**Modeling and Predicting Cyber Hacking Breaches**

**ABSTRACT:**

Analysing 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 modelled 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.

**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 analysed in only covered the time span from 2000 to 2008 and does not necessarily contain the breach incidents that are caused by cyber-attacks; the dataset analysed in is more recent, but contains two kinds of incidents: negligent breaches (i.e., incidents caused by lost, discarded, stolen devices and other reasons) and malicious breaching. Since negligent breaches represent more human errors than cyber-attacks, we do not consider them in the present study. Because the malicious breaches studied in contain four sub-categories: hacking (including malware), insider, payment card fraud, and unknown, this study will focus on the hacking sub-category (called hacking breach dataset thereafter), while noting that the other three sub-categories are interesting on their own and should be analysed separately

**DISADVANTAGES:**

* This not only will deep our understanding of data breaches, but also shed light on other approaches for mitigating the damage, such as insurance.
* we do not consider them in the present study. Because the malicious breaches studied in contain four sub-categories: hacking (including malware), insider, payment card fraud, and unknown, this study will focus on the hacking sub-category.

**PROPOSED 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.

**ADVANTAGES:**

* we conclude that the Gumbel copula can adequately describe the dependence between the inter-arrival times and the breach sizes.
* AIC and BIC are the most commonly used criteria in the model selection in the statistics. AIC is meant to balance the goodness-of-fit and the penalty for model complexity (the smaller the AIC value, the better the model).
* 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.

**SYSTEM REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

* System : Intel Core i7.
* Hard Disk : 1 TB.
* Monitor : 15’’ LED
* Input Devices : Keyboard, Mouse
* Ram : 16 GB.

**SOFTWARE REQUIREMENTS:**

* Operating system : Windows 10.
* Coding Language : Python
* Tool : PyCharm, Visual Studio Code
* Database : SQLite