FORECASTING CYBER HACKING BREACHES: MODELING AND PREDICTION

[1] M.Suresh, Assistant Professor, Department of CSE, Chalapathi Institute of Technology, Guntur.
[2] Badisha Bhargav, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.
[3] Chimata Pavan Kumar, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.
[4] Chepuri Aravind, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.
[5] Gattu Revanth Akhil, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.

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.

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 Office 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 confidential 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 prob lem. 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. 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 inci dents. Maillart and Sornette [7] studied the statistical prop erties 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 firms is independent of time, but the frequency of large breach incidents occurring to non-US firms 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.

2. LITERATURE SURVEY

1) Prior Works Closely Related to the Present Study: Maillart and Sornette [7] analyzed a dataset [8] of 956 per sonal 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) ∼ n −α where α = 0.7±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-skew normal 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 inci dents 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. Böhme and Kataria [12] studied the dependence between cyber risks of two levels: within a company (internal depen dence) 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 hack ing breach incidents). This perspective is important because it reflects the consequence of cyber hacking (including mal ware). The new methodology found for the first time, that both the incidents inter-arrival times and the breach sizes should be modeled by stochastic processes rather than distri butions, and that there exists a positive dependence between them. 2) Other Prior Works Related to the Present Study: Eling and Loperfido [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 Office 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. Bakdashetial. [26] extended these studies into related cyber security 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 of data breaches, by using tools that include the opportunity theory of crime, the institutional anomie theory, and the institutional theory

3. EXISTING SYSTEM

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.

4. PROPOSED SYSTEM

In this paper, we make the following three contributions. First, we show that both the hacking breach incident inter arrival 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.

SYSTEM ARCHITECTURE

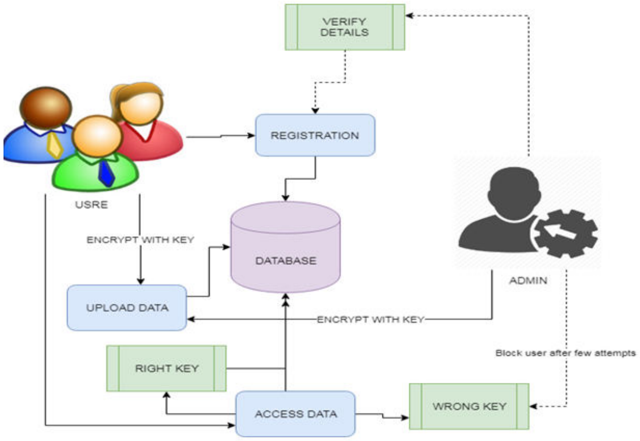


Fig 1: System Architecture

5. UML DIAGRAMS

1. CLASS DIAGRAM

The cornerstone of event-driven data exploration is the class outline. Both broad practical verification of the application's precision and fine-grained demonstration of the model translation into software code rely on its availability. Class graphs are another data visualisation option.

The core components, application involvement, and class changes are all represented by comparable classes in the class diagram. Classes with three-participant boxes are referred to be "incorporated into the framework," and each class has three different locations:

- The techniques or actions that the class may use or reject are depicted at the bottom.

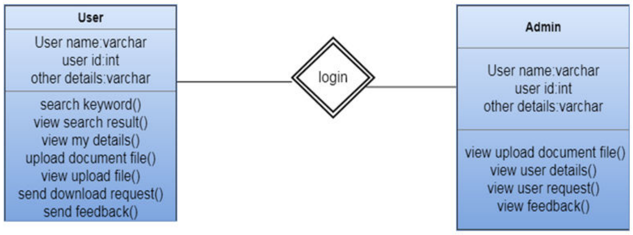


Fig 5.1 shows the class diagram of the project

2. USECASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

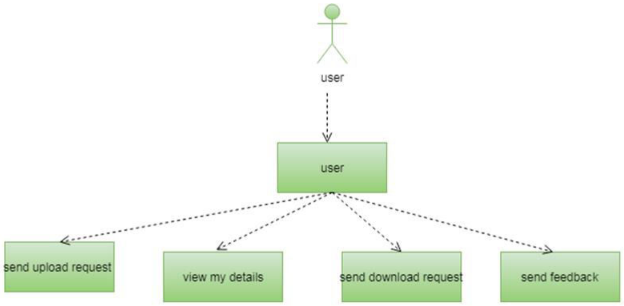


Fig 5.2 Shows the Use case Diagram for User

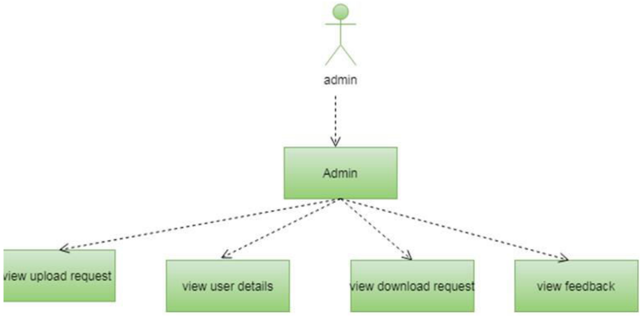


Fig 5.3 Use case Diagram for Admin

3. SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

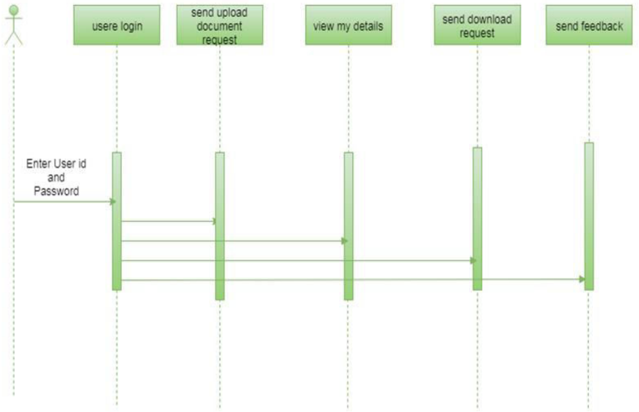


Fig 5.4 Shows the Sequence Diagram

6. RESULTS

7.1 Output Screens

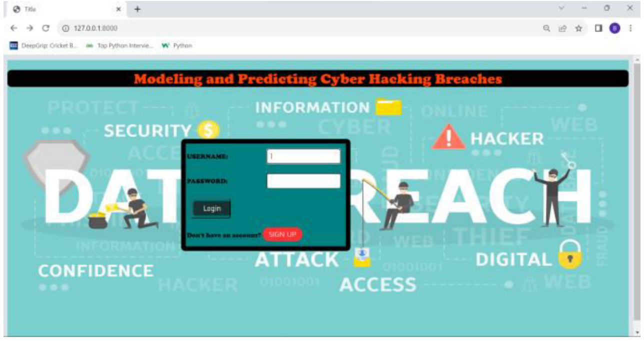


Fig 6.1 User Login

In above screen we enter the user name and password for user login.

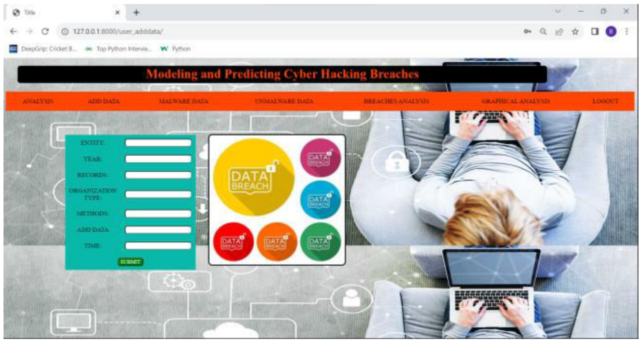
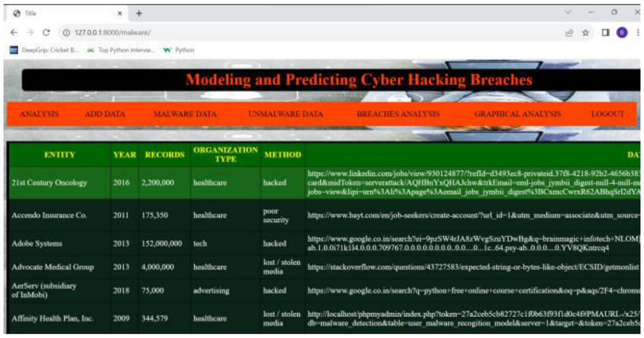


Fig 6.2 Details of Data Breaches

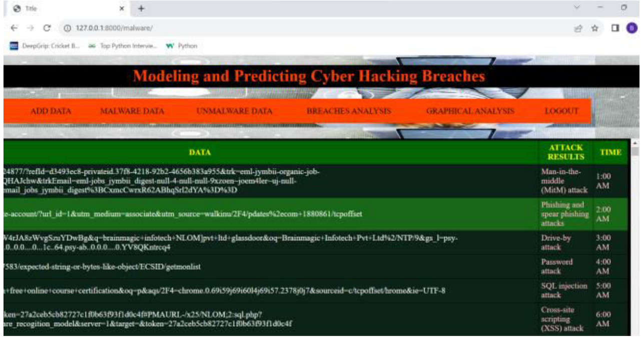
In above screen we can enter the details about the data breaches.



ENTITY	YEAR	RECORDS	ORGANIZATION TYPE	METHOD	DA
21st Century Oncology	2016	2,200,000	healthcare	hacked	https://www.linkedin.com/jobs/view/930124877/?refId=d3493ec8-privateid.37f8-4218-92b2-4656b383 card&midToken=serverattack/AQHBnYxQHAJchw&trkEmail=eml-jobs_jymbii_digest-null-4-null-nu jobs~view&lipi=urn%3Ali%3Apage%3Aemail_jobs_jymbii_digest%3BCxmcCwrxR62ABhqSrl2dYA
Accendo Insurance Co.	2011	175,350	healthcare	poor security	https://www.bayt.com/en/job-seekers/create-account/?url_id=1&utm_medium=associate&utm_source=
Adobe Systems	2013	152,000,000	tech	hacked	https://www.google.co.in/search?ei=9pzSW4rJA8zWvgSzuYDwBg&q=brainmagic+infotech+NLOM] ab.1.0.0i71k1l4.0.0.0.709767.0.0.0.0.0.0.0.0.0..0.0....0...1c..64.psy-ab..0.0.0....0.YV8QKntrcq4
Advocate Medical Group	2013	4,000,000	healthcare	lost / stolen media	https://stackoverflow.com/questions/43727583/expected-string-or-bytes-like-object/ECSID/getmonlist
AerServ (subsidiary of InMobi)	2018	75,000	advertising	hacked	https://www.google.co.in/search?q=python+free+online+course+certification&oq=p&aqs/2F4~chrome
Affinity Health Plan, Inc.	2009	344,579	healthcare	lost / stolen media	http://localhost/phpmyadmin/index.php?token=27a2ceb5cb82727c1f0b63f93f1d0c4f#PMAURL-/x25/ db=malware_detection&table=user_malware_recogition_model&server=1&target=&token=27a2ceb5c

Fig 6.3 Malware and Un malware Data

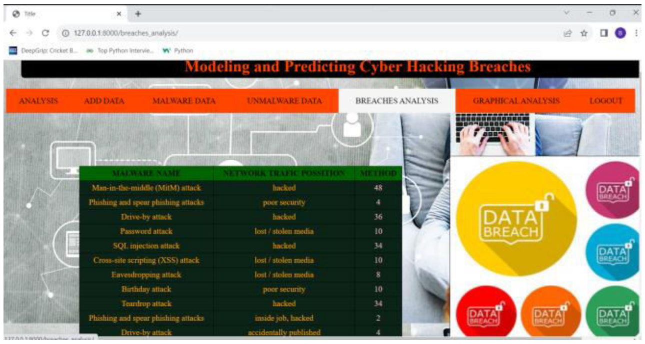
In above screen shows the malware and un malware data.



DATA	ATTACK RESULTS	TIME
24877/?refId=d3493ec8-privateid.37f8-4218-92b2-4656b383a955&trk=eml-jymbii-organic-job- QHAJchw&trkEmail=eml-jobs_jymbii_digest-null-4-null-null-9xzoen~joem4ler~uj-null- email_jobs_jymbii_digest%3BCxmcCwrxR62ABhqSrl2dYA%3D%3D	Man-in-the-middle (MitM) attack	1:00 AM
e-account/?url_id=1&utm_medium=associate&utm_source=walkinu/2F4/pdates%2ecom+1880861/tcpoffset	Phishing and spear phishing attacks	2:00 AM
W4rJA8zWvgSzuYDwBg&q=brainmagic+infotech+NLOM]pvt+ltd+glassdoor&oq=Brainmagic+Infotech+Pvt+Ltd%2/NTP/9&gs_l=psy- .0..0.0....0...1c..64.psy-ab..0.0.0....0.YV8QKntrcq4	Drive-by attack	3:00 AM
7583/expected-string-or-bytes-like-object/ECSID/getmonlist	Password attack	4:00 AM
+free+online+course+certification&oq=p&aqs/2F4~chrome.0.69i59j69i60l4j69i57.2378j0j7&sourceid=c/tcpoffset/hrome&ie=UTF-8	SQL injection attack	5:00 AM
ken=27a2ceb5cb82727c1f0b63f93f1d0c4f#PMAURL-/x25/NLOM;2:sql.php? re_recogition_model&server=1&target=&token=27a2ceb5cb82727c1f0b63f93f1d0c4f	Cross-site scripting (XSS) attack	6:00 AM

Fig 6.4 Attacks Results

In above screen shows the attacks results.



MALWARE NAME	NETWORK TRAFIC POSSITION	METHOD
Man-in-the-middle (MitM) attack	hacked	48
Phishing and spear phishing attacks	poor security	4
Drive-by attack	hacked	36
Password attack	lost / stolen media	10
SQL injection attack	hacked	34
Cross-site scripting (XSS) attack	lost / stolen media	10
Eavesdropping attack	lost / stolen media	8
Birthday attack	poor security	10
Teardrop attack	hacked	34
Phishing and spear phishing attacks	inside job, hacked	2
Drive-by attack	accidentally published	4

Fig 6.5 Breaches Analysis

In above screen shows the result of breaches analysis.

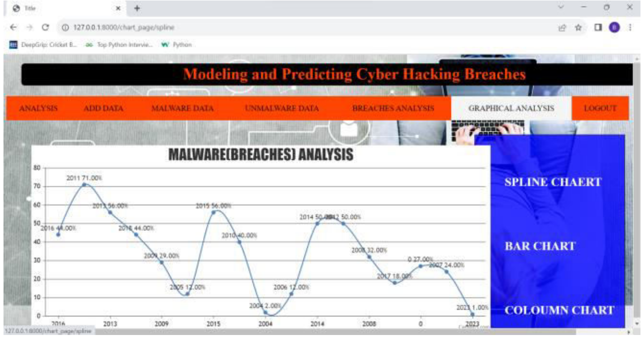
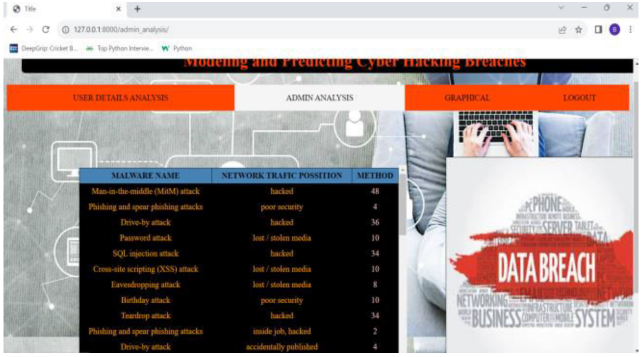


Fig 6.6 Graphical Analysis

In above screen shows the breaches results in graph representation.



MALWARE NAME	NETWORK TRAFIC POSSITION	METHOD
Man-in-the-middle (MitM) attack	hacked	48
Phishing and spear phishing attacks	poor security	4
Drive-by attack	hacked	36
Password attack	lost / stolen media	10
SQL injection attack	hacked	34
Cross-site scripting (XSS) attack	lost / stolen media	10
Eavesdropping attack	lost / stolen media	8
Birthday attack	poor security	10
Teardrop attack	hacked	34
Phishing and spear phishing attacks	inside job, hacked	2
Drive-by attack	accidentally published	4

Fig 6.7 Admin Analysis

In above screen shows the admin analysis report.

7. 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 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. 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 to analyze datasets of a similar nature.

8. REFERENCES

[1] P. R. Clearinghouse. Privacy Rights Clearinghouse's Chronol ogy of 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: http://www.idtheftcenter.org/2016databreaches.html

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

[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] M. Eling and W. Schnell, "What do we know about cyber risk and cyber risk 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: https://blog.datalossdb.org

[9] B. Edwards, S. Hofmeyr, and S. Forrest, "Hype and heavy tails: A closer look 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 Extremal Events: For Insurance and Finance, vol. 33. Berlin, Germany: Springer-Verlag, 2013.

[12] R. Böhme and G. Kataria, "Models and measures for correlation in cyber-insurance," in Proc. Workshop Econ. Inf. Secur. (WEIS), 2006, pp. 1–26.

[13] H. Herath and T. Herath, "Copula-based actuarial model for pricing cyber-insurance policies," Insurance Markets Companies: Anal. Actuar ial Comput., 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 Syst., vol. 56, pp. 11–26, Dec. 2013.

[15] M. Xu and L. Hua. (2017). Cybersecurity Insurance: Modeling and Pricing. [Online]. Available: https://www.soa.org/research-reports/20 17/cybersecurity-insurance

[16] M. Xu, L. Hua, and S. Xu, "A vine copula model for predicting the effectiveness 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 cybersecurity risks," 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 security incident data using time series models," in Proc. 19th Int. Symp. Softw. Rel. Eng. (ISSRE), Nov. 2008, pp. 77–86.