JES
Journal of Engineering Sciences

Vol 13, Issue 06, June/2022
ISSN NO:0377-9254

# A NOVEL METHOD FOR DETECTING CYBER BREACH

[1]Ms.P.Priyanka [2]Thuvva Deepika [3]Thiruvidi Jahnavi [3]Varjala Rohitha

[4]Palle Venkata Haritha

[1]Assistant professor [2,3,4,5] U.G. SCHOLAR

Department of CSE

RAVINDRA COLLEGE OF ENGINEERING FOR WOMEN

## 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 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.

## I. 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] eports 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 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 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

A data breach is the intentional or unintentional release of secure or private/confidential information to an untrusted environment. Other terms for this phenomenon include unintentional information disclosure, data leak, information leakage and also data spill. Incidents range from concerted attacks by black hats, or individuals who hack for some kind of personal gain, associated with organized crime, political activist or national governments to careless disposal of used computer equipment or data storage media and unhackable source.

Definition: "A data breach is a security violation in which sensitive, protected or confidential data is copied, transmitted, viewed, stolen or used by an individual unauthorized to do so."[1] Data breaches may involve financial information such as credit card & debit card details, bank details, personal health information (PHI), Personally identifiable information (PII), trade secrets of

corporations or intellectual property. Most data breaches involve overexposed and vulnerable unstructured data – files, documents, and sensitive information.[2] Data breaches can be quite costly to organizations with direct costs (remediation, investigation, etc) and indirect costs (reputational damages, providing cyber security to victims of compromised data, etc.)

According to the nonprofit consumer organization Privacy Rights Clearinghouse, a total of 227,052,199 individual records containing sensitive personal information were involved in security breaches in the United States between January 2005 and May 2008, excluding incidents where sensitive data was apparently not actually exposed.[3]

Many jurisdictions have passed data breach notification laws, which requires a company that has been subject to a data breach to inform customers and takes other steps to remediate possible injuries. A security breach is any incident that results in unauthorized access to computer data, applications, networks or devices. It results in information being accessed without authorization. Typically, it occurs when an intruder is able to bypass security mechanisms.

Technically, there's a distinction between a security breach and a data breach. A security breach is effectively a break-in, whereas a data breach is defined as the cybercriminal getting away with information. Imagine a burglar; the security breach is when he climbs through the window, and the data breach is when he grabs your pocketbook or laptop and takes it away.

Confidential information has immense value. It's often sold on the dark web; for example, names and credit card numbers can be bought, and then used for the purposes of identity theft or fraud. It's not surprising that security breaches can cost companies huge amounts of money. On average, the bill is nearly $4m for major corporations.

It's also important to distinguish the security breach definition from the definition of a security incident. An incident might involve a malware infection, DDOS attack or an employee leaving a laptop in a taxi, but if they don't result in access to the network or loss of data, they would not count as a security breach.

## II. OBJECTIVE OF THE PROJECT

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.

## III. 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.Recently, researchers started modeling data breach incidents. Maillart and Sornettestudied 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. Edwards et al. analyzed a dataset containing 2,253 breach incidents that span over a decade (2005 to 2015). They found that neither the size nor the frequency of data breaches has increased over the years. Wheatley et al., analyzed a dataset that is combined from 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.

## IV. 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. 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.

## V. Modules

**ADMIN**

In this module Admin will predict malware in this application.

After login admin can view user data prediction, admin analysis and graphical analysis

**USER**

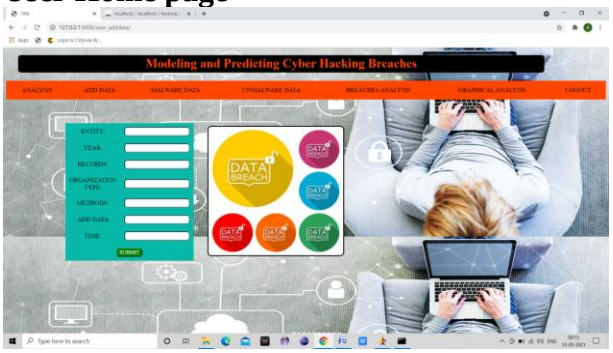
In this module user will data ,Analysis ,malware data, unmalware data ,branched analysis and graphical Analysis

## VI. RESULTS

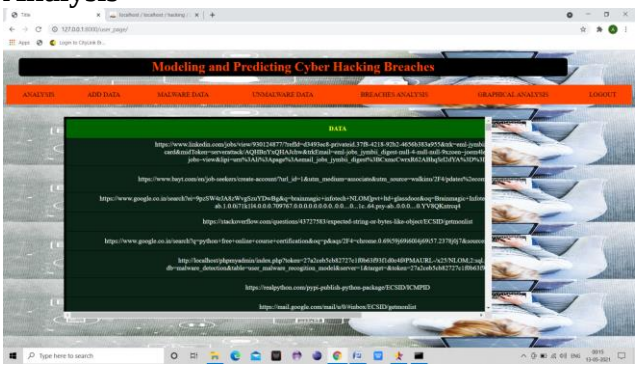
**Home page**


**Login**


**User Home page**

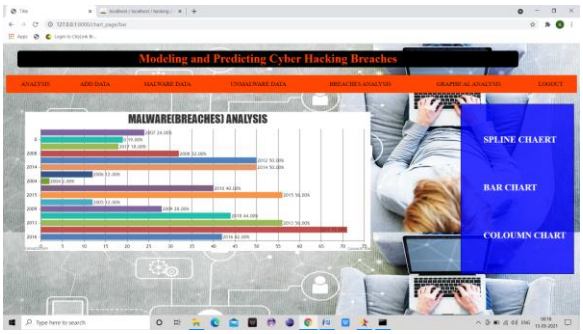



**Analysis**

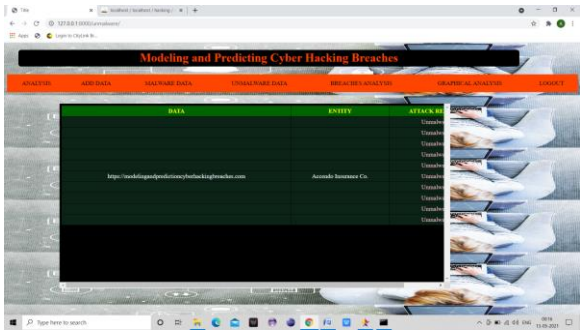



**Malware Data**


Unmalware Data




Coloumn chart



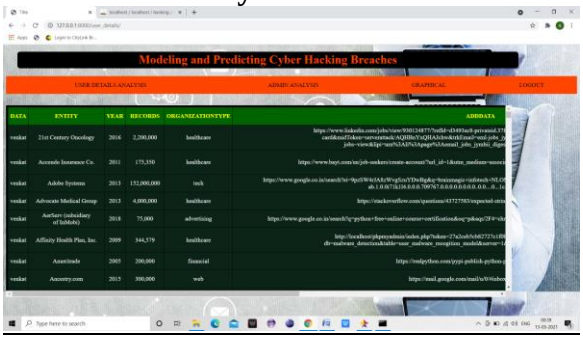

Breaches Analysis


Admin Login

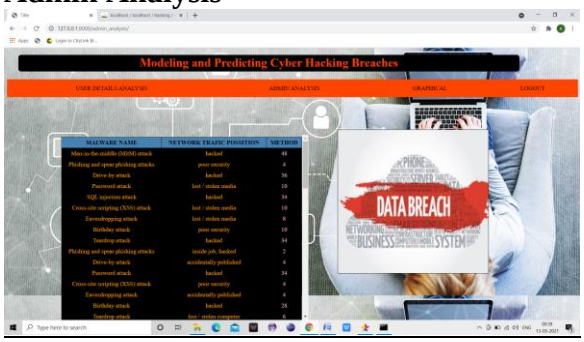

Graphical Analysis


User details Analysis


Bar chart



**Admin Analysis**




## VII. 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.

## VIII. FUTUREWORK

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).

## IX. 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] 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. Actuarial 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/ 2017/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.