MODELLING AND PREDICTING CYBER HACKING BREACHES

E.Rupa[1], Thatikonda Vineela[2], Singu Pooja[3], Sunkireddy Sai teja[4], S.Akshita[5], S.Mohan[6]

[1]Assistant professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[2,3,4,5,6] IVth Btech Student, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

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 auto correlations. 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.

1. Introduction

Introduction Data breaches are one of the most devastating cyber incidents. The Privacy Rights Clearinghouse 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) 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 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. While technological solutions can harden cyber systems against attacks, data breaches continue to be a big problem. This motivates usto 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.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.SURVEY

Hammouchi et. Al 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 thebest technical predictors, while spreadability and agreeability are the best social predictors. Mandal et. Al 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 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.

3.SYSTEM ANALYSIS

The present study is motivated by several questions that have notbeen 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 intothe overall situation of cyber threats. This question was not answered by previous studies. Specifically, the dataset analyzed in 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 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 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 ontheir 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 increasesfrom 2000 to July 2006 but remains stable thereafter. Edwards etal. analyzed a dataset containing 2,253 breach incidents that spanover a decade (2005 to 2015). They found that neither the size northe frequency of data breaches has increased over the year.

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. Wefind 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 processmodels can predict the inter-arrival times and the breach sizes. Tothe 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.

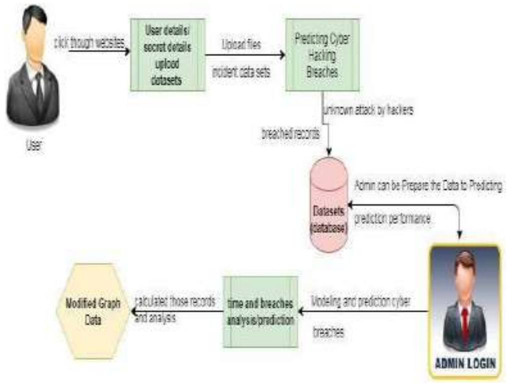
Second, we discover a positive dependence between the incidents inter-arrival times and the breach sizes, and show that this dependence can be adequately described by a particular copula. We also show that when predicting inter-arrival times and breach sizes, it is necessary to consider the dependence; otherwise, the prediction results are notaccurate. 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 thatthe 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 individualhacking breach incidents will not get much worse.

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 processmodels can predict the inter-arrival times and the breach sizes. Tothe 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.

4. SYSTEM ARCHITECTURE

Our project makes use of a variety of algorithms to help us achieve a precise result. The Python programming language, which is among the most often used and popular languages in AI and ML because it comes with all of the necessary tools and libraries has been used in our project. It has several libraries, like pandas for the filtering process, matplotlib for plotting the data, data visualization, and exploratory data analysis. We have also used sklearn which is Scikit-learn which includes several clustering, regression, and classification algorithms that are commonly used in AI and machine learning. Numpy is used to deal with the multidimensional array and data structures. Seaborn library has been used for data visualization. The model then applies this technique to pre- defined data set including all the information about our customers. In a l

linear pattern algorithms are executed one after the other. The data is then analyzed, segregated, and provided into the model to train it. As shown in Figure. 2. After each algorithm, the precision rate is displayed. We have trained our model with many algorithms to get a precise result. The Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, SVM, and K Neighbors algorithm will all be used, with a 70% training set and a 30% testing set. We have discovered that logic regression, decision trees, and random forests have superior precision. Following the testing procedure, the model predicts if the current candidate based on the conclusion is a good candidate for


User
click though websites
User details/ secret details upload datasets
Upload files
incident data sets
Predicting Cyber Hacking Breaches
unknown attack by hackers
breached records
Datasets (database)
Admin can be Prepare the Data to Predicting
prediction performance
ADMIN LOGIN
Modeling and prediction cyber
breaches
time and breaches analysis/prediction
calculated those records and analysis
Modified Graph Data

5.IMPLEMENTATION

1. UPLOAD DATA:

The data resource to database can be uploaded by bothadministrator and authorized user.or request for files.

2. ACCESS DETAILS:

The access of data from the database can be given by administrators.

3. USER PERMISSIONS:

The data from any resources are allowed to access the data with only permission from administrator.

4. DATA ANALYSIS:

Data analyses are done with the help of graph.

6.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 stochasticprocesses 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 periodof 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 toanalyze datasets of a similar nature.

7.SCOPE OF 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.It is wroth while to estimate the exact occuring times of breach incidents.Finally,more research needs to be conducted towards understanding the predictability of breach incidents.

8. REFERENCES

1.F.Y. Leu, J.C. Lin, M.C. Li, C.T Yang, P.C Shih, "Integrating Grid with Intrusion Detection," Proc. 19thInternational Conference on Advanced InformationNetworking and Applications, pp. 304-309, 2005.

2. White paper, "Intrusion Detection:A Survey," ch.2, DAAD19-01, NSF, 2002.

3. K. Scarfone, P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," NIST Special Publication800-94, Feb. 2007.

4. IBM Security.Accessed:Nov.207
2016 Cyber ClaimsStudy. Accessed: Nov.
201710/P02_NetDiligence-2016- Cyber-Claims-Study-ONLINE.pdf

6.M. Eling and W. Schnell, "What do we know about cyber risk and cyber riskinsurance?" J. Risk Finance, vol. 17, no. 5, pp. 474– 491, 2016.

7. https://ieeexplore.ieee.org/document/836017
2#:~:text=Modeling%20and%20Predicting%2
0Cyber%20Hacking%20Breaches%20Abstract
%3A%20Analyzing,t
opic2C%20and%20many%20studies%20remain%20to%20be%20done.

8. https://github.com/nansunsun/Cybersecurity-incident-prediction-and-discovery-data

9. https://www.datacenters.com/news/hacking-data-breaches-cyber-warfare

10. http://truevolts.com/wpcontent/uploads/2019/12/28-5.docx