and Predicting Cyber Hacking Breaches," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 288-293, doi: 10.1109/ICICCS51141.2021.9432175.