MODELING AND PREDICTING CYBER HACKING BREACHES

MOHAN RAJESH[1], G.RAMESH KUMAR [2]

[1]MCA Student, B V Raju College, Kovvada, Andhra Pradesh, India.

[2]Assistant Professor, B V Raju College, Kovvada, Andhra Pradesh, India.

ABSTRACT:

Analyzing cyber incident datasets 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 dataset corresponding to 12 years (2005-2017) of cyber hacking activities that include malware attacks. We show that, in contrast to the fifindings reported in the literature, both the hacking breach incident *inter-arrival times* and the *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 fifit 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 dataset. We draw a set of cybersecurity 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 —Hacking breach, data breach, cyber threats, cyber risk analysis, breach prediction, trend analysis, time series, cybersecurity data analytics.

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 [2] reports 1,093 data breach incidents in 2016, which is 40% higher than the 780 data breach incidents in 2015. The United States Offifice of Personnel Management (OPM) [3] 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 [4]

reports that in year 2016, the global average cost for each lost or stolen record containing sensitive or confifidential information was $158. NetDiligence [5] 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 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

M. Xu is with Department of Mathematics, Illinois State University, USA. K. Schweitzer and R. Bateman are with U.S. Army Research Laboratory South. S. Xu is with Department of Computer Science, University of Texas at San Antonio. Correspondence: shxu@cs.utsa.edu 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) [6].

Recently, researchers started modeling data breach incidents. Maillart and Sornette [7] studied the statistical properties of the personal identity losses in the United States between year 2000 and 2008 [8]. They found that the number of breach incidents dramatically increases from 2000 to July 2006 but remains stable thereafter. Edwards et al. [9] analyzed a dataset containing 2,253 breach incidents that span over a decade (2005 to 2015) [1]. They found that neither the size nor the frequency of data breaches has increased over the years.

Wheatley et al. [10] analyzed a dataset that is combined from [8] and [1] and 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 fifirms is independent of time, but the frequency of large breach incidents occurring to non

US fifirms exhibits an increasing trend.

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.

LITERATURE SURVEY

Prior works closely related to the present study. Maillart and Sornette [7] analyzed a dataset [8] of 956 personal identity loss incidents that occurred in the United States between year 2000 and 2008. They found that the personal identity losses per incident, denoted by X, can be modeled by a heavy tail distribution $\Pr(X > n) \sim n^{-\alpha}$ where $\alpha = 0.7 \pm 0.1$. This result remains valid when dividing the dataset per type of organizations: business, education, government, and medical institution. Because the probability density function of the identity losses per incident is static, the situation of identity loss is stable from the point of view of the breach size. Edwards et al. [9] analyzed a different breach dataset [1] of 2,253 breach incidents that span over a decade (2005 to 2015). These breach incidents include two categories: *negligent breaches* (i.e., incidents caused by lost, discarded, stolen devices, or other reasons) and *malicious breaching* (i.e., incidents caused by hacking, insider and other reasons). They showed that the breach size can be modeled by the log-normal or log-skewnormal distribution and the breach frequency can be modeled by the negative binomial distribution, implying that neither the breach size nor the breach frequency has increased over the years.

Wheatley et al. [10] analyzed an organizational breach incidents dataset that is combined from [8] and [1] and spans over a decade (year 2000 to 2015). They used the Extreme Value Theory

[11] to study the maximum breach size, and further modeled the large breach sizes by a doubly truncated Pareto distribution. They also used linear regression
to study the frequency of the data breaches, and found that the
frequency of large breaching incidents is independent of time for the United States organizations, but shows an increasing trend for non-US organizations. There are also studies on the dependence among cyber risks.

B ohme and Kataria [12] studied the dependence between cyber risks of two levels: within a company (internal dependence) and across companies (global dependence). Herath and Herath
[13] used the Archimedean copula to model cyber risks caused by virus incidents, and found that there exists some dependence between these risks. Mukhopadhyay et al. [14] used a copula-based Bayesian Belief Network to assess cyber vulnerability. Xu and Hua [15] investigated using copulas to model dependent cyber risks. Xu et al. [16] used copulas to investigate the dependence encountered when modeling the effectiveness of cyber defense early-warning. Peng et al. [17] investigated multivariate cybersecurity risks with dependence. Compared with all these studies mentioned above, the present paper is unique in that it uses a new methodology to analyze a new perspective of breach incidents (i.e., cyber hacking breach incidents). This perspective is important because it reflects the consequence of cyber hacking (including malware).

The new methodology found for the fifirst time, that both the incidents inter-arrival times and the breach sizes should be modeled by stochastic processes rather than distributions, and that there exists a positive dependence between them.

Other prior works related to the present study. Eling and Loperfifido [18] analyzed a dataset [1] from the point of view of actuarial modeling and pricing. Bagchi and Udo [19] used
a variant of the Gompertz model to analyze the growth of computer and Internet-related crimes. Condon et. al [20] used the ARIMA model to predict security incidents based on a dataset provided by the Offifice of Information Technology at the University of Maryland. Zhan et al. [21] analyzed the posture of cyber threats by using a dataset collected at a network telescope. Using datasets collected at a honeypot, Zhan et al. [22], [23] exploited their statistical properties including long
range dependence and extreme values to describe and predict the number of

attacks against the honeypot; a predictability evaluation of a related dataset is described in [24]. Peng et al. [25] used a marked point process to predict extreme
attack rates. Bakdash et al. [26] extended these studies into
related cybersecurity scenarios. Liu et al. [27] investigated
how to use externally observable features of a network (e.g.,mismanagement symptoms) to forecast the potential of data breach incidents to that network. Sen and Borle [28] studied the factors that could increase or decrease the contextual risk f data breaches, by using tools that include the opportunity theory of crime, the institutional anomie theory, and the institutional theory.

C. Paper outline

The rest of the paper is organized as follows. In Section II we describe the dataset and research questions. In Section III we present a basic analysis of the dataset. In Section IV we develop a novel point process model for analyzing the dataset. In Section V, we discuss the prediction performance of the proposed model. In Section VI we present qualitative and quantitative trend analyses. In Section VII we conclude our paper with future research directions. We defer formal description of the main statistical notions to the Appendix, and discuss their intuitive meanings when they are mentioned for the first time.

SYSTEM ANALYSIS AND DESIGN

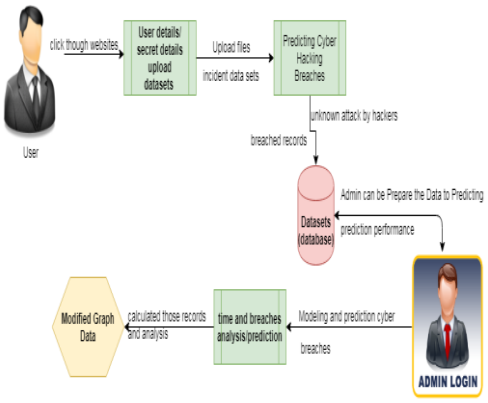


Fig :Architecture

The life cycle of a software application is defined as the time it takes to design, test, and then implement the software.

During the first round of research, we refine the design.

Initiation of the Application Process

For software applications, there is an initial evaluation and a comprehensive evaluation. The Expert performs a preliminary analysis to determine what is needed and whether there are any cost advantages to be had. Research studies that include all of the relevant variables aid in the development and expansion of the programmed.

It's called the Criterion System (SRS).

Software Application Requirement Specification is a document that completely describes what the recommended should do, but does not specify exactly how the software application programmed accomplishes this task. "

The Requirement for Performance...

1) High throughput and a fast procedure time are required.
2) The outcome should be both quick and exact.

Top Qualities of a Great Software Application.

This application is very easy to maintain because it is directly linked to the data source.

If you've ever had trouble finding a new app due to the fact that there are simply way too many options, this collection is for you.

For future enhancements, this programmed is quite flexible.

Requirements in terms of the available technology.

Demands placed on the software.

In the software application's

The second step in the life cycle of a computer system is its style, which is the fundamental layout. The system's capabilities are still being established and tested. The first step is to generate a list of programmed requirements. This document outlines the information inputs, circulation, and outcomes generation processes.

There is a shift from individual-oriented data to system data throughout the layout phase. Physical procedures, equipment, and computer programmers are all given responsibilities at the layout phase. In the initial stages of research, flow diagrams are generated and then dissolved until all system facets appear to be working properly

Data organization, software development (such as formulas), and partnerships between various components of a system are all part of design as a multi-step process. The formation of a style is a multifaceted process that includes both logical and physical components. Using reviews, linkages are made between the current system and the needs that have been gathered. The physical plan specifies the software and hardware required to meet the needs of the local layout.

Modularization has taken place at this point in time. The quality of each component's preparation is critical to the

overall success of the integrated system. Step-by-step alterations in a work are the norm when it comes to altering it. Such an ingredient must be administered throughout the interphone as well. The design approach is constantly evolving as new techniques, improved evaluation, and also a greater understanding of software application design are developed.

A wide variety of software format approaches exist, each with its own set of design quality standards. – Software programmed style have three technological tasks: design, coding, and testing.

As the software's requirements vary, so does its integrity. The format system transforms the academic solution of the expediency study into a real-world reality.

The ability to put things together in a variety of ways.

Representations for the Object-Oriented Modeling Language.

The acronym for the Unified Modeling Language (UML) is UML. Rational Software Application Corporation, James Rumbaing, and Invar Jacobson were all involved in the development of this object-oriented symbols system. This group of eminent computer experts designed a complex technology mix. To model object-oriented software, the Things Management Group (OMG) has established UML as a necessity.

They are classified into three categories:

Planned routines for the average person. This is a visual representation of the way systems and processes behave in the real world. Also included are representations of the user's job, status devices and use contexts.

Illustrations illustrating the exchange of information. The relationships between objects are the focus of this type of habit arrangement. This area includes interactive, series, and temporal representations.

Structure diagrams, if you will. A non-time-bound visual representation of a set of requirements. Classes, composite structures, and product layouts are all covered in this group.

3 .Problem Statement

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.

4 .ALGORITHM:

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.

5. MODULES:

ADMIN.

In this module, the administrator will be able to predict malware in this application.

There is a visual examination of the user projections and assessments after a client logs in.

This module provides the client with all of the information they need, including details, evaluation, malware data, enamelware data, branched evaluation, and aesthetic analysis.

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.

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.

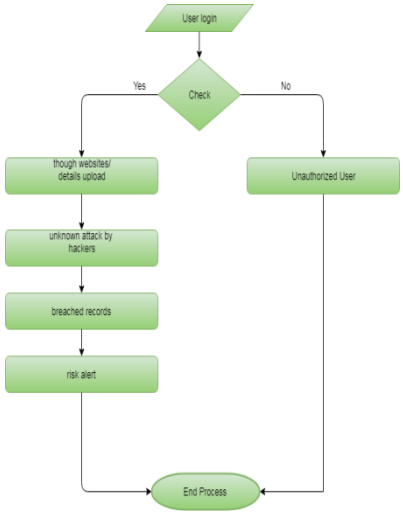


Fig: Flow Chart

6 .METHODOLOGY

Specifying, designing, and coding are the three stages of the software development process that are thoroughly evaluated during software screenings. A closer look reveals an intriguing anomaly in the software's source code. Getting the software from a conceptual design to a working prototype was a primary objective early on in development.

During the screening phase, the generated system is tested on a variety of data sets. System screening relies heavily on how test results are shown. When testing this system, the results of the tests were used to determine how well it performed. Test data was used to identify and correct systemic errors. Tests were required before the planned system could be put into operation.... A variety of techniques are used in testing, including:

The system must be fixed if it isn't working properly.

Interoperability evaluations

Individual System Acceptance Testing Validation

The early stages of device testing

In software development, testing is concentrated on the module because it is the smallest system. A series of tests has been written here by the programmer in preparation for the system's eventual integration into a larger system. Coders carried out a preliminary check. In order to make certain that each module is on track to accomplish its objectives, we perform a thorough evaluation on each individual one.

Using a multi-pronged approach, we make certain that none of the types are infected. The following is a list of the test cases that were run.

Users will be prompted to input a value if the use rid and password fields are left blank in the login form.

Participants will see an error notice if they attempt to log in using an invalid use rid or password. "The login credentials you entered are incorrect. Do it again if the first attempt fails "There are so many questions that I have no idea where to begin...

Every field on the book's entry page and the new student/teacher screen requires a value to be input. Error notifications could instead be sent through customer care.

A member id, book number, concern day, and return date are necessary for publishing transactions. If the fields are left blank, the user will see the error message "Fields need not to be blank."

Verification/Checking of Combinations

Creating a program's structure and testing for faults in the user interface are both referred to as "integration screening." As a result of extensive testing, all components of a predetermined programmed structure will be included. This action is the culmination of everything that has come before it. The entire programmed has been put through its paces. Users may find the interfaces difficult to navigate. There are bound to be a slew of errors in this situation.

There are two ways to perform a combination screening:

Assimilation in the reverse direction

A bottom-up approach to integration

In order to build a more sophisticated system, the first step is to downgrade huge modules to smaller ones. As opposed to this, overhead assimilation utilizes a mechanism in which smaller bits are mixed with larger ones. Here's an example of a combo that's been turned upside down. It was tough to adapt because there were so many variables. In spite of this, each error was fixed and then passed on to the next phase of testing.

It has been thoroughly tested to ensure that the system's user may easily switch between different screens...

The database and forms have been thoroughly tested to guarantee that they work together seamlessly. The user will be notified if there is a problem with the gadget.

End-user acceptance testing (AET) (UAT)

A system must have the support of its users before it can be considered a success. With the person who would be using it on a regular basis, the system was thoroughly tested to ensure that it would meet the needs of its target audience. This is the case as a result of the following factors:

It is possible to define a system as a set of separate instructions...

In order to move forward, the application system's technical requirements must be specified in great detail before any work can begin. End-user supervisors from several departments were consulted before finalizing the system specifications.

Users' requirements are taken into account during the development of the product. The current situation and the intended outcome of putting in place the new information system

Based on the issue at hand, and how thoroughly an inquiry is conducted, a system's performance is determined. Determining a client's needs, rather than their desires, through in-depth analysis An organized system lays out exactly how and when each task or activity should be accomplished and who is responsible for it.

were put through its paces.

System screening is the process of verifying that all of the software/system components match the requirements given in the software specification (SRS).

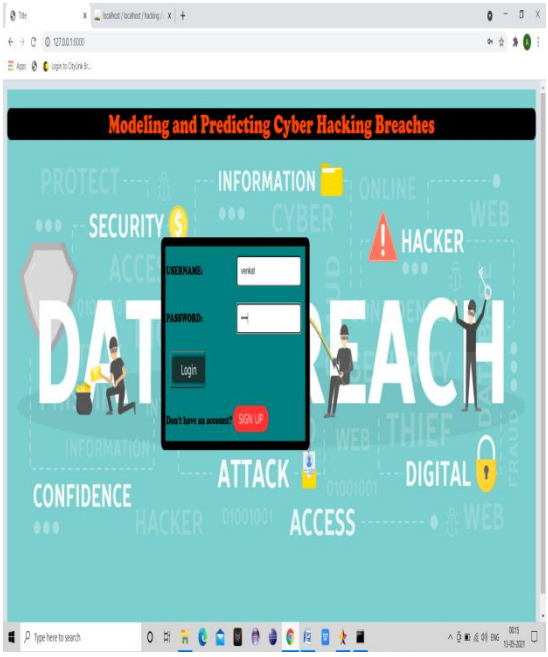
After the system has gone through integration testing, a last round of testing is conducted. Non-functional requirements should also be tested as part of a system's development. Methods of system screening differ greatly from one enterprise to the next.

The software and hardware requirements for this project have been met and tested. Specifications for the hardware and software were strictly adhered to.

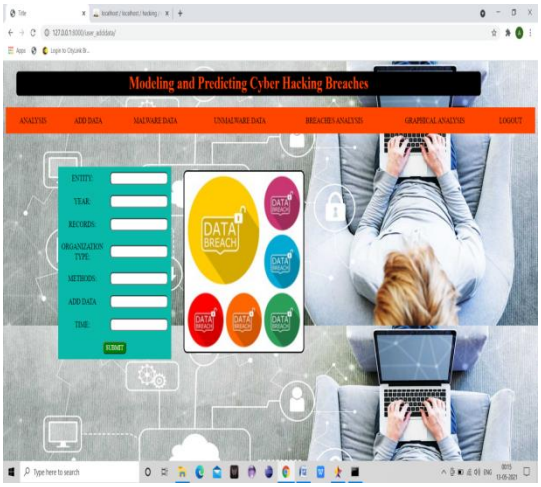
7 .RESULT

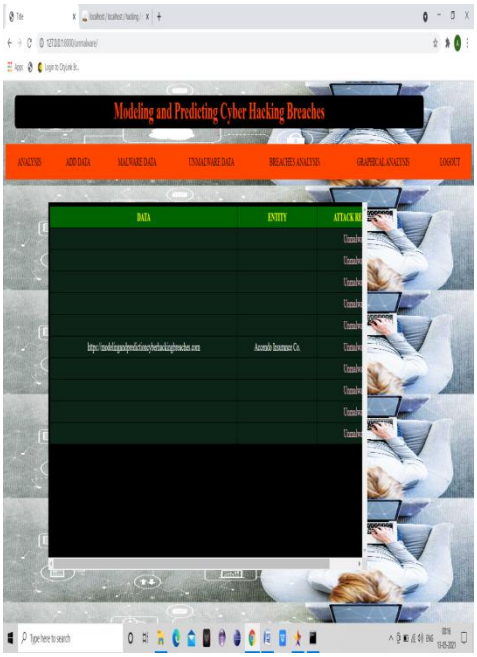
Home page

Login

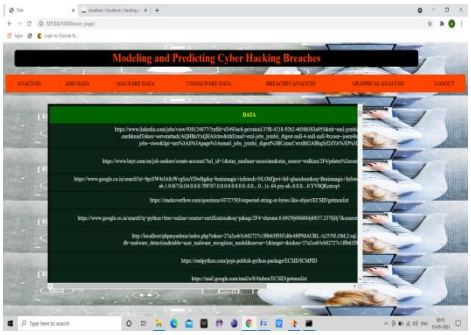


User Home page

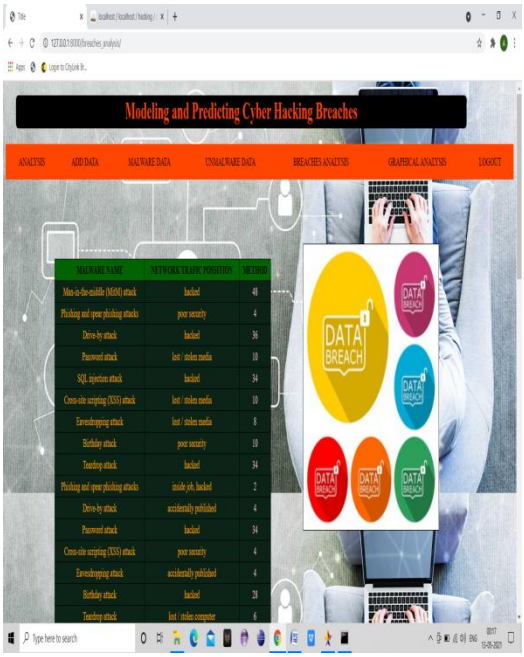




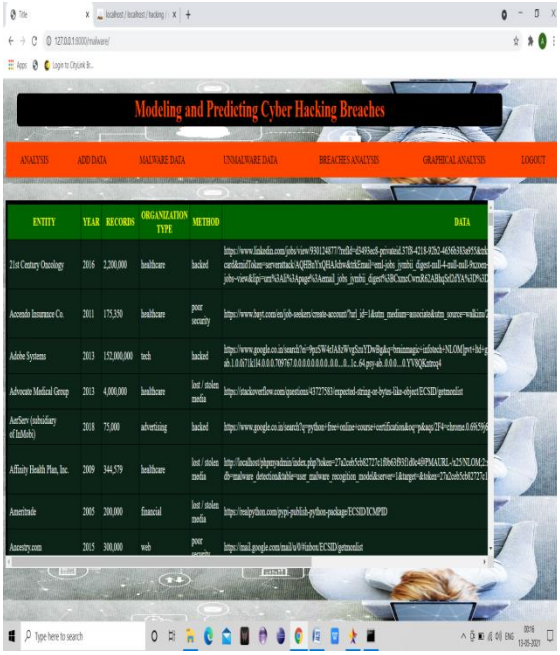
Analysis



Breaches Analysis

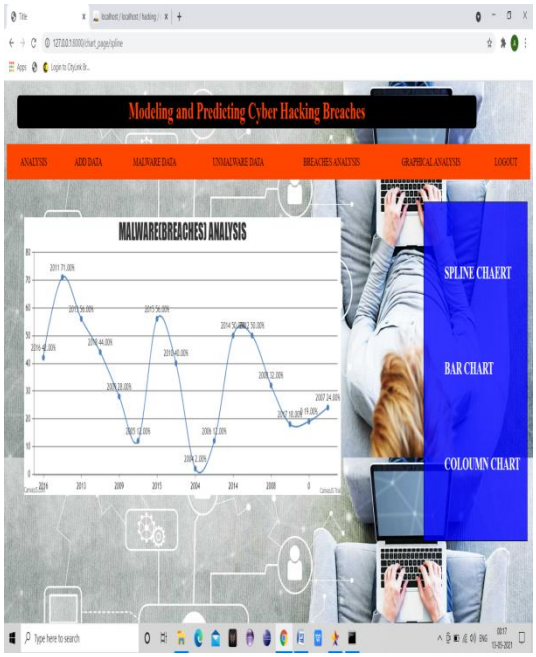


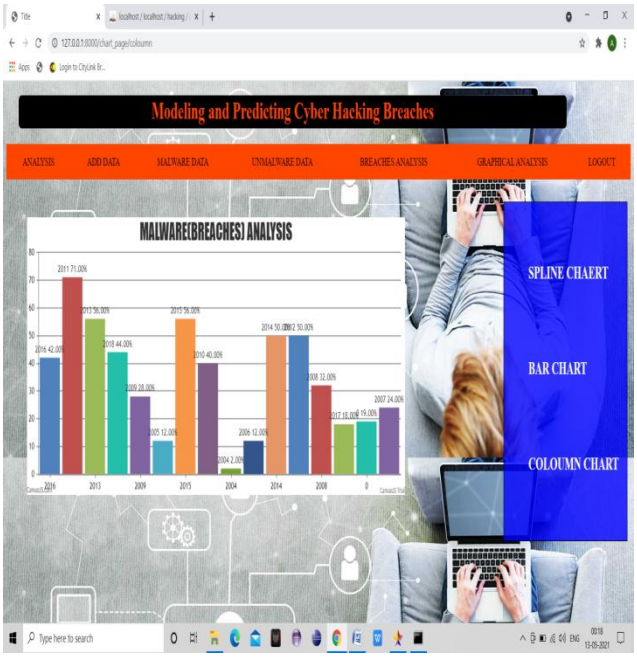
Malware Data



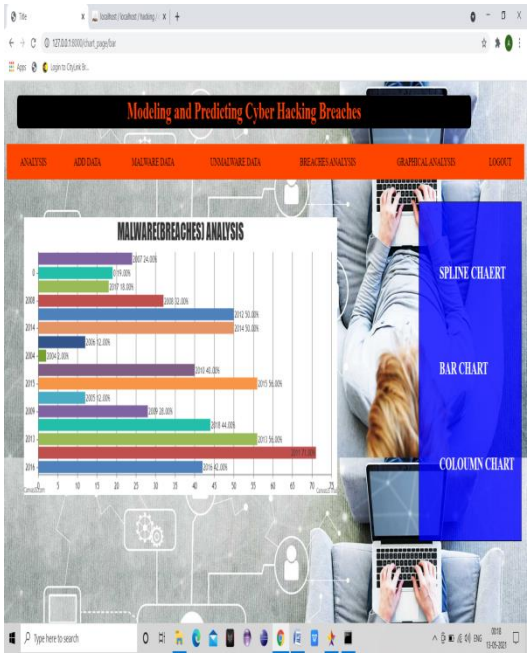
Graphical Analysis

Unaware Data



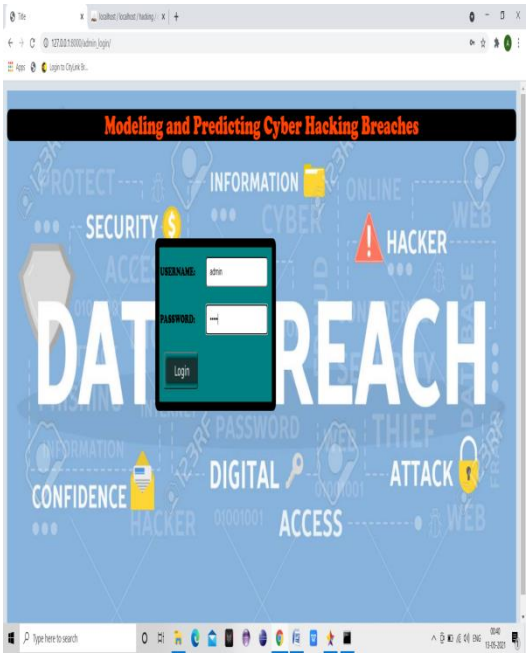


Bar chart



Column chart

Admin Login



User details Analysis

Admin Analysis

4.5 Test Cases

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 fifitting 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 interarrival 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.

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 [24]).

Acknowledgement. We thank the reviewers for their construc
tive comments that helped improve the paper. In Section V,
we incorporated some insightful comments of one reviewer on
how to connect the prediction models to real-world cyber de
fense quantitative risk management. This work was supported
in part by ARL grant #W911NF-17-2-0127.

FUTURE SCOPE:

There are still many unanswered questions. Examples of this can be found in the research of predicting extremely large numbers and handling missing data, for example (i.e., violation cases that are not reported). It's also a good idea to figure out the exact time of the breach events. A further investigation is needed to establish the probability of a security breach taking place (i.e., the upper bound of prediction precision).

9 .REFERENCES

[1] P. R. Clearinghouse, “Privacy rights clearinghouse’s chronology of data breaches.”https://www.privacyrights.org /data-breaches, Last accessed on November 9, 2017.

[2] I. T. R. Center, “http://www.idtheftcenter.org/2016datab reaches.html.”

[3] C. R. Center, “Cybersecurity incidents.” https://www.opm.gov/ cybersecurity/cybersecurity-incidents/, Last accessed on November 9,
2017.

[4] I. Security, “https://www.ibm.com/security/data-breach/index.html.”

[5] N. . C. C. Study, “https://netdiligence.com/wp-content/uploads/2016/10/
P02 NetDiligence-2016-Cyber-Claims-Study-ONLINE.pdf.”

[6] M. Eling and W. Schnell, “What do we know about cyber risk and cyber risk insurance?,” *The Journal of Risk Finance*, vol. 17, no. 5, pp. 474–
491, 2016.

[7] T. Maillart and D. Sornette, "Heavy-tailed distribution of cyber-risks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 75, no. 3, pp. 357–364, 2010.

[8] R. B. Security, "Datalossdb." https://blog.datalossdb.org/, Last accessed on November 9, 2017.

[9] B. Edwards, S. Hofmeyr, and S. Forrest, "Hype and heavy tails: A closer look at data breaches," *Journal of Cybersecurity*, 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," *The European Physical Journal B*, vol. 89, no. 1, p. 7, 2016.

[11] P. Embrechts, C. Kl ¨uppelberg, and T. Mikosch, *Modelling extremal events: for insurance and fifinance*, vol. 33. Springer Science & Business Media, 2013.

[12] R. B ¨ ohme and G. Kataria, "Models and measures for correlation in cyber-insurance.," in *Workshop on the Economics of Information Security (WEIS)*, 2006.

[13] H. Herath and T. Herath, "Copula-based actuarial model for pricing cyber-insurance policies," *Insurance Markets and Companies: Analyses and Actuarial Computations*, vol. 2, no. 1, pp. 7–20, 2011.

[14] A. Mukhopadhyay, S. Chatterjee, D. Saha, A. Mahanti, and S. K. Sadhukhan, "Cyber-risk decision models: To insure it or not?," *Decision Support Systems*, vol. 56, pp. 11–26, 2013.