Intelligent Cyber Security Solutions through High Performance Computing and Data Sciences : An Integrated Approach

Sandhya Maitra*
Dr. Sushila Madan**

Abstract

The recent advances in Data Sciences and HPC despite transforming the ongoing digitization to have a positive impact on the social and economic aspect of our lives, have at the same time, given birth to several security issues. Thus the face of Cyber security has changed in the recent times with the advent of new technologies such as the Cloud, the internet of things, mobile/wireless and wearable technology. The technological advances in data science which help develop contemporary cyber security solutions are storage, computing and behavior. On the other hand high performance computing power facilitates the usage of sophisticated machine learning techniques to build innovative models for identification of malware. Big data holds vital importance in building analytical models which identify cyber attacks. Besides High performance computing is necessary for supporting all aspects of data-driven research. An integrated approach combining the technological benefits provided by predictive power of data sciences and the aggregated parallel processing power of high performance computing would help devise intelligent and powerful cyber security solutions supporting proactive and dynamic approach to threat management to counteract the multitude of potentially new emerging cyber attacks.

Keywords: High Performance computing, Data Sciences, Machine Learning, Cyber Security

I. Introduction

The researchers all over the world face challenges related to upsurge of voluminous data of many areas such as Bioinformatics, Medicine, Engineering & Technology, GIS and Remote Sensing, Cognitive science and Statistical data. Advanced algorithms, visualization techniques, data streaming methodologies and analytics are the need of the hour. These have to be developed within the constraints of storage and computational power, algorithm design, visualization, scalability, distributed data architectures, data dimension reduction and implementation to name a few. The other issues to be considered include optimization, uncertainty quantification, systems theory, statistics and types of model development methods. This requires contextual problem solving based on multidisciplinary approaches. The scale, diversity, and complexity of Big Data necessitates the advent of new architecture, techniques, algorithms, and analytics to manage it and extract value or hidden knowledge from it. Analytics research encompasses a large range of problems of data mining research[1]. Data is increasingly becoming cheap and ubiquitous. The rapid growth in computer science and information technology in the recent times has led to the generation of massive amount of data. This avalanche of data has made a strong impact on almost all aspects of human life and fundamentally changed every field in science and technology. A multitude of new types of data is collected from web logs, sensors, mobile devices, transactions and various instruments. The emerging technologies such as data mining and machine learning enable us to interpret this massive data. The High Performance Computing (HPC) techniques are increasingly being used by organizations to efficiently and effectively deal with processing and storage challenges thrown by explosive growth of such

Sandhya Maitra*
Research Scholar
Banasthali Vidyapith

Dr. Sushila Madan**
Professor
Lady Shri Ram College for Women

enormous data. Advances in Networking, High End Computers, Distributed and Grid computing, Large-scale visualization and data management, Systems reliability, High-performance software tools and techniques, and compilation techniques are taking a new era of high performance, parallel and distributed computing. Over the past few decades security concerns are becoming increasingly important and extremely critical in the realm of communication and information systems as they become more indispensable to the society. With the continuous growth of cyber connectivity and the ever increasing number of applications, remotely delivered services, and networked systems digital security has become the need of the hour. Today government agencies, financial institutions, and business enterprises are experiencing security incidents and cyber-crimes, by which attackers could generate fraudulent financial transactions, commit crimes, perform an industrial espionage, and disrupt the business processes. The sophistication and the borderless nature of the intrusion techniques used during a cyber security incident, have generated the need for designing new active cyber defense solutions, and developing efficient incident response plans. With the number of cyber threats escalating worldwide, there is a need for comprehensive security analysis, assessment and actions to protect our critical infrastructures and sensitive information[1].

II. Cyber Security

The spectacular growth of cyber connectivity and the monumental increase of number of networked systems, applications and remotely delivered services cyber security has taken top precedence amongst other issues. Attackers are able to effect fraudulent financial transactions, perform industrial espionage, disrupt business processes and commit crimes with much ease. Additionally government agencies are also experiencing security incidents and cyber-crimes of dangerous proportions which can compromise on Nations Security. The sophisticated intrusion techniques used in the cyber security incidents and their borderless nature have provided the impetus to design new active cyber defense solutions, and develop efficient and novel incident response plans. The number of cyber threats are escalating globally, necessitating comprehensive security analysis, assessment and action plans for protecting our critical infrastructures and sensitive information[1].

Cyber security in recent times demand secure systems which help in detection of intrusions, identification of attacks, confinement of sensitive information to security zones, data encryption, time stamping and validation of data and documents, protection of intellectual property, besides others. The current security solutions require a mix of software and hardware to augment the power of security algorithms, real time analysis of voluminous data, rapid encryption and decryption of data, identification of abnormal patterns, checking identities, simulation of attacks, validation of software security proof, patrol systems, analysing video material and many more innumerable actions [2].

Analysis of new and diverse digital data streams can reveal potentially new sources of economic value, fresh insights into customer behavior and market trends. But this influx of new data creates challenges for IT Industry. We need to have Information Security measures to ensure a safe, secure and reliable cyber network, for the transmission and flow of information[1].

III. High Performance computing

The re-emergence of need for supercomputers for cyber security stems from their computing capacity ability to perform large number of checks in an extremely short time particularly in the case of financial transactions for the identification of cyber crimes using techniques featuring cross-analysis of data coming from several different sources[2]. The knowledge gained through HPC analysis and evaluation can be instrumental providing comprehensive cyber security as it helps interpret the multifaceted complexities involved in cyber space comprising complex technical, organizational and human systems[3].

A combined system of Distributed sensor networks and HPC cybersecurity systems such as exascale computing helps in real-time fast I/O HPC accelerated processing. This covers various issues such as data collection, analysis and response to takes care of the issues of data locality, transport, throughput, latency, processing time and return of information to defenders and defense devices.

An important set of HPC jobs has involved analytics, discovering patterns in the data itself as in cryptography. The data explosion fueling the growth of high performance data analysis originates from the following factors:

1. The efficiency of HPC systems to run data-intensive modeling.
2. Advent of larger, more complex scientific instruments and sensor networks such as "smart" power grids.
3. Growth of stochastic modeling (financial services), parametric modeling (manufacturing) and iterative problem-solving methods, whose cumulative results are large volumes of data.
4. Availability of newer advanced analytics methods and tools: MapReduce/Hadoop, graph analytics, semantic analysis, knowledge discovery algorithms and others the escalating need to perform advanced analytics by commercial applications in near-real-time such as cloud.

Data-driven research necessitates High performance computing. Big Data fuels the growth of HP data analysis[3]. Research on High Performance Computing includes mainly networks, parallel and high performance algorithms, programming paradigms and run-time systems for data science apart from other areas. High-performance computing (HPC) refers to systems that can rapidly solve difficult computational problems across a diverse range of scientific, engineering, and business fields by virtue of their processing capability and storage capacity. HPC being at the forefront of scientific discovery and commercial innovation, holds leading competitive edge for nations and their enterprises[4]. India in an endeavour to meet its stated research and education goals is making every effort towards doubling up its high performance computing capacity and is exploring opportunities to integrate with global research and education networks.

Cyber Security and Data Sciences

The challenge of protecting sensitive data increased exponentially in recent times because of the non existence of a secure perimeter as before where it was confined to secure data centers as data leaks out of massive data centers into cloud, mobile devices and individual PCS . Most companies do not have policies prohibiting storage of data in mobiles while people on the other hand prefer storing them on to their mobiles with huge computing and storage power for convenience and efficiency of operations.

Cloud-based data mostly exists in commercial data centers, on shared networks, on multiple disk devices in the data center, and multiple data centers for the purpose of replication. The extremely difficult task of developing Cloud security is now made possible with new technologies such as HPC and machine learning.

Data from data centers should be moved to cloud only for business reasons with benefits outweighing the costs of providing cloud security to protect it. Data Inventories should be maintained in encrypted form, tracked and managed well on mobile devices to prevent theft of data. Additionally Cloud networks should be subjected to thorough penetration testing[5].

The value of cyber security data plays a major role in constructing machine learning models. Value of a data is the predictive power of a given data model as well as the type of hidden trends which reveal as a result of meticulous data analysis. The value of cyber security data refers to the nature of data which can be positive or negative. Positive data such as malicious network traffic data either from malware or varied set of cyber attacks hold higher value than data science problems as it can be used to build machine learning based network security models. From cyber security view point the predictive power of effective data models lies in the ability to differentiate normal network traffic from abnormal malicious traffic indicating active cyber attack. Machine learning builds classifiers to identify network traffic as good or bad based on the analysis. The spam filters are based on these techniques to identify normal emails from ad's, phishing and other types of spam. Big Data helps build Classifiers to train a machine learning algorithm and also helps evaluate the classifiers performance. The positive data that a spam classifier needs to detect is behavior exhibited by a spam email. Similarly the network traffic exhibiting behavior of real cyber attacks is positive data for a network security model. Negative data refers to normal data such as legitimate emails in case of spam classifier and normal traffic data for a network security model. In both the cases the classifier should be able to detect bad behavior without incorrectly classifying genuine mails or network traffic to be harmful. The various cyber security problems differ on the basis of quick availability of positive data. In the case of spam emails positive data is easily available in abundance for building a classifier. On the other hand despite increased cyber attacks across various organizations positive data from real cyber attacks and malware infections can seldom be accessed. This is true for especially targeted attacks. The pace at which the hackers modify their techniques to create increasingly sophisticated attacks render libraries of malware samples quickly obsolete. In case of targeted attacks malware is custom built to steal or destroy data in a secret manner. The predictive power of a machine learning model relies on the high value of positive samples in terms of its general nature for identifying potentially new cyber attacks. Additionally performance on these models is highly influenced by the choice of features used to build them. The prerequisites for interpreting huge amount of positive samples are feature selection and appropriate training techniques. The highly unbalanced nature of training data for a machine learning model is owing to negative samples always being many orders of magnitude more abundant than positive data samples. The application of proper evaluation metrics, sophisticated sampling methods and proper training data set balancing helps us find out if we have the appropriate quantity of positive samples or not. The lengthy process of collecting positive samples is one of the first and most important tasks for building machine learning based cyber security models. This is how big data is relevant to cyber security[6].

Intelligent Cyber Security Solutions Powered by HPC and Data Sciences

The advances in Data Sciences and HPC have extended innumerable benefits and conveniences to our day to day activities and transformed the ongoing digitization to deeply impact the social and economic aspects of our lives. At the same time these dependencies have also given rise to many security issues. The attackers in the cyber world are also getting more creative and ambitious in exploitation of techniques and causing real-world damages of major dimensions by making even proprietary as well as personally identifiable information equally vulnerable. The problem is further compounded as designing effective security measures in a globally expanding digital world is a demanding task. The issues to be addressed include defining the core elements of the cyber security, Virtual private network security solutions, Security of wireless devices, protocols and networks, Security of key internet protocols, protection of information infrastructure and database security. The advent of the Internet of Things (IoT) also increased the need to step up cyber security. The Io T is a network of physical objects with embedded technology to communicate, sense or interact with their internal states or the external environment where a digitally represented object becomes something greater than the object by itself or possesses ambient intelligence. Despite its manifold advantages the rapid adoption of IoT by various types of organizations escalated the importance of security and vulnerability. The computing world underwent a major transformation in terms of increased reliability, scalability, quality of services and economy with emergence of cloud computing. Nevertheless, remote storage of data in cloud away from owner can lead to loss of control of data. The success and wide spread usage of cloud computing in future depends on effective handling of data security issues such as accountability, data provenance and identity and risk management. The face of Cyber security has changed in the recent times with the advent of new technologies such as the Cloud, the internet of things, mobile/ wireless and wearable technology[1].

The static data once contained within systems have now become dynamic and travel through a number of routers, hosts and data centers. The hackers in cyber criminals have started using Man-in-the-Middle attacks to eavesdrop on entire data conversations Spying software and Google Glass to track fingerprint movements on touch screens, Memory-scraping malware on point-of-sale systems, theft of specific data by Bespoke attacks.

Context-aware behavioral analytics treats unusual behavior as a symptom of an ongoing nefarious activity in the computer system.

These cases can no longer be handled by tool based approaches fire walls or antivirus machines. The previous solutions no more succeed in managing risk in recent technologies, there is an imperative need for brand new solutions. Analytics help in identifying unusual or abnormal behaviors. Behavior based analytics approaches include Bio Printing, mobile location tracking, behavioral profiles, third party Big Data and external threat intelligence. Now a days hackers carefully analyze a system defenses and use Trojan horses and due to the velocity volume and variety of big data security breaches cannot be identified well in time. Solutions based on new technologies combining machine learning and behavioral analytics help detect breaches and trace the source. User profiling is built and machine behavior pattern studied to detect new type of cyber attacks, the emphasis is on providing rich user interfaces which help in interactive exploration and investigation. These tools can detect strange behavior and changes in data.

This problem can be solved by Virtual dispersive technologies which split the message into several encrypted parts and routed on different independent servers, computers and/or mobile phones depending on the protocol.

This problem can be solved by Virtual dispersive technologies which split the message into several encrypted parts and routed on different independent servers, computers and/or mobile phones depending on the protocol.

The traditional bottlenecks are thus completely avoided. The data dynamically travels on optimum random paths also taking into consideration network congestion and other issues as well. Hackers find it difficult to find data parts. Furthermore in order to prevent cyber criminals exploiting the weak point of the technology which is the place where two endpoints must connect to a switch to enable secure communication, hidden switches are used by VDN making them hard to find.

Critical infrastructures can be protected by security measures and standards provided by Smart Grid technologies. The cloud based applications which are beyond the realm of firewalls and traditional security measures can be secured by using a combination of encryption and intrusion detection technologies to gain control of corporate traffic. Cloud data can be protected by Security assertion Markup language, an XML based open standard format, augmented with encryption and intrusion detection technologies. This also helps control corporate traffic.

Proxy based systems designed through SAML secure access and traffic, log activity, watermark files by embedding security tags into documents and other files for tracking their movement and redirect traffic through service providers. Such solutions neither require software to load on endpoints nor changes to end user configurations. Any kind of suspicious activity such as failed or unexpected logins etc are alerted by notifications. The security administrators can instantaneously erase corporate information without effecting personal data of users. Active defense measures such as counter intelligence gathering, sink holing, honey pots and retaliatory hacking can be adopted to track and attack hackers. Counter intelligence gathering is a kind of reverse malware analysis in which a cyber expert secretly finds information about hackers and their techniques. Sink holing servers hand out non routable addresses for all domains within sink hole. Malicious traffic is intercepted and blocked for later analysis by experts. Isolated systems called Honey pots such as computer, data or network sites are set up to attract hackers. Cyber security analysts to catch spammers to prevent attacks etc.. Retaliatory hacking is most dangerous security measure which usually considered illegal as it may require infiltration into a hacker community, build a hacking reputation to prove the hacking group of your credentials. None of these things being legal raises debate over active defense measures. Early warning systems forecast sites and server likely to be hacked using machine learning algorithms. These systems are created with the help of machine learning and data mining techniques. Most of the algorithms take into the account a website software, traffic statistic, file system structure or webpage structure. It uses a variety of other signature features to determine the presence of known hacked and malicious websites.

Notifications can be sent to website operators and search engines to exclude the results. Classifiers should be designed to adapt to emerging threats. Such security measure is growing in its scope. The more data that absorbs the better will be its accuracy[7].

The cyber threats in recent times necessitate state of the art dynamic approach to threat management. The Cyber security threats rapidly changing with technological advancements. An application vulnerability free today may be exposed to a major unanticipated attack tomorrow. A few of recent examples are of Adobe Flash vulnerability allowing remote code execution, NTP (Network Time Protocol) issue allowing denial-of-service attacks, Cisco ASA firewall exposure allowing for denial-of-service attacks, and Apple, thought for a long time to be invulnerable, releasing iOS 9, quickly followed by additional releases to correct newly discovered exposures. The dynamic threats are the key challenges to information security and necessitate dynamic security approaches for their mitigation. Neither were these a resultant of negligence on the part of affected parties nor was it the result of a change affected by these parties in the products. The information security programs should be proactive, agile and adaptive. A few of the strategies for moving from static to a dynamic is by making vulnerability checks a regular and frequent task with monthly external scans and internal scans conducted on same schedule or when software or configuration changes are made, whichever happens first, paying attention to fundamentals such as checking logs and auditing access rights. Firmware updates should be top priority as many of the exposures we face today result from issues found in the firmware of devices attached to our network score devices such as routers and firewalls, or Internet of Things devices, such as printers and copiers. Threat sources should be studied on a regular basis[8].

Data science techniques help in the prediction of types of security threats decides reacting to these threats. Data sciences and cyber security were highly isolated disciplines until recent times. The cyber security solutions are usually based on signatures which use pattern matching with prior identified malware to capture cyber attacks. But these signature based solutions could not prevent zero day attacks for unidentified malware as they lack predictive power of data science. Data science effectively uses scientific techniques to draw knowledge from data. The ongoing security breaches accentuate the need for new approaches for identification and prevention of malware. The technological advances in data science which help develop contemporary cyber security solutions are storage, computing and behavior. The storage aspect eases the process of collection and storage of huge data on which analytic techniques are applicable. On the other hand high performance computing power assists machine learning techniques to build novel models for identification of malware. The behavioral aspect had shifted from identification of malware with signatures to identify the specific kind of behaviors exhibited by an infected computer. Big data plays a key role analytical models which identify cyber attacks. Any rule based model based on machine learning requires large number of data samples to be analyzed in order to unearth the set of characteristics of a model. Subsequently data is required to cross check and assess the performance of a model.

Application of machine learning tools to enterprise security gives rise to a new set of solutions. These tools can analyze networks, learn about them, detect anomalies and protect enterprises from threats[9].

Machine learning increased in its popularity with the advent of high performance computing resources. This has resulted in the development of off-the-shelf machine learning packages which allow complex machine learning algorithms to be trained and tested on huge data samples. The aforementioned characteristics render machine learning as an indispensable tool for developing cyber security solutions. Machine learning is a broader data science solution for detecting cyber attacks. Minor changes in malware can leave Intrusion Prevention Systems and Next-generation Fire wall perimeter security solutions performing signature matching in network traffic ineffective. The rigorous analytical methods of data sciences differentiate abnormal behavior defining an infected machine after identifying normal behavior through repetitive usage. Therefore contemporary cyber security solutions require big data samples and advanced analytical methods to build data-driven solutions for malware identification and detection of cyber attacks. This results in spectacular improvement of cyber security efficacy[10].

Conclusions

- Cyber Security Solutions should be more proactive and dynamic.
- Effective Cyber Security Solutions for future threats can be achieved by exploiting the processing and storage power of High Performance Computing.
- Intelligent Cyber Security Solutions can be built by exploring the predictive power of machine learning and data mining approaches.
- Machine learning approaches require Big Data for training models.
- Big Data can be efficiently processed in real time using High Performance Computing.
- Cloud Computing, IoT can be highly risk prone in the absence of effective security framework.
- The Solution to Future security needs lies in integrating the processing and storage power of High Performance Computing with predictive power of machine learning and data mining techniques.

References

1. S. Maitra, “NCETIT’2017", *iitmipu.ac.in*, 2017. [Online]. Available: http://iitmipu.ac.in/wp-content/uploads/2017/02/NCETIT-2017-Brochure.pdf. [Accessed: 14- Feb- 2017].
2. “HPC solutions for cyber security”, *Eurotech.com*, 2017. [Online]. Available: https://www.eurotech.com/en/hpc/industry+solutions/cyber+security. [Accessed: 11- Feb- 2017].
3. C. Keliiaa and J. Hamlet, “National Cyber Defense High Performance Computing and Analysis: Concepts, Planning and Roadmap”, Sandia National Laboratories, New Mexico, 2010.
4. S. Tracy, “Big Data Meets HPC”, *Scientific Computing*, 2014. [Online]. Available: http://www.scientificcomputing.com/article/2014/03/big-data-meets-hpc. [Accessed: 11- Feb- 2017].
5. R. Covington, “Risk Awareness:The risk of data theft — here, there and everywhere”, *IDG Contributor Network*, 2016.
6. D. Pegna, “Cybersecurity, data science and machine learning: Is all data equal?”, *Cybersecurity and Data Science*, 2015.
7. “Hot-technologies-cyber-security”, *cyberdegrees*, 2017. [Online]. Available: http://www.cyberdegrees.org/resources/hot-technologies-cyber-security/. [Accessed: 04-Feb- 2017].
8. R. Covington, “Risk Awareness:Is your information security program giving you static?”, *: IDG Contributor Network*, 2015.
9. B. Violino, “Machine learning offers hope against cyber attacks”, *Network World*, 2016.
10. D. Pegna, “Cybersecurity and Data Science:Creating cybersecurity that thinks”, *IDG Contributor Network*, 2015.