ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Cyber Attack Modeling and Prediction

[1] M. Pradeepthi, [2] T. Naga Lakshmi

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract

We can learn more about how the threat landscape has changed over time by analyzing data sets of cyber incidents. There is still a lot of ground to cover as this is a new area of study. We provide the results of a statistical study of a data collection of breach incidents covering twelve years (2005–2017) of malware assaults and other forms of cyber hacking. We demonstrate that, contrary to what has been said in the literature, the inter-arrival times of hacking breach incidents and the magnitude of breaches should be represented by stochastic processes instead of distributions due to the fact that they display autocorrelations. Then, for the inter-arrival periods and the breach sizes, we provide specific stochastic process models. We also demonstrate that these models are capable of forecasting the magnitude of breaches and the intervals between arrivals. We use qualitative and quantitative trend analysis to the dataset to get a better understanding of how hacker breach instances have developed over time. An important conclusion among our cybersecurity findings is that cyberattacks are becoming more often, while they are not increasing in severity.
Index Terms—Hacking breach, data breach, cyber threats, cyber risk analysis, breach prediction, trend analysis, time series, cybersecurity data analytics. I.

INTRODUCTION DATA

Among the most catastrophic cyber disasters are breaches. There were 7,730 data breaches reported by the Privacy Rights Clearinghouse [1] between 2005 and 2017, which equates to 9,919,228,821 information compromised. There were 1,093 data breach occurrences in 2016, up 40% from 780 in 2015, according to the Identity Theft Resource Center and Cyber Scout [2]. In 2015, the United States Office of Personnel Management (OPM) [3] discovered that hackers gained access to the personal information of 4.2 million federal employees and contractors, as well as their background investigation records. This data included 21.5 million Social Security numbers. Not only that, but data breaches can cost a pretty penny. The average cost per lost or stolen record containing sensitive or private information was $158 in 2016, according to IBM [4]. Cybersecurity [5] The assistant editor who oversaw the manuscript's evaluation and final approval to publish was In 2016, there were 1,339 compromised records on average, with a median cost of $39.82 per record, an average cost of $665,000, and a median cost of $60,000. Data breaches are still a major issue, even if technical solutions may make cyber systems more resistant to assaults. The need to describe the development of data breach occurrences is prompted by this. Not only will this help us better comprehend data breaches, but it will also illuminate alternative ways to lessen the impact, including insurance. The present knowledge of data breaches (e.g., the absence of modeling methodologies) is unable to produce appropriate cyber risk measurements to guide the assignment of insurance premiums, despite the widespread belief that insurance would be beneficial [6].

Data breach incidents have just recently been modeled. The statistical properties of personal identity losses in the United States between the years 2000 and 2008 were explored by Maillart and Sornette [7]. They discovered that the number of breach occurrences spikes from 2000 to July 2006, but then stays the same after that. A dataset of 2,253 breach instances spanning a decade (2005–2015) [1] was examined by Edwards et al. [9]. Data breaches have remained relatively stable in both magnitude and frequency throughout the years, according to their findings. From 2000 to 2015, Wheatley et al. [10] examined a dataset that correlates to organizational breach incidences. The dataset is a combination of [8] and [1]. Their research shows that significant breach events (those compromising more than 50,000 data) happen to US organizations at a constant rate regardless of time, whereas non-US enterprises are seeing a rising trend in the frequency of these attacks. Several unanswered concerns inspired this research, including: Is the frequency of data breaches due to cyberattacks growing, decreasing, or staying about the same? An ethical response to this topic will provide light on the state of cyber dangers as a whole. None of the prior research addressed this specific issue. In particular, the dataset examined in [7] only encompasses events that occurred between 2000 and 2008 and does not

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

include breaches that were the result of cyber assaults. In contrast, the dataset examined in [9] is more current and includes two types of breaches: negligent breaches (i.e., incidents caused by devices that were lost, discarded, stolen, or otherwise unintentionally compromised) and malicious breaching. We exclude careless breaches from the current analysis as they are more indicative of human mistake than cyber assaults. While the other three sub-categories are intriguing and merit independent analysis, this study will concentrate on the hacking sub-category (hence referred to as the hacking breach dataset) because there are four sub-categories in the malicious breaches studied in [9]: hacking (including malware), insider, payment card fraud, and unknown. What We've Achieved We provide three new findings in this study. We begin by demonstrating that, instead of using distributions to represent the hacking breach event inter arrival times (which indicate the incidence frequency) and breach magnitude, stochastic processes are the way to go. Our research has shown that specific point processes and ARMA-GARCH models can adequately explain the changing sizes of hacking breaches and the times between incidents, respectively. ARMA stands for "AutoRegressive and Moving Average" and GARCH for "Generalized AutoRegressive Conditional Heteroskedasticity." We show that these models of stochastic processes can foretell both the magnitude of breaches and the intervals between their occurrences. As far as we are aware, this is the first publication to propose modeling these cyber threat elements using stochastic processes instead of distributions. Secondly, we find that the incident inter-arrival times positively correlate with the breach sizes, and we demonstrate that this correlation may be properly captured by a certain copula. We also demonstrate that neglecting to account for dependency leads to inaccurate prediction results for inter-arrival periods and breach sizes. As far as we are aware, this is the first study to demonstrate both the presence of this dependency and the fallout from disregarding it. Our third step is to analyze the trends in cyber hacking breaches using both qualitative and quantitative methods. As the number of hacking breach incidents rises, we see that the situation is deteriorating in terms of the time it takes for incidents to arrive, but things are looking up in terms of the size of the incidents, suggesting that the damage from individual breaches will not worsen. We are hopeful that this study will encourage other research that will provide valuable insights into different ways to reduce risk. Insurance firms, government organizations, and regulators might benefit from these insights as they need a thorough understanding of the risks associated with data breaches.

Related Work

Works Done That Are Highly Relevant to This Investigation: Between 2000 and 2008, 956 instances of per sonal identity loss were examined in a dataset [8] by Maillart and Sornette [7]. The researchers discovered that a heavy tail distribution $\Pr(X > n) \sim n^{\alpha}$, where $\alpha = 0.7 \pm 0.1$, may be used to describe the personal identity losses per occurrence, represented by X. This finding holds true even after separating the dataset into four categories: commercial, academic, governmental, and healthcare. Given that the probability density function of identity losses per event remains constant, the situation of identity loss remains stable when considering the extent of the breach. A separate dataset [1] of 2,253 breach occurrences spanning a decade (2005–2015) was examined by Edwards et al. [9]. Two types of breaches occur: negligent breaches, which may be caused by careless actions like devices being misplaced or stolen, and malicious breaches, which can be caused by malevolent actors like hackers or insiders. Both the magnitude and frequency of breaches have remained relatively constant throughout the years, as shown by the fact that they may be represented by log-normal or log-skewnormal distributions for the former and the negative binomial distribution for the latter. Using data collected from 2000–2015, a dataset of organizational breach incidents was examined by Wheatley et al. [10]. This dataset was compiled from two sources: [8] and [1]. After studying the maximum breach size using Extreme Value Theory [11], they proceeded to simulate the huge breach sizes using a doubly truncated Pareto distribution. They also looked at the data breach frequency using linear regression and discovered that, although it exhibits no trend for non-US firms, the incidence of significant breaching instances for US organizations is independent of time. The interdependence of cyber hazards has also been the subject of research. Two degrees of cyber risk dependency were investigated by Böhme and Kataria [12]: interdependence between enterprises on one level and interdependence between companies on a global level.
Cyber hazards produced by viral occurrences may be modeled using the Archimedean copula. Herath & Herath [13] discovered a relationship between these risks. A Bayesian Belief Network based on copula was used by Mukhopadhyay et al. [14] to evaluate cyber vulnerability. To model dependent cyber threats, Xu and Hua [15] looked into copulas. The reliance discovered while modeling the efficiency of cyber security early-warning was investigated by Xu et al. [16] using copulas. Cybersecurity hazards with

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

dependency and several variables were studied by Peng et al. [17]. The current research stands out from the others because it employs a novel technique to examine breach episodes from a different angle—specifically, cyber hacking breach incidents. Cyber hacking (including malware) has serious consequences, and this viewpoint reflects those consequences. It was discovered using the new technique that there is a positive connection between the incident inter-arrival durations and the breach sizes, and that stochastic processes, not distributions, should be used to represent them. 2) Additional Research That Is Relevant To This Work: In their examination of a dataset [1], Eling and Loperfido [18] used an actuarial modeling and pricing perspective. For their analysis of the rise of cybercrime, Bagchi and Udo [19] used a modified version of the Gompertz model. Using data supplied by the University of Maryland's Office of Information Technology, Condon et al. [20] used the ARIMA model to forecast security events. Using data gathered from a network telescope, Zhan et al. [21] assessed the state of cyber threats. A analogous dataset is reported in [24], and Zhan et al. [22], [23] used datasets gathered at a honeypot to characterize and estimate the frequency of assaults on the honeypot by using their statistical aspects, such as long-range dependency and extreme values. Extreme assault rates were predicted using a marked point approach by Peng et al. [25]. Related cyber security situations were explored by Bakdash et al. [26]. Data breach event forecasting using externally visible network characteristics (e.g., management symptoms) was the subject of research by Liu et al. [27]. Using frameworks such as the institutional theory, the opportunity theory of crime, and the institutional anomie theory, Sen and Borle [28] investigated what variables may raise or lower the contextual risk of data breaches.

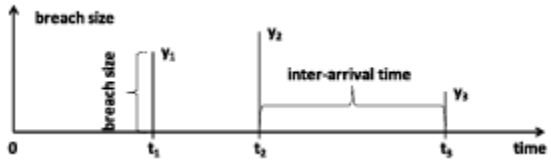


Fig. 1. Illustrative description of cyber hacking breach incidents.

Section C. Outline of the Article How the remainder of the article is structured is as follows. The dataset and research topics are detailed in Section II. A preliminary examination of the data set is detailed in Section III. To analyze the dataset, we create a new point process model in Section IV. We go over how well the suggested model predicts in Section V. Section VI showcases our trend analysis, both quantitative and qualitative. In Section VII, we wrap up our work by outlining potential avenues for further research. We talk about the intuitive interpretations of the basic statistical ideas when they are first stated and save the formal description for the Appendix. Chapter Two: Research Questions and the Dataset Section A: Research Prompts An example of a cyber hacking breach occurrence is shown in Figure 1. At periods t1, t2, and t3, three separate incidences occur, each of which exposes a distinct quantity of data records. Because t2 t1 = t3 t2, the occurrences are not regularly spaced. Time series d1 = t1, d2 = t2 t1, d3 = t3 t2,... represent the arrival times between two successive occurrences; time series y1, y2, y3,... represent the breach sizes, which are the number of data records exposed as a result of an incident.

We are interested in using a dataset that contains cases of cyber hacking breaches to address the following inquiries. First, in describing the inter-arrival times of breach episodes, should we use a distribution or a stochastic process? If so, which one? Answering this question would immediately enhance our understanding of the ever-changing cyber hacking breach scenario from a temporal viewpoint, which is why it is vital. (Parts Three and Four) 2) How about we define the breach sizes using a distribution or a stochastic process? And if so, which one should we use? In order to have a better grasp of the ever-changing cyber hacking breach scenario from a magnitude standpoint, the response to this question is crucial. (Parts Three and Four) 3) Is there no correlation between the breach magnitude and the incident inter-arrival times? If that's not the case, how can we define the relationship between them? In order to have a better grasp of the ever-changing cyber hacking breach scenario from both a temporal and scale standpoint, the answer to this issue is crucial. (Part Four) 4) Is it possible to foretell the scope and timing of the next hacking incident? The capacity to anticipate events and maybe engage in proactive defense on a short time frame (e.g., days or weeks ahead of time) is shown by our response to this question, which is why it is crucial. If the defender determines that there's a good chance a major breach occurrence will happen next week, they may dynamically change their defensive posture, such as implementing more stringent regulations. In the real world, this is analogous to the effects of weather forecasting. Chapter V 5) What patterns may be seen in episodes of hacker breaches? Asking this can help us see the big picture and determine whether things

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

are improving or worsening over a long period of time (say, 10 years) and, if so, by how much. (Part Six) B. Table of Data We used the most comprehensive and widely used publicly accessible dataset on cyber breaches, culled from the Privacy Rights Clearinghouse (PRC) [1], to conduct our analysis in this study. Negligent breaches and the other types of harmful breaches (such as insider fraud, credit card fraud, and unknown) are not taken into consideration since our emphasis is on hacker breaches. We also exclude partial records with unknown, unreported, or missing cyber breach amounts from the remaining raw data. This is because we are interested in studying breach magnitude. From January 1, 2005, to April 7, 2017, 600 hacking incidents in the US are included in the resultant dataset. Organizations in the following sectors have been the targets of hacking attacks: BSF (financial and insurance services); BSR (retail and merchant, including online retail); BSO (other); EDU (educational institutions); GOV (government and military); MED (healthcare, medical providers, and insurance services); and NGO. Here, ti is the day on which an incident of breach size yti (i.e., the number of private data records that are breached) occurs, and t0 is the day on which observation starts (i.e., t0 does not correspond to the occurrence of any incident). The sequence (ti, yti) for each integer from 0 to 600 represents the dataset. With i ranging from 1 to 600, the inter-arrival times are given by di = ti ti 1. There is a daily average of one incident report among the ti's, with 52 days having two, seven days having three, and one day (02/26/2016) having seven. Due to the possibility of unreported hacker breach instances, we must warn that the dataset may not include all of them. Also, instead of the actual dates of the occurrences, the dates that are associated with them are the days that they were reported. The greatest dataset available in the public domain is this one, however, and it is available from this data source [1] [9], [29]. Therefore, it will be possible to assess the severity of the data breach risk by analyzing it, and when more accurate datasets of this kind become available in the future, the approaches may be used to evaluate them. C. Initial Steps You may choose to consider these instances as a single "combined" event (i.e., putting the quantity of compromised data together) because, as we said before, there are days when there are numerous hacker breach incidences.

TABLE I SUMMARYOFNOTATIONS(r.v. STANDSFORRANDOMVARIABLE)

t	time, which is used when describing a general model
$C(\cdot)$	copula function, which is used to model the dependence
$\{(t_i, y_{t_i})\}_i$	the ith incident occurring at time t_i with breach size y_{t_i}
d_i	breach incidents inter-arrival time $d_i = t_i - t_{i-1}$
$\mathrm{VaR}_\alpha(t)$	the Value-at-Risk at level $0 < \alpha < 1$ for r.v. X_t: $\mathrm{VaR}_\alpha(t) = \inf\{l : P(X_t \le l) \ge \alpha\}$

The problem is that many victims may have different cyber systems, so this approach isn't foolproof. Multiple occurrences may be recorded at various times within the same day (e.g., 8pm vs. 10pm) due to the dataset's temporal resolution being a day. In light of this, we suggest creating tiny random intervals of time to divide the occurrences that occur on the same day. To be more precise, we do a random sorting of the incidents that occur on a given day, insert a small random interval between two consecutive incidents (with midnight as the starting point for the first interval), and then make sure that these incidents occur on the same day (for example, if there are two incidents on a given day, they could be assigned at 8am and 1pm).

MODELINGTHE HACKING BREACH DATASET

Here, we take the breach dataset—more especially, the in-sample of 320 incidents—and apply a new statistical model to it. The 280 occurrences that were not part of the sample will be utilized to assess the fitted model's predictive abilities (Section V). A. Inter-Arrival Time Modeling According to the first insight, we should use an autoregressive conditional mean (ACD) model to predict the duration between XUetal.: MODELINGANDPREDICTINGCYBERHACKING BREACHES 2861 stock transactions[30] and later use it to model duration processes (e.g., [31], [40]). 1 Think back to the fact that the dataset is represented as a series (ti,yti) 0≦i≦n, where n = 600, and ti for i ≥1 is the day on which an occurrence of magnitude yti occurs. The periods between arrivals are denoted as di = ti ti 1, where i=1,2,...,n. Using historical data, the conditional mean model seeks to normalize the inter-arrival timedi=ti ti 1 for every i=1,2,...,n. We define in particular

$$d_i = \Psi_i \epsilon_i, \qquad \text{(IV.1)}$$

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Choosing the Right Model: Fi 1 stands for the historical data up to time ti 1, and the◌ʿ i's are innovations that are both independent and identically distributed (i.i.d.), with E(ɬi)=1. Based on our preliminary investigation, we have chosen the following ACD models for model selection: (i) they are basic and can be estimated rapidly in reality; and (ii) they are flexible enough to handle the development of the inter-arrival times. Traditional ACD (ACD) model [30]:

$$\Psi_i = \omega + \sum_{j=1}^{p} a_j d_{i-j} + \sum_{j=1}^{q} b_j \Psi_{i-j},$$

$$\log(\Psi_i) = \omega + \sum_{j=1}^{p} a_j \log(\epsilon_{i-j}) + \sum_{j=1}^{q} b_j \log(\Psi_{i-j}$$

- The type-II log-ACD model ($LACD_2$) [41]:

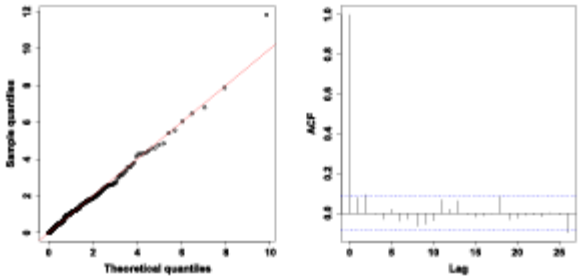
$$\log(\Psi_i) = \omega + \sum_{j=1}^{p} a_j \log(d_{i-j}) + \sum_{j=1}^{q} b_j \log(\Psi_{i-j}$$

Our analysis is further limited to the scenario of p=q=1 in what follows since a higher order does not always increase the prediction accuracy [42]. Standardized Gamma distribution is considered to apply to the ɬi's inventions. We shall verify this assumption below. We assume this to be true since it is both flexible and suggested in previous works as a means of representing data with uneven spacing[40], [42]. It should be remembered that the density function of the generalized gamma distribution is

$$f(x|\lambda,\gamma,k) = \frac{\gamma\, x^{k\gamma-1}}{\lambda^{k\gamma}\,\Gamma(k)} \exp\left\{-\left(\frac{x}{\lambda}\right)^{\gamma}\right\}, \qquad \text{(IV.2)}$$

TABLEIV
MODELFITTINGRESULTSOFTHEACDANDLOG-ACDMODELSTO THEINTER-ARRIVALTIMESOFHACKINGBREACHINCIDENTS. THENUMBERSINTHEPARENTHESESARETHE ESTIMATEDSTANDARDDEVIATIONS

Model	ω	a_1	b_1	k	γ	AIC	BIC
ACD	3.0559 (3.367)	0.0682 (0.065)	0.5705 (0.427)	0.5802 (0.121)	1.2061 (0.170)	1997.81802 -	2016.65963 -
LACD1	**3.825** (0.2254)	**0.058** (0.0241)	**-0.767** (0.0971)	**0.556** (0.1136)	**1.254** (0.1748)	**1993.01132** -	**2011.85293** -
LACD2	0.5931 (0.7333)	0.0505 (0.0506)	0.6977 (0.3541)	0.5787 (0.1202)	1.2073 (0.1692)	1998.07453 -	2016.91613 -



(a) (b)

Fig. 5. Theqq-plot andsampleACFof the residuals for the inter-arrival times. (a)Theqq-plotof residuals. (b)ACFof residuals.

for example, the exponential, the Weibull, the half-normal, and the gammad distributions. For our estimate to guarantee that E(ɬi)=1, we must set

$$\lambda = \frac{\Gamma(k)}{\Gamma(k+1/\gamma)}.$$

This model's parameters are fitted using the maximum likelihood estimation (MLE) approach [31]. The fitting results are described in Table IV. We find that AIC and BIC, two model selection criteria, intuitively quantify how well the proposed model fits the data (i.e., the lower the values, the better the fitting) (as discussed in Appendix B), suggesting that LACD1 should be chosen. Additionally, we note that the estimated standard deviation of 0.0971 is a statistically significant coefficient of LACD1. This proves that the previous inter-arrival timings indeed impact the present inter-arrival times to a considerable degree. Additionally, we note that kγ1 implies that the conditional hazard function of arrival times is U-shaped. As shown in Figure 5, we display the fitting residuals to officially assess the fitting accuracy of LACD1. All points, with the exception of one, cluster around the 45-degree line in Figure 5(a), which is the qq-plot of the residuals, indicating that the fitting is correct. Figure 5(b) displays the sample ACF of the residuals, which allows us to test whether the suggested LACD1 model adequately captures the dependency between the inter-arrival times. We find

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

that the correlations at all lags are quite minor. Specifically, the p-values of the formal McLeod-Li and Ljung-Box statistical tests[31], [36] are shown in the right-hand side of Table V. These tests intuitively examine if there are any remaining correlations in the individuals, as discussed in Appendix C. It is evident that these

PREDICTION

We now look at ways to forecast these variables after demonstrating how to fit the breach sizes and inter-arrival intervals. A. Evaluation Criteria for Predictions We should not forget the Value-at-Risk (VaR) [53] statistic. Assuming that 0 <α, the VaR at level α for the relevant random variable Xt

$$\mathrm{VaR}_{\alpha}(t) = \inf\{l : P(X_t \leq l) \geq \alpha\}.$$

The first algorithm is for predicting the VaRαs of hacking incidents by separately inputting the inter-arrival times and breach sizes. The (dti, yti) i=1,...,m used for fitting and the (dti, yti) i=m+1,...,n utilized for assessment and prediction accuracy are historical incident inter-arrival durations and breach sizes, represented by (dti, yti) i=1,...,m+n. α position. 1. Perform 2 for each integer i from m+1 to n. Predict the conditional mean of the following: i = exp(ϋ+a1log(ℓi 1)+b1log(i 1), using the LACD1 model to estimate the events' inter-arrival durations for ds s = 1,...,i 1. Find the ARMA-GARCH of the size after log transformation, and then forecast the future mean ˆi and standard error ˆ σi. Using the AIC-based bivariate residuals from the prior models, choose an appropriate Copula; In order to determine the incident inter-arrival times, convert the simulated dependent samples u(k) 1,i's into the z(k) 1,i's using the inverse of the estimated generalized gamma distribution, with k = 1,...,10000. To determine the breach sizes, convert the simulated dependent samples u(k) 2,i's into the z(k) 2,i's using the inverse of the estimated mixed extreme value distribution, with k =1,...,10000. Finally, calculate the y(k) Determine the expected 10,000 2-dimensional breach data points d(k) i i,k = 1,...,10000 using Eq. (IV.1) and (IV.3), hence; Use the simulated breach data to calculate the VaRα,d(i) for the incident inter-arrival durations and VaRα,y(i) for the log-transformed breach sizes. 16. If the value of d(k) i is more than VaRα,d(i), then the incident's inter-arrival time is violated. Similarly, if the value of y(k) i is greater than VaRα,y(i), then the breach size is violated. end for Output: Numbers of violations in inter-arrival times and breach sizes.

Table VARTESTSOFPREDICTEDINTER-ARRIVALTIMESAND BREACHSIZESAT LEVELSα = .90,.92,.95

	α	Ob.	Exp	LRuc	LRcc	DQ
inter-arrival time	.90	26	28	.6871	.8522	.9523
inter-arrival time	.92	21	22	.7554	.5157	.6931
inter-arrival time	.95	12	14	.5743	.4979	.4352
breach size	.90	31	28	.5561	.8099	.9996
breach size	.92	27	22	.2336	.4881	.9999
breach size	.95	20	14	.1210	.2673	.9999

To make advantage of rolling prediction, which means that training data is expanded as the prediction operation progresses, it is necessary to re-fit the most recent training data, which may need using alternative copula models. Therefore, additional dependent structure has to be considered. This clarifies the need of re-selecting the copula structure using the AIC criteria (refer to Step 4 of Algorithm 1) so that it better fits the freshly updated training data. You may find the outcomes of the predictions in Table VIII. At the.1 level of significance, we find that the prediction models pass every test. Specifically, for any level of α, the models are able to forecast the future inter arrival times. Regarding the breach sizes, at level α =.90, there are 28 violations in the model predictions and 31 violations in the actual data, which are rather similar. The estimated number of violations according to the model is 14, however there are 20 violations based on the actual data for α =.95. That means the models are being cautious when they estimate the sizes of future breaches. All 280 out-of-sample predictions are shown in Figure 9. The results of the predictions for the incident inter-arrival periods are shown in Figure 9(a). The initial breach sizes are shown in Figure 9(c), although they are difficult to see visually. We exhibit the log-transformed breach sizes in Figure 9(b) for a better visualizing impact. Figure 9(c) shows that there are a number of very large breach sizes that do not match the VaR.95 predictions. As a result, the forecast remains a mystery because it failed to account for some of the very major breaches. Finally, the three statistical tests show that the suggested models are able to accurately forecast the VaRs of the event magnitude and the time it takes for an occurrence to arrive. The

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

suggested models may not be able to accurately forecast the exact values of the extremely large inter-arrival times or the extremely large breach sizes because there are a number of them, and they are significantly larger than the predicted VaR.95 values. However, as shown in Section V-C below, our models are able to forecast the combined likelihoods of a breach incidence of a certain scale happening at a future date.

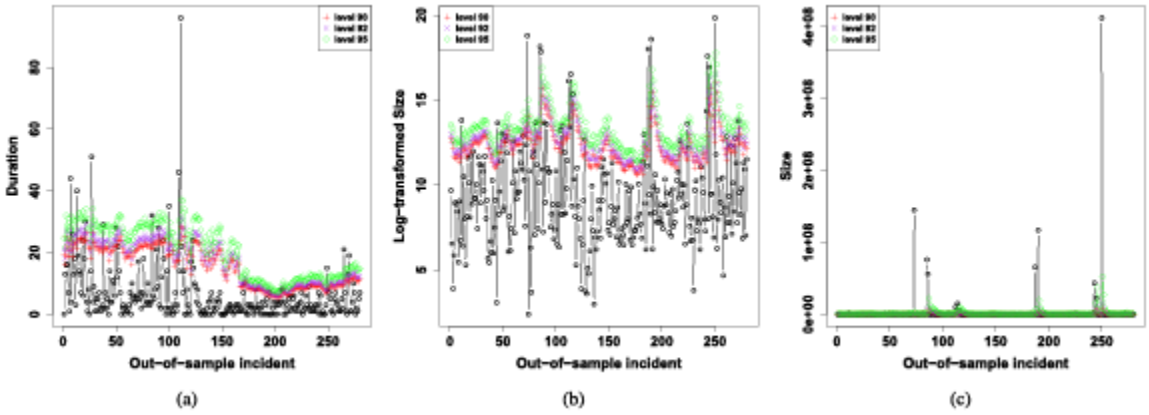


Fig. 9. Predicted inter-arrival times and breach sizes, where black-colored circles represent the observed values. (a) Incidents inter-arrival times. (b) Log-transformed breach sizes. (c) Breach sizes (prior to the transformation)

.

TABLE IX PREDICTEDJOINTPROBABILITIESOF INCIDENTSINTER-ARRIVALTIMESAND BREACHSIZES,WHERE"PROB." ISTHEPROBABILITY OF BREACHSIZEACERTAINPREDICTEDyt OCCURRINGWITHTHENEXTTIMEdt ∈ (0,∞)

Inter-arrival time / Breach size	Copula model					
	Prob.	$d_t \in (30,\infty)$	$d_t \in (14,30]$	$d_t \in (7,14]$	$d_t \in (1,7]$	$d_t \in (0,1]$
$y_t \in (1\times10^6,\infty)$	0.0460	0.0002	0.0042	0.0084	0.0233	0.0099
$y_t \in (5\times10^5,1\times10^6]$	0.0217	0.0001	0.0013	0.0038	0.0129	0.0036
$y_t \in (1\times10^5,5\times10^5]$	0.1107	0.0002	0.0075	0.0200	0.0530	0.0300
$y_t \in (5\times10^4,1\times10^5]$	0.0890	0.0005	0.0055	0.0163	0.0463	0.0204
$y_t \in (1\times10^4,5\times10^4]$	0.2544	0.0005	0.0166	0.0409	0.1240	0.0724
$y_t \in (5\times10^3,1\times10^4]$	0.1156	0.0003	0.0075	0.0178	0.0551	0.0349
$y_t \in (1\times10^3,5\times10^3]$	0.2089	0.0005	0.0110	0.0305	0.1035	0.0634
$y_t \in [1,1\times10^3]$	0.1537	0.0001	0.0068	0.0212	0.0732	0.0524
Total	1	0.0024	0.0604	0.1589	0.4913	0.2870
	Benchmark model					
$y_t \in (1\times10^6,\infty)$	0.0339	0.0002	0.0018	0.0064	0.0176	0.0079
$y_t \in (5\times10^5,1\times10^6]$	0.0224	0.0002	0.0019	0.0033	0.0102	0.0068
$y_t \in (1\times10^5,5\times10^5]$	0.1112	0.0003	0.0070	0.0160	0.0560	0.0319
$y_t \in (5\times10^4,1\times10^5]$	0.0913	0.0002	0.0061	0.0163	0.0425	0.0262
$y_t \in (1\times10^4,5\times10^4]$	0.2568	0.0003	0.0165	0.0439	0.1260	0.0701
$y_t \in (5\times10^3,1\times10^4]$	0.1121	0.0003	0.0070	0.0160	0.0554	0.0334
$y_t \in (1\times10^3,5\times10^3]$	0.2170	0.0009	0.0116	0.0356	0.1066	0.0623
$y_t \in [1,1\times10^3]$	0.1553	0.0007	0.0102	0.0261	0.0779	0.0404
Total	1	0.0031	0.0621	0.1636	0.4922	0.2790

using the previous information about (di, yti), where di = ti ti 1 and yti is the breach size at time ti for i = 1,...,n. The following time intervals are used to categorize the predicted inter-arrival time of the next breach incident: (i) more than one month, for which dt ∈ (30,∞); (ii) between two weeks and one month, for which dt ∈ (14,30]); (iii) between one and two weeks, for which dt ∈ (7,14]); (iv) between one day and one week, for which dt ∈ (1,7]; (v) within one day, for which dt ∈ (0,1]. We also use the following size intervals to break down the anticipated breach

magnitude of the next incident: (i) greater than one million records or yt ∈ (1 × 106,∞), indicating a large breach; (ii) yt ∈ (5 × 105,1 × 106]; (iii) yt ∈ (1 × 105,5 × 105]; (iv) yt ∈ (5 × 104,1 × 105]; (v) yt ∈ (1 × 104,5 × 104]; (vi) yt ∈ (5 × 103,1 × 104]; (vii) yt ∈ (1 × 103,5 × 103]; (viii) smaller than 1000 or yt ∈ [1,1 × 103], indicating a small breach. In order to forecast the combined event, we fit these bivariate data using the aforementioned models and use Algorithm 1 (steps 2–8). Both the copula model and the benchmark model, which assumes independence between incident arrival times and breach sizes, provide descriptions of the expected probability of joint events (dt, yt) in Table IX. These probability diverge from the benchmark model, as we can see. As an instance, the benchmark model yields a chance of only.0339 for data breaches, but the actual likelihood for breach sizes surpassing one million, sometimes known as serious breach occurrences, is.0460 (yt ∈ (1×106,∞). In addition, the benchmark model predicts a probability of.0255, while the copula model gives a probability of.0332 for the combined event of inter-arrival duration dt ∈ (0,7) and breach size yt ∈ (1×106,∞). Because of this, the benchmarkmodel fails to accurately reflect the seriousness of data breach occurrences. Additionally, we note that both models anticipate a breach occurrence happening within a month, with the benchmark model predicting a likelihood of.9969 and the copula model predicting a probability of.9976. This strongly suggests that some kind of data breach will occur within the next thirty days. Additionally, XU et al.: A breach occurrence is predicted to occur within a week by the benchmark model, with a chance of.7783, according to the copula model.

MODELING AND PREDICTING CYBER HACKING BREACHES

This likelihood is.7712 out of 2867. This bodes well for the likelihood of a data breach event occurring within the next seven days. Based on our review of the PRC database, it seems that 1.3 million records were compromised in a data breach that occurred on April 12, 2017. Keep in mind that the data used by our model ends with tn= April 7, 2017, therefore the event occurred within a week, as anticipated by our model. A second intriguing finding is that the model foretold the following: the likelihood of a new event happening within one day (specifically, on April 8, 2017) with a chance of 0.287. We were unable to locate any event reports for April 8, 2017, after reviewing the original dataset. Hence, there is a 28.7 percent probability that a cyber event would go unrecorded. In addition, the forecast shows that, should an unrecorded event occur, the likelihood of the breach's magnitude exceeding 500,000 is very low (0.047); it was less than 50,000 with a chance of 0.7774. From what has been said thus far, we may deduce: Sixth Insight: The suggested method can reliably forecast the combined likelihood that a certain interval will be occupied by hacking incidents of a given size, meaning that incidents of a certain magnitude are likely to occur within that time frame. Using the former technique with the "caveat" that the projected value has a no-more-than-5% probability of being less than the actual value that will be seen is the best approach to anticipate the specific breach size at a certain future point in time. The second approach is the way to go if you want to know how likely it is that a breach of a certain size will occur at a specific future date. Comparable to weather forecasting, this type of prediction capability allows cyber defenders to adapt their defense posture in real-time to lessen the impact of an attack. This could involve temporarily disabling services that aren't necessary or investing more resources into analyzing network traffic, such as costly but effective deep packet inspections or large-scale data correlation analyses. Additionally, military strategy budget estimates could be aided by the prediction model. Given that the probability and severity of an assault determine the amount of time and money needed to protect a company, this is crucial (i.e., quantitative risk management). So, for example, if the model indicates that a massive data breach is highly improbable, the defenses can be less robust (in terms of ratio cost-effectiveness). On the other hand, if the model indicates that a massive data breach is probable, the defender can implement more nuanced defenses, such as honeypots and more precise audit systems. The practical use of weather forecasting lends credence to our belief that predictive-defense, specifically dynamic defense enabled by prediction capabilities, should be a focus of future study.

CONCLUSION

Based on our analysis of a dataset of hacker breaches, we demonstrated that both the amount of the breach and the period between occurrences should be represented by stochastic processes instead of distributions. Both the fitting and prediction accuracies of the statistical models presented in this work are good. Specifically, we suggest to use a copula-based method to forecast the combined likelihood that an event of a certain breach size

would transpire at a later date. Results from statistical analyses demonstrate that the approaches suggested here outperform those previously published in the literature, which failed to account for the interdependence of incidents' arrival timings and magnitude as well as the temporal correlations between them. To get a better picture, we used both quantitative and qualitative methods of analysis. Among the cybersecurity lessons we derived is the fact that cyber hacking breach instances are becoming more often, yet with little increase in damage severity. It is possible to use or modify the methods described in this study to examine comparable datasets. Numerous unanswered questions remain for the benefit of researchers in the future. Finding solutions to problems like missing data (such as unreported breach occurrences) and predicting very big numbers are intriguing and difficult areas to study. If possible, it is also beneficial to try to predict when exactly breach instances would happen. Lastly, more studies are required to determine the predictability of breach incidences, often known as the upper limit of prediction accuracy [24].

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, 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. 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/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.

[21]. Z. Zhan, M. Xu, and S. Xu, "A characterization of cyber security posture from network telescope data," in Proc. 6th Int. Conf. Trusted Syst., 2014, pp. 105–126. [Online]. Available: http://www.cs.utsa.edu/~shxu/socs/intrust14.pdf

[22]. Z. Zhan, M. Xu, and S. Xu, "Characterizing honeypot-captured cyber attacks: Statistical

framework and case study," IEEE Trans. Inf. Forensics Security, vol. 8, no. 11, pp. 1775–1789, Nov. 2013.

[23]. Z. Zhan, M. Xu, and S. Xu, "Predicting cyber attack rates with extreme values," IEEE Trans. Inf. Forensics Security, vol. 10, no. 8, pp. 1666–1677, Aug. 2015.

[24]. Y.-Z. Chen, Z.-G. Huang, S. Xu, and Y.-C. Lai, "Spatiotemporal pat terns and predictability of cyberattacks," PLoS ONE, vol. 10, no. 5, p. e0124472, 2015.

[25]. C. Peng, M. Xu, S. Xu, and T. Hu, "Modeling and predicting extreme cyber attack rates via marked point processes," J. Appl. Stat., vol. 44, no. 14, pp. 2534–2563, 2017.