MODELING AND PREDICTING CYBER HACKING BREACHES

Mr. B. Narasimha Rao[1], Dr. Meerasharif Sheik[2], Pitani Venkata Suresh[3],

Roshan Adhikari[4], Ujjwal Timilsina[5], Pothuraju Rekha[6]

[1] *Associate Professor, Dept. of Computer Science and Engineering, Odalarevu*

[2] *Head of the Department for Computer Science and Engineering, Odalarevu*

[3-6] *B.Tech. Student, Dept. of Computer Science and Engineering, Odalarevu*

[1-6] *Bonam Venkata Chalamayya Engineering College, Odalarevu*

Abstract:-

This paper presents a statistical analysis of a 12-year data set (2005-2017) on incidents of cyber hacking, encompassing both malware and non-malware attacks. The examination of these incidents has revealed patterns over time regarding the timing and magnitude of breaches, suggesting that simple distributions may not suffice. This study proposes specific stochastic models that can effectively predict breach timings and their sizes. Furthermore, both qualitative and quantitative trend analyses indicate that the occurrence of cyber breaches is on the rise, while the average size of these breaches remains manageable. This research aims to enhance understanding of the evolution of cyber threats and underscores the necessity for dynamic modeling approaches within cyber security data analysis.

1. INTRODUCTION:

Cyber security has become increasingly vital as organizations depend on digital systems to handle and safeguard sensitive information. Data breaches, which arise from malicious cyber attacks such as hacking, can lead to severe repercussions. Recently, the frequency and complexity of cyber hacking incidents have surged, highlighting the urgent need for dynamic strategies to comprehend and forecast these events. The primary objectives of this project include assessing whether cyber hacking breaches are becoming more common, less frequent, or remaining stable over time, as well as developing predictive models to estimate the timing and impact of future hacking breaches. This research offers valuable insights into the evolution of cyber threats and establishes a statistical basis for incident mitigation and response planning.

2. THE EXISTING SYSTEM:

This study addresses several crucial research questions that have been overlooked in the cyber security field. A significant question is whether data breaches due to cyber-attacks are increasing, decreasing, or holding steady over time. Answering this question can provide alarming insights into the evolution of cyber threats. Understanding these dynamics is essential for effectively forming cyber security policies, managing resources for defence strategies, and promoting awareness of long-term trends in malicious online activities. Previous research did not adequately address this question; studies covering 2000 to 2008 rely on outdated datasets with limited scope, failing to clearly distinguish if the data breaches were direct consequences of cyber attacks. Consequently, conclusions drawn from these studies do not accurately reflect the evolving patterns of cyber threats, as the frequency and complexity of attacks have risen. A more recent study uses a broader and updated dataset, categorizing breaches as either malicious or benign. Benign breaches, resulting from human errors like misplaced devices or accidental data exposure, are excluded from this study as they do not represent intentional cyber threats. Within the malicious breach category, four sub-types are specified, with this research concentrating primarily on hacking breaches that involve targeted cyber attacks, such as malware usage, unauthorized access, and exploitation of vulnerabilities. Other sub-categories, while significant, are set aside in this study to facilitate a more precise analysis of hacking trends and the development of predictive models for future cyber threats.

3. THE PROPOSED SYSTEM

This paper outlines three primary advancements to enhance the effectiveness of prior research. First, we illustrate that both the times between breaches and the sizes of breaches are best depicted using stochastic processes rather than traditional statistical distributions. Specifically, we employ a point process model to capture the timing of breaches and an ARMA (AutoRegressive and Moving Average)-GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) model to characterize the magnitude of breaches. These models provide accurate historical results alongside valuable predictions for future breach patterns. Second, we discovered a positive relationship between the time intervals of breaches and their sizes. This relationship can be aptly modeled using a specific copula function, facilitating the joint modeling of these two variables. Our findings indicate that this dependence is vital for generating accurate predictions; neglecting it results in significant prediction errors. This represents the first study to explicitly recognize and highlight this issue within the realm of cyber incidents or breaches. Lastly, we perform qualitative and

quantitative trend analyses across the 12-year dataset, finding that hacking breaches are increasingly common, signifying a worsening overall cyber threat. However, the average size of each breach has not significantly changed, suggesting that cyber threats may be stabilizing. Our research findings are instrumental in enhancing decision-making regarding cyber risk and we hope this study sparks further investigation into innovative methods for understanding and mitigating rising cyber threats.

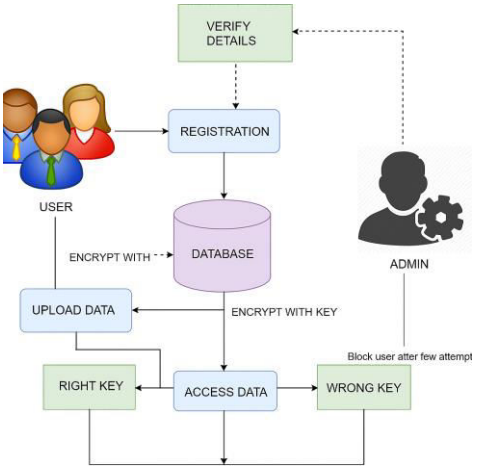
4. ALGORITHM

Support Vector Machine(SVM)

Support Vector Machine (SVM) is a type of supervised learning model that belongs to learning algorithms. It aids in data analysis and pattern recognition, making it useful for both classification and regression analyses. SVM represents each data point from the dataset in an N-dimensional space, where N corresponds to the number of features (or variables). The algorithm seeks a hyperplane that not only distinguishes the classes but also maximizes the margin between them, indicating the greatest distance to the closest data points. The position and pattern of the hyperplane heavily depend on the support vectors, which are the nearest points to the hyperplane.

A wider margin typically enhances model performance as it lowers the likelihood of errors and boosts the model's capacity to generalize to unseen data. However, in many practical scenarios, the data may not be linearly separable in its original format. To tackle this issue, SVM employs a technique known as the kernel trick, which transforms the original data into a higher-dimensional space where linear separation becomes more feasible.

5. SYSTEM ARCHITECTURE



5.1: Registration and Verification

Users must register within the system, and their details are verified either by an administrator or through an automated system before granting access for further procedures.

5.2:Data Upload

Once registered, users can submit data, which is secured using specific keys before

being stored in the database to maintain its security and prevent unauthorized access.

5.3: Data Access

To access stored data, users must input the correct key. If the key is valid, the data is decrypted for the user; if not, access is denied. After several incorrect attempts, the system may block the user.

5.4:Admin Role

The administrator is responsible for verifying user information during registration, monitoring access attempts, and blocking users who have multiple failed access attempts.

6. IMPLEMENTATION

6.1: Upload Data

Only authorized users and administrators can access the database to upload data resources. Data is protected by requiring uploads to be linked with keys, thus preventing unauthorized access. User authorization is based on the information they provide, and only the admin can grant this authorization. Access to the system and the ability to upload or request contents from the database is limited to these authorized users.

6.2: Access Details

The database is a valuable resource containing crucial data, with only administrators holding the keys to this treasure. The admin manages all uploaded data and has exclusive rights to permit access and authorize users based on their specified details.

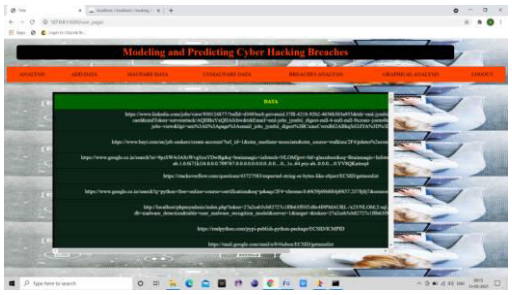
6.3: User Permissions

Users can retrieve data from various resources with the admin's permission, but they must first be authorized based on the verification of their provided details. Any attempts to access data using incorrect credentials will result in a block, which can be lifted by the admin after evaluating the user’s activity in the database.

6.4: Data Analysis

Data gathered from authorized users is analyzed through graphical representations. Graphs facilitate effective analysis, allowing us to predict data trends and policies. Analyzing datasets visually enhances understanding, leading to better model formation. In this paper, we primarily utilize graphical analysis, bar charts, and column charts for representation.

7.Result



8. CONCLUSION

In this study, the dataset of hacking breach incident is analyzed by focusing on two key aspects: the time between breaches and the size of each breach. Our findings conclude that both of these factors are better while using stochastic processes instead of traditional probability distributions because of the presence of temporal correlations. We developed statistical models that fits the data as well as offer accurate predictions. We used a copula-based method to estimate the chance of a breach of a certain size happening within a specific time period. Our result covers all the existing methods in better way which used to ignore the time-based patterns and the relationship between breach frequency and size of the breach. By qualitative and quantitative analysis, we discovered that the cyber breach incidents are becoming more frequent, and the average size of each breach is stable. The methods proposed in this paper can be applied or changed accordingly to analyze the similar type of cyber security data in the future.

9. FUTURE WORK

As every research will have some problems left behind for the upcoming research as well. This research doesn't completely solves the issue with the extreme large values so it will be interesting and some hard work to solve the issues like this. Also, the exact repetition time of the breaches can also be estimated in the future work of this research and more research is needed for solving the unpredictability of the breach incidents. The model can be updated and changed according to the advanced cyber breaches.