ANALYTICS FRAMEWORK FOR CYBER DEFENSE USING DATA FROM MULTIPLE SOURCES

Hitesh Vemulapalli [1], Dr.Saikat Gochhait[2]

Symbiosis Institute of Digital and Telecom Management, constituent of Symbiosis International (Deemed University)

Abstract:

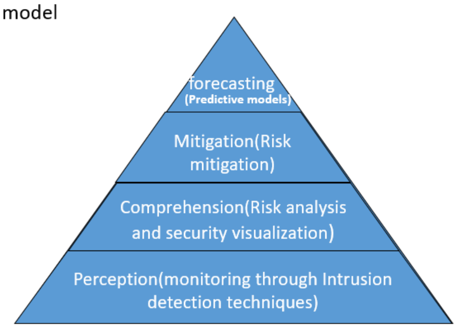
The concept of analytics to improve data security management is the key component for cyber defense against all possible attack vectors. Based on the IT information available and their possible affordances, a research model can be constructed to analyze the mechanism behind Analytics usage for better information security management. At the same time, the model takes care of the position of IT convergence and data-driven community and has been tested empirically using real-time data using partial least squares structural equation model. The data-driven culture and incorporation of IT processes provide a constructive collaboration impact on the dependencies between business analytics and management of data security. However, in the current IT environment, it becomes necessary to define and forecast the intent of the sophisticated targeted attacks using noisy multisource data (Gochhait ,2011). So we discuss ways to merge this heterogeneous data and perform correlation analysis, which can be used in the proposed analytics framework for better detection and prevention against targeted cyber attacks. The framework also recommends using attack graph analysis and several security metrics to understand the effectiveness of our protection systems. This framework can be extended to cloud technologies as well, enhancing the management of cloud computing data security (Gochhait, Shou & Fazalbhoy, 2020). The key to creating a successful framework using analytics is not the amount of data but mining that generates insights. Thus from the perspective of cloud computing, analytics support decision rationality affordance through the decision making affordance for better security management practices.

Introduction:

Under current network settings, Advanced Persistent Threats (APTs) are becoming popular since they are more secretive and sophisticated and they cannot be detected by traditional measures. They also pose a huge security risk to enterprises which can cause serious financial implications. Security managers are not aware of their network being impaired until weeks, months, or years later. There are two aspects to the importance of timely detection and analysis of cyberattacks. First is the early monitoring of intruder behaviors, once the device has been breached unauthorized services if any can be stopped and crash recovery becomes easy. Secondly, APT attacks can take place in many stages, so detection at any stage can stop the entire attack sequence. So prompt identification of pre-attack actions and predicting followup attacks can help us place the right safeguards in place to avoid further damages from the threats. There are two types of intrusion detection technologies, host-based, and network detection. Host-based relies on payload and checks for suspicious code embedded and network-based detection analyzes traffic flows for unknown vectors. Since conventional detection systems based on sole source produces a lot of false intrusion alerts and needs to be manually analyzed further, using multisource data has become a need. Existing methods are unable to detect attacks because of a lack of multisource data correlation and the combinatorial relationship between variables. Since semantic information about data is not effectively expressed manual intervention is required.

Since these APTs are using zero-day exploits (vulnerabilities that have not been disclosed yet), it has become necessary for security teams to focus on unexpected and proactive measures.Through building correct stochastic simulations and building relationships between vulnerabilities life cycle events, detecting attack patterns and resource allocation for protection of organizational assets becomes efficient. By the suggested analytics framework we will have a single platform for vulnerability scoring framework like CVSS[4-6]. Attack graph analysis could also be used in conjunction with various approaches such as graph theory and probabilistic analysis for calculating security metrics. Models can be developed which considers time-dependent variables like the age of vulnerability/exploit. This helps us identify business-critical systems with high-risk exposure based onthe probability of attacks. This systematic strategy for identifying risks and threats is called ‘Cyber Situational Awareness’ and various levels in this approach are shown in the **figure**.

- Cyber situational awareness model


forecasting
(Predictive models)
Mitigation(Risk mitigation)
Comprehension(Risk analysis and security visualization)
Perception(monitoring through Intrusion detection techniques)

The proposed framework for cyberattacks detection based on the integration of heterogeneous data from multiple sources makes use of information gained through correlation analysis is similar to “HeteMSD: A Big Data Analytics Framework for Targeted Cyber-Attacks Detection Using Heterogeneous Multisource Data”.This heterogeneous multi-source data can be classified into three groups containing semantic, non-semantic, and security knowledge. Attackgraphs are a simple representation of possible ways and paths attacker might take to gain access to the system. This is an effective measure since it considers the relationships between vulnerabilities as well. The proposed framework uses the Markov model for security analysis and risk evaluation. This framework helps in identifying the threat levels through security metrics and this can be used to be lookup the heterogeneous data of this possible entry point for the attacker without compromising resource allocation.

Literature review:

Jiageng Chen (2019) paper on “AI-driven Cyber Security Analytics and Privacy Protection” have presented on the developments of cybersecurity in the current internet scenario. It also talks about the various evolving sources of data that are driving the industry towards anticipating better business needs and how the privacy of data can be a real concern. It also talks about how AI-driven cybersecurity solutions can help in cryptanalysis and building new encryption algorithms. The rise of personally identifiable information thefts, network penetrations, and the spread of malicious software can be countered by the use of AI security solutions that use machine learning, big data analytics, statistics, and cryptography (Gochhait and Rimal, 2019).. It also talks about the use of biobehavioral characteristics and deep learning for user authentication on computers to counter insider threats. It also talks about the Secure Information Sharing System(SISS) model with a group key subsystem method. Usage of software metrics for software buffer overflow vulnerability prediction method is mentioned and the decision tree algorithm was proposed.

Ankang (2019) paper on “HeteMSD: A Big Data Analytics Framework for Targeted Cyber-Attacks Detection Using Heterogeneous Multisource Data” focuses on advanced characteristics of data and timely detection along the lines of multistep targeted attacks. It also discusses the limitations of existing methods to compile heterogeneous data and detect anomalies in network traffic. They combined cyberattack analysis methods with big data correlation techniques for comprehensive security awareness. It also characterizes the heterogeneous data and proposed several research points like accuracy, efficiency, Automated intelligence for better detection. Their proposed framework divides the detection into five layers namely Sensing, Event, Alert, Context, Scenario. Event-Event correlation, Alert-Alert correlation, Pattern-Knowledge, and Alert-context correlation are all the techniques used for investigating cyberattacks.

Pelin (2019) paper on “Big Data Analytics for Cyber Security” discusses on the data generated in various platforms in this modern IoT era and various exploits that result in compromising that data. It also visualizes the usage of CPPS (Cyber-Physical Power Systems) and several algorithms for dataset features reduction for intrusion detection. The role-based access control and role mining have been used to find the relationship between permissions to avoid vulnerabilities. Malware detection using neural networks in the byte stream is proved to be better than several existing machine learning models.

Subil Abraham(2014) paper on “A predictive framework for cybersecurity analytics using attack graphs” helps in understanding the importance of security metrics in vulnerability assessments. It proposes a stochastic model for vulnerabilities by taking dynamic attributes like the age of exploit that changes over time into account. It shows how the attack graphs are evolved based on interconnected vulnerabilities and how the security states must be dealing with them. Types of security metrics like Core, Vulnerability based, time-based, structural, probability-based are all explained in detail to construct vulnerability lifecycle models. Evolving security states assessment through node rank analysis, expected path length metric, Probabilistic path metric, and Impact analysis was all discussed in detail.

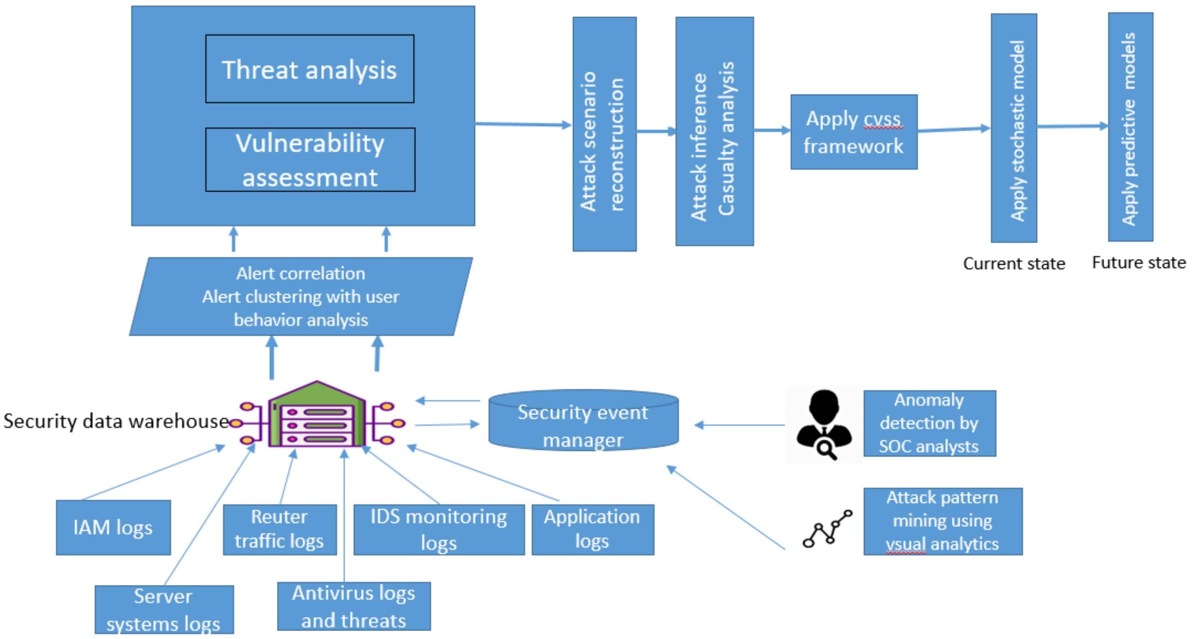
Goodall and Sowul (2009) paper on “Visual Analytics for Cyber Defense”aims at a better understanding of cyber events, enhancing analyst’s awareness through reporting and analysis. Multiple visualizations help provide an overall view of network data in all contexts. It also talks about various tools that enable tagging key elements for sharing, collaboration, and to maintain context. Cognitive Task Analysis, Data integration with data tools form a framework to provide a full understanding of the domain. Overview through various coordinatedmultiple views like geographic and details on demand makes use of the network-related metadata to quickly recognize suspicious behavior and reporting them.

Ben Thorne al;"Using Attack Graphs to Understand Vulnerabilities" focuses on the vulnerability dependencies by laying out the potential exploits. It also discusses the algorithms that prioritize vulnerabilities based on simulated attacks and detailed explanations of how to extend the attack graph-based model to different systems already in place. The experiment reveals various stages of protection before, after, during an attack and benefits of attack graphs to understand the connection between the vulnerabilities helping with the defense. After the attack, the course of action has been explored with better-informed decisions from the attack graph model developed from various complex systems.

Goodall and Sowul (2009) paper on "VI assist: Visual Analytics for cyber defense" talks about tools that can provide adequate support for cyber defense workflows through visual interpretation, reporting, and IP geolocation. Tools like NVisionIP's display of class B IP addresses as a scatterplot and the capability to drill down on them for host details and Flow tagging was included as well for comparison with VI assist. This paper highlights the human's innate perceptual ability to identify attack patterns when presented graphically and how VI assist can offer better protection than other tools by fitting into the existing cyber defense infrastructure.

All these papers discussed above uses only a single source of data from any of the tools like firewalls or access logs or audit trials and detection/prevention techniques which are done on a single layer of perception without nothing concrete to enforce predictive analytics techniques to stop future intrusion attempts. The proposed framework is supposed to evaluate the vulnerabilities from recorded heterogeneous data through multiple sources through all the possible techniques available at our disposal and helps quantify the threats through stochastic models so that future threats can be successfully evaded while monitoring current compromises in security.

Concepts and Proposed Framework:



Heterogeneous data characteristics: The data gathered from heterogeneous multiple sources have various limitations and might need classification accordingly. Most of the attacks carried out today need a vulnerability in the software or operating system that needs to be exploited. So security knowledge data should include vulnerability information so that we can build on various attack patterns and external threat intelligence. Non-semantic data that does not contain any security-related information like traffic logs, operation logs, IAM logs that needs separate analysis for suspicious behavior needs to filter. Semantic data include some sense of attack detection and violation of security policies, taken from operation system defender scans, IDS or IPS logs in the internal network are key to provide an overview of the situation and making the attack patterns obvious.

Architecture Design: For architectural purposes we can divide the defense into different layers, each one supporting others in the overall threat prediction analysis. The sensing layer mainly deals with data aggregation from multiple sources involved and preparing the data for fusion and feature extraction. The event layer takes care of the non-semantic data mentioned above. This will not be an easy task since the data from IAM and traffic logs needs to be analyzed to form a correlation between them that can be later used for event correlation and anomaly detection techniques. Alert layer deals with semantic data from antivirus software, firewalls,and other security monitors. The alerts and notification coming from this data should all be fed into the event security manager for further aggregation and information. The context layer is for obtaining reinforcing information on an already confirmed hypothesis about any possible attack from the alert layer thereby can produce an alert level to introduce human participation for manual analysis. Since traditional machine learning algorithms can only analyze data from a single source, attack context and inference should be perceived at this level. Scenario layer is the topmost and based on attack modeling, casualties are determined and the scenario is closely monitored for threat intelligence expression. The main functions ofthe proposed framework are data preprocessing, analysis, and correlation. Preprocessing is for raw data wrangling to convert it into security semantic data for feature selection and extraction. Analysis function is to rank the notifications generated from semantic data to eliminate any false alerts and increase the reliability of detected anomalies being attacks. Correlation helps in figuring out the attack patterns from establishing multiple correlations among different levels of data and improves situation awareness through human cognition. There are several aspects of correlation like event-event, alert-alert, pattern knowledge, Alert-Context correlation. Multiple datasets from different sources are joined to create a single data format for anomaly detection through data fusion and ensemble learning which provides an event-event correlation. Similarity-based, condition-based, Scenario-based methods are used to establish alert correlation and after removing redundant semantic information of alarms attack patterns are established. As the number of alerts doesn’t correlate with attacks, this significantly decreases the false alarms raised helping security analysts understand better the situation. Representation learning oriented to calculate relationships between entities and their relationships through similarities in descriptive data is used in pattern knowledge correlation. Alert-context correlation is supposed to help eliminate the hardship of manual analysis by removing irrelevant factors of the attack and help for better attack modeling techniques. These techniques include attack chain model, pyramid model, attack graphs, and so on. Here the model proposes stochastic techniques using attack graphs for optimizing the security state of the network.

Attack Graph Modelling: As network systems grow more complicated every day, thereby increasing the number of nodes it becomes increasingly important to graphically model the cyber attacks through attack graphs. Constructing attack graphs has become highly automated through computer-aided tools making it the best approach for carrying a vulnerability analysis for the system. Each node in the attack graph represents an attack state and edges represent a change of state caused by an attacker's decision. Network security metrics that help in security evaluation are usually divided into three categories.

*Metrics:*The first is Core metrics, aggregation metrics that don't have any structure or formula to quantify the security. Examples include Total vulnerability measure calculated from existing vulnerability and Aggregated historical vulnerability measures. CVSS(Common Vulnerability Scoring System) is an open standard used by the US government as well to quantify the IT security vulnerabilities and it updates the database for forty-five thousand known vulnerabilities. Structural metrics use the structure of the attack graph like Shortest Path, which is required by the attacker to choose in our modeled attack graph to reach his end goal, Mean of Path lengths is the average of all the paths in graph to reach attacker's end goal and Number of Paths is the number of ways to reach the end goal. There are also other kinds of metrics including probability-based, which deals with certainties of individual entities being compromised and time-based metrics that depend on the response times of attacks like Mean Time to Breach, Mean Time to Recovery, Mean Time to First Failure. The drawback to these metrics is that they are static and cannot be modeled into stochastic techniques that use the CVSS framework. The attack graph can be generated through heterogeneous data from multiple sources. After its generation, the probabilities can be assigned to the edges to find the most likely vulnerability exploited by the attacker. The stochastic process is then applied to the graph which helps us gain insights over the security state of the network. Since the attack graph is having properties like one absorbed or goal state and every state can lead to an absorbing state, we prefer a model for generating graph as a Markov chain. This works through performing node rank analysis, probabilistic path score, and Temporal attack graph scores and vulnerability exploitability scores through the CVSS scoring framework. After this security analysts can take preventive measures regarding that vulnerability or delay that attacker's goalstate in the graph.

Visual Analytics for security analysts: In the proposed framework there's also a component of anomaly detection by SOC(Security Operations Center) analysts using visual analytics. Although the goal of cyber defense has always been to automate and eliminate human intervention as much as possible, in this case where we deal with complex datasets that are not uniform in nature, it is better to use information visualization by skilled analysts with algorithmic support for augmenting digital security.

Tool specifications: For this field of visualization, various evaluation methods and techniques are developed in the areas of collaboration, Human-Computer Interaction(HCI), situation awareness. But the visualization tools employed for this purpose need to have qualities like usability and learnability, ensuring that the learning curve is flat to enable analysts to quickly adapt to the interface and managing cognitive workloads to ensure the performance of the system is optimal. The main challenges are component interoperability since this needs to be fed into security event managers and feature set utility to make use of all the available features. After selecting a suitable system based on the goals and features mentioned above, the challenge would be to select the right visualization. Several visualizations provide interactive analysis through the TRIAGE algorithm. This algorithm uses graph-based visual analysis and data fusion methods on these weighted graphs to form multi-dimensional clusters. These clusters represent the collection of events correlated by several features, and the specific combinations of them might reveal characteristics of an attack campaign, in some cases the root cause as well. So analysts should be able to decide on which features to include/ look for in the clusters, the aggregation functions to form the clusters and cluster interpretation to formulate insights to feed to the security event manager.

Visualizations and views: The tool might contain dashboards, charts, cross tabs for feature selection to help with the whole process and analysts are free to decide on the views including treemaps and graph views. Treemaps could consist of rectangles with colors representing features and frequency is presented through sizes. They can also provide a drill-down option to reveal the details of event co-occurrences. Usually, treemaps are used to analyze malware and spyware or forensic scan data on file systems. Chord views are used or figuring out the relationships between features in clusters. Circle segments show values and different colors are used for different features. Wherever there is an occurrence of correlated events, they can be highlighted and drilled own further for complex datasets. Graph views are similar to network analysis views in the sense that node is the value, size of the node representing the number of events in the cluster, and edges thickness represents the number of co-occurrences of those events.

Conclusion and Future Work:

In the paper, we have discussed the design of a framework using data from multiple sources for the detection and prevention of targeted cyber attacks to help security professionals thwart potential threats. The correlations mentioned in the proposed framework are discussed based on the existing models, solutions, and provides theoretical fundamentals for future research in effective mitigation techniques. The framework integrates human experts into the analysis and knowledge enrichment layers to move him from being an observer to an active participant. Furthermore, research is required on standardizing metrics from these heterogeneous data and proposed correlation methods which can help in attack investigations as well. Also, the stochastic model for the metrics has several shortcomings in precisely quantifying the security, so trying to bring in other complementary metrics would give a better evaluation of security situation. The TRIAGE algorithm mentioned is to analyze multidimensional clusters and support investigative tasks of analysts using visualization tools. The challenges for this approach include inconsistent design in tools, parametrisation for aggregating datasets to form clusters. For future needs, this can be extended to predictive and forecasting models for proactive monitoring of networks.

References:

AnkangJu, Y. (2019). items: A Big Data Analytics Framework for Targeted Cyber-Attacks Detection Using Heterogeneous Multisource Data. *Hindawi Security and Communication Networks*.

F.Fischer, J. D. (2014). A Visual Analytics Field Experiment to Evaluate Alternative Visualizations for Cyber Security Applications. *EuroVis Workshop on Visual Analytics*.

Gochhait,S. (2011). "Strategic impact of synergy between Information technology and Business processes on the performance of the companies in India" published in refereed International Journal of Innovation, Management and Technology, ISSN: 2010-0248, Vol 2: Issue 4, 2011.

Gochhait,S., Rimal, Y. (2019). "Machine Learning Neural Analysis Noisy Data", International Journal of Engineering and Advanced Technology , ISSN: 2249-8958, 8(6),08/2019

Gochhait, S., Shou, D. T., & Fazalbhoy, S. (2020). *Cloud Computing Applications and Techniques for E-Commerce*. IGI Global. http://doi:10.4018/978-1-7998-1294-4

Goodall, J.R and Sowul,M.(2009) "VIAssist: Visual analytics for cyber defense," *2009 IEEE Conference on Technologies for Homeland Security*, Boston, MA, 2009, pp. 143-150, doi: 10.1109/THS.2009.5168026.

Jiageng Chen, C. S. (2019). AI-Driven Cyber Security Analytics and Privacy Protection. *Hindawi Security and Communication Networks*.

Nair, S. A. (2015). A PREDICTIVE FRAMEWORK FOR CYBER SECURITY ANALYTICS USING ATTACK GRAPHS. *International Journal of Computer Networks & Communications (IJCNC) Vol.7*.

Pelin Angin, B. B. (2019). Big Data Analytics for Cyber Security.

Subil Abraham, S. N. (2014). Cyber Security Analytics: A Stochastic Model for Security Quantification Using Absorbing Markov Chains. *Journal of Communications Vol. 9, No. 12*.

Thorne, B. (2018). Using Attack Graphs to Understand Vulnerabilities.

Zhiying Wang, N. W. (n.d.). An empirical study on business analytics affordances enhancing the management of cloud computing data security. *International Journal of Information Management*.

Ullah, F., and Babar, M.A (2018) 'Architectural Tactics for Big Data Cybersecurity Analytic Systems: A Review'