International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No. 2, 68-74
Available online at http://pubs.sciepub.com/iteces/4/2/4

DOI:10.12691/iteces-4-2-4

Big Data in Network Security Systems

Lidong Wang*

Department of Engineering Technology, Mississippi Valley State University, Mississippi, USA
*Corresponding author: lwang22@students.tntech.edu

Abstract The purpose of this paper is to introduce several issues of network security including network security systems, techniques and approaches, events fusion, and real-time data processing. Artificial intelligence methods such as data mining and machine learning in network security are also presented. Big data in network security systems is also introduced, which includes big heterogeneous data, big data processing technologies (stream processing, batch processing, and micro-batch processing), The encryption and security mechanism of big data and some applications of Big Data technologies in network security are also offered. Challenges of Big Data in network security are also deliberated.

Keywords: *big data, network security, stream processing, data mining, data engineering, information technology*

Cite This Article: Lidong Wang, "Big Data in Network Security Systems." *International Transaction of Electrical and Computer Engineers System*, vol. 4, no. 2 (2017): 68-74. doi: 10.12691/iteces-4-2-4.

1. Introduction

The protection of critical infrastructure (CI) and critical information infrastructure (CII) has been regarded as an important matter of national security and cybersecurity. CI consists of all critical sectors of a nation's infrastructure. According to the definition of the European Union for critical infrastructure, 'critical infrastructure means an asset, system or part thereof located in Member States which is essential for the maintenance of vital societal functions, health, safety, security, economic or social well-being of people, and the disruption or destruction of which would have a significant impact in a Member State as a result of the failure to maintain those functions' [1]. Infrastructures are described as complex integrated systems. CII is a subset of CI. CII has been described as a system which is part of a global or national information infrastructure that is essential for the continuity of critical services, which includes hospitals, public transport, banking and finance, telephone networks, data centres, the Internet, and utility services, etc. There are two aspects of CII: a physical component aspect which includes equipment such as phones, radios, televisions, computers, high speed networks, satellites, and wireless communication networks; and an immaterial aspect which is the information and content that is stored and flows through the physical component. Attacks on CII are increasing in volume and sophistication with destructive consequences. With the increase of mobile and user connectivity, the rise of the Internet of Things (IoT), and the various types of threats and vulnerabilities, the Internet and its infrastructures have become more complex and vulnerable to attacks. In 2014, the Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) reported that the Energy sector and the Critical Manufacturing Sector were two of the primary targets and took the highest attack vector. More than half of the attacks were advanced persistent threats (APT) [1,2].

Organizations usually have anti-virus software, firewalls, computer workstations, intrusion detection systems (IDSs), and end-user applications, etc. However, there is rarely any integration among traditional IDSs in the context of monitoring for security breach attempts and rarely any sort of integrated security monitoring approach across a large proportion of an organization's information systems [3]. Network data acquisition can be performed using a combination of web crawler, task system, word segmentation system, and index system, etc. As one widely used data collection method, log files are record files automatically generated by the data source system to record activities in designated file formats for subsequent analysis [4]. Automated intrusion detection refers to methods of detecting unwanted network access, for example, producing alarms or events to report possible attacks through scanning detailed network activity logs [5]. There are often six phases for intrusion into a network [6]:

- Probing phase: an intruder collects information regarding the operating system, firewall and the user profile. The information will narrow the intruder's options for finding the weaknesses within the system.
- Gaining the initial access phase: this phase includes parameters such as invalid password attempt, user's terminal (network address), and user networking hours.
- Gaining full system access: the following activities occur which include illegal file/directory access attempt, illegal password file access attempt, and illegal application access.
- Performing the hacking attempt: the intruder is going to use system facilities and information in this phase.

- Covering hacking tracks: the intruder will erase all the track or clues that would result in the exposure of his/her access routes and identity.
- Modifying utilities to ensure future access: the intruder will create a backdoor in the system for his or her future access.

Social engineering is defined as "one of the simplest methods to gather information about a target through the process of exploiting human weakness that is inherit to every organization" [7]. Five of the most common types of social engineering attacks to target victims include Phishing, Baiting, Pretexting, Quid pro quo, and Tailgating [7]. The industry has been using IP reputation for years to identify malicious destinations on the Internet. Security researchers evaluate each IP address and determine whether it is 'bad' or 'good' via automated methods based on the activity observed across a massive network of sensors. IP reputation has shown itself to be a good indicator that an address has already been used for malicious activity at some point. However, malicious IP addresses (and even domains) are not active for long because attackers cycle through domains and addresses frequently to avoid detection. IP reputation is not sufficient to identify all the Command and Control (C&C) traffic on the network — many malicious sites used in targeted attacks are not shown in IP reputation lists [8]. The purpose of this paper is to investigate the advances of network security systems and the applications and challenges of Big Data in network security systems.

2. Network Security Systems and Methods

2.1. Categorization of Network Security Systems

Intrusion detection and prevention systems (IDPS) can be categorized into four various types shown in Table 1 [9]. An HIPS detects malicious activities on a single computer. Network-based intrusion detection can be categorized into two types: flow-based anomaly detection and packet-based anomaly detection. Flow-based anomaly detection usually relies on existing network elements such as switches and routers to make the flow of information available for analysis. On the other hand, packet-based anomaly detection doesn't rely on other network components; it observes network traffic to detect anomalies. The software of packet-based anomaly detection does not use third party elements to generate the metadata of network traffic. Instead, the entire packet-based analysis looks at raw packets as they traverse network links. Network behaviour analysis (NBA) is an effective approach to intrusion detection. Without NBA systems added to the security model, the architecture could require three to four times more intrusion prevention system devices. Although intrusion detection and intrusion prevention systems can spot common and signature-based attacks such as certain viruses, port scans, and denial of services, they cannot trap the security attacks of fast spreading such as zero-day worms [10].

Table 1. Categorization of IDPS

Categories	Description
Host based intrusion prevention system (HIPS)	Monitors single host for suspicious activity by analysing events occurring within the host.
Network based intrusion prevention system (NIPS)	Analyses the traffic of entire network by analysing protocol activities and take appropriate actions.
Network behaviour analysis system (NBAS)	Examines traffic to identify threats that generate unusual traffic flow, such as malware, policy violation, and DDOS attack.
Wireless intrusion prevention system (WIPS)	Analyses the traffic of wireless network by analysing protocol activities and take appropriate actions.

Snort is an open and free source network intrusion detection and prevention system. Sax2 is a network-based IDS. It is a professional intrusion detection and prevention system that performs constant daily and hourly network monitoring, real-time packet capturing, advanced protocol analysing, and automatic expert detection. Sax2 and Snort are real-time traffic analysers. A comparison between Snort and Sax2 indicates: 1) Snort is an open source IDS and Sax2 is shareware IDPS; 2) Snort is supported by all the major OS while Sax2 is only supported by Windows; 3) Snort analyses all of the protocols while SAX2 analyses IP, UDP, TCP, FTP, HTTP, POP3, and SMTP protocols, etc. [11].

2.2. Wireless Local Area Network

A wireless local area network (WLAN) IDS is like an NIDS; it can analyse network traffic. However, it also analyses wireless-specific traffic including connection to access points (AP), rogue APs, users outside the physical area of the company, and WLAN IDSs built into APs. With networks increasingly support wireless technologies at various points of a topology, WLAN IDS will play greater roles in security. Because WLANs have other functionality and vulnerabilities, a WLAN IDS needs to monitor network-based attacks as well as wireless-specific attacks. A wireless IDS contains several components such as sensors, management logging databases, and consoles like a NIDS. Wireless IDSs are unique in that they can be run centralized or decentralized. In centralized systems, data is correlated at a central location and decisions and actions are made based on the data. In decentralized systems, decisions are made at sensors. WLAN IDS sensors can monitor several types of events such as wireless specific events and events monitored on wired networks. WLAN sensors can detect anomalies such as unusual usage patterns, DoS attacks, poorly secured WLAN devices, unauthorized WLANs and wireless devices, wireless scanners war driving tools, and man in the middle (MITM) attacks. The limited scope of these events means that WLAN IDS results are usually more accurate than wired IDS results [12].

A wireless IDPS that monitors wireless network traffic and analyses it to identify suspicious activity is involved with wireless networking protocols. Wireless IDPS technologies may also be needed if organizations

determine that their wireless networks need additional monitoring or if organizations plan to ensure that rogue wireless networks are not in use in the organization's facilities. NBA technologies can also be used if organizations desire additional detection capabilities for worms, denial of service attacks, and other threats that NBAs are particularly well-suited to detect [13].

2.3. Techniques and Approaches in Network Security

Anomaly-based detection has been used to find attackers for over a decade. It typically uses Netflow. Normal traffic patterns (source/destination/protocol) are profiled for users on the network and then traffic variations from the baseline which exceed tolerances are checked [8]. Signature-based detection (also called misuse detection) identifies patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks [14]. Hybrid-based approach has been proposed to improve the capabilities of an IDPS by combining the two methods (anomaly-based detection and signature-based detection). The main idea is that the signature-based approach can detect known attacks while anomaly-based approach can detect unknown attacks [15].

The use of information-theoretic measures such as entropy, relative entropy, conditional entropy, information gain, and information cost has been explored to capture the intrinsic characteristics of normal data and such measures have been used to guide the process of building and evaluating anomaly detection models. Efficient approaches were also developed which use statistics on packet header values for network anomaly detection [14]. The operations of network intrusion detection systems (NIDS) rely on network traffic analysis. Network traffic from several protocols (HTTP, DNS, and SIP, etc.) is inspected to find anomalies. These anomalies are defined by rules that rely on either signatures or anomalous traffic behaviour. Current IDSs analyse several protocols; data and events observed by them are correlated by SIEM (Security Information and Event Management) to detect intrusions. However, one shortcoming is that current solutions performing in-depth packet analysis are not scalable and adaptable to big networks that generate high volumes of data. Because these scalability problems exist, Internet Service Providers (ISPs) usually collect IP flow data since this represents an aggregated view of traffic by discarding the payload. Main IP flow record attributes are source and destination ports, source and destination IP addresses, the version of IP protocol, timestamp, the number of bytes, and the number of packets exchanged. However, not only the volume of data and its velocity, but also the variety of data is a challenge [16]. The data is high dimensional typically with a mix of continuous as well as categorical attributes. A challenge faced by anomaly detection techniques in this domain is that the nature of anomalies keeps changing over time as the intruders adapt their network attacks to evade the existing intrusion detection solutions [17].

Anomaly-based technologies detect "abnormal" behaviours; therefore, this approach may catch the unknown attacks, but typically has a high false positive rate. Signature-based technologies catch known attacks effectively, but are not effective for the rapid growing new types of attacks [18]. Some of the methods that can be used to identify signature are [19]:

- Email containing a specific virus. The IDS can compare the subject of each email with the subject associated with the virus-laden email, or it can look for an attachment with a specific name.
- Connection attempt from a reserved IP address: This is easily identified by checking the source address field in an IP header.
- Denial of service attack on a POP3 server caused by issuing the same command thousands of times: A signature of this attack should be kept track of how many times the command is issued and an alert should be issued when the number exceeds a certain threshold.
- Packet with an illegal TCP flag combination: This can be found by comparing the flags set in a TCP header with the combinations with a known good or bad flag.
- File access attack on an FTP server by issuing file and directory commands to it without first logging in: A state-tracking signature can be identified which would monitor FTP traffic for a successful login and would alert if certain commands were issued before a user had authenticated properly.
- Domain Name System (DNS) buffer overflow attempt contained in the payload of a query: By parsing the DNS fields and checking the length of each of them, the IDS can identify an attempt to perform a buffer overflow using a DNS field; another method is looking for exploit shell code sequences within the payload.

2.4. Log Files, Events Fusion and Real-time Data Processing for Network Security

Most of the data that is used to analyse network security comes from log files of every event. Log files may include the records of attempts to access a website or to download a file, system logins, email transmissions and authentication attempts. The high volume of data generated by log files enables researchers to identify malfunctions, attacks, or suspicious activity in the system. Intel Security gathers log file data from clients, network devices, servers, specialised sensors, and specific applications. It also collects contextual information that helps security experts to interpret the events captured in log files. Data compiled in log files and contextual information is maintained in various formats and should be put into a consistent format and inserted into a system for analysis. For each network event, data is extracted, put into standard formats, and loaded into a data warehouse [20]. Logs may consist of many various types such as event logs, database logs, and sever logs, etc. The contents of logs can be numerical data as well as non-numerical data. For some kinds of logs like server logs, numerical data (e.g. CPU load, memory) is sufficiently representative as features. But meaningful non-numerical data (e.g. the indicators about the state of systems) are also valuable. Features reflecting the text contents of the logs are often the focus of research [21].

Fusion of IDS events is an active research area. For example, it can be helpful in combining data from multiple IDS sources, or examining other flows that involve suspected attackers and/or compromised clients. Problems and limitations of automatic approaches to intrusion detection [5] are: 1) many alerts make makes text-based manual analysis of the IDS output tiresome and error-prone; 2) threshold adaptation is difficult because small changes often have unpredictable effects; and 3) missing context information often makes the interpretation of some alerts problematic or even impossible and therefore can potentially lead to misjudgement of any threats.

Most studies have been conducted in an off-line learning fashion and all the features of training instances are given a priori. Such assumptions may not always hold for real-world applications where training examples may arrive in an online manner. For example, training data usually arrive sequentially in an online spam email detection system, which makes it difficult to employ a regular batch feature selection technique in an efficient, scalable, and timely manner [22]. There are often five phases of real-time data processing [23]:

1) Data distillation — This includes combining heterogeneous data sources, filtering for populations of interest, extracting features for unstructured text, selecting relevant features and outcomes for modelling, and exporting sets of distilled data to a local data mart.
2) Model development — Includes sampling and aggregation, variable transformation, model estimation, model refinement, and model benchmarking.
3) Validation and deployment — Encompasses re-extracting fresh data, running it against the model, and comparing the results with outcomes obtained based on the data which has been withheld as a validation set.
4) Real-time scoring — Scoring is triggered by actions at the decision layer, and the actual communications are brokered by the integration layer. Hadoop is not particularly well-suited for real-time scoring although it can be used for "near real-time" applications. New technologies such as Cloudera's Impala have been designed to improve Hadoop's real-time capabilities.
5) Model refresh — Data is always changing; therefore, it is necessary to refresh the data and the model built on the original data.

3. Artificial Intelligence in Network Security

Data mining and machine learning are important methods in artificial intelligence. Even though the offline processing has a few significant advantages, data mining techniques can be used to improve IDSs in real-time processing [14]. Data mining based on the association rule is one of the approaches to solving intrusion detection problems. Size and dimensionality of the feature space are the two primary complications in the IDS development [6]. Data mining algorithms can be used for anomaly detection and misuse detection. In misuse detection, training data are labelled as either "normal" or "intrusion." A classifier can then be derived to detect known intrusions. Anomaly detection builds models of normal behaviour and automatically identifies significant deviations from it [24]. Anomaly-based detection is designed to uncover abnormal patterns of behaviour. There are different categories of anomaly-based detection methods and three of them are most commonly used [15]:

- Statistical: The system observes the activity of subjects (such as CPU usage or the number of TCP connections) according to the statistical distribution and creates profiles which represent their behaviors. Two profiles are made: one is made during the training phase and the other is the current profile during the detection. An anomaly can be detected if there is a significant difference between the two profiles.
- Data mining based: Data mining techniques can help improve the process of intrusion detection by unfolding associations, changes, patterns, anomalies, and important events and structures in data. Clustering, classification, outlier detection, and association rule discovery are data mining techniques used in IDPS.
- Machine learning based: System call-based sequence analysis, Markov models, and Bayesian network are the most frequently used techniques.

The LOF (Local Outlier Factor) algorithm has been used to perform 'outlier-oriented' studies. It has been reported that LOF is superior to any other outliers detecting algorithm for identifying network intrusion. Another reason to use LOF is that the machine learning framework Jubatus has the LOF algorithm for its standard repertoire. Jubatus is a distributed machine learning platform. It was developed for real-time, deep analysis in a distributed environment. Some examples of detecting cyber-attacks using an outlier detecting algorithm are: there are possibilities of DoS attacks when the period of packet transmission is too short; or there are possibilities of BufferOverFlow attacks when the length of the packet is too long [25].

Mining information from heterogeneous databases and global information systems is very important. Local- and wide-area computer networks (such as the Internet) connect many sources of data and form distributed, huge, and heterogeneous databases. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are very difficult to be discovered by simple query systems. The discovery of knowledge from various sources of structured, semi-structured, or unstructured data with diverse data semantics is a great challenge for data mining. Web mining that uncovers interesting knowledge about Web structures, Web contents, Web dynamics, and Web usage has become a very challenging and fast-evolving field in data mining [24].

4. Big Data in Network Security Systems

4.1. Big Heterogeneous Data in Network Security

The ability of Big Data analytics to correlate data from a wide range of data sources over significant time periods has resulted in a lower false positive rate and allows the

APT (advanced persistent threats) signal to be detected in the noise of authorized user activities. Before Big Data analytics can be used for the detection of sophisticated threats, there is a need for new detection algorithms capable of processing large amounts of data from diverse data sources. There is also a need to further handle issues that are related to the specific problem of malicious activity detection using correlated data sources such as collecting information from untrustworthy sources [26].

Security events should be correlated with each other to improve alerting accuracy as well as give a more comprehensive overview of cyber threats. A challenge of large organizations in big data is handling an incredible amount of host log event data. In addition, it is very difficult to correlate events with a large amount of data, especially data with many different formats. A more comprehensive approach for monitoring diverse heterogeneous event sources for intrusion detection can yield a better situational awareness of the threats, minimize false alarms, and improve detection accuracy by correlating security events among diverse sources. While a more comprehensive security monitoring system across heterogeneous systems could improve security, it would further exacerbate the big data challenge in intrusion detection. Integrating across more security sensors would increase the problems of big data in: 1) volume and velocity — more data flow in and out of the monitoring system at a high rate; 2) variety — data with various types and formats come from different sources and collectively generating high dimensionality. Feature selection could figure considerably where numerous diverse heterogeneous sources are analysed and correlated because feature reduction can drastically reduce volume, velocity, and variety. It is an important technique in addressing big data challenges in intrusion detection and can notably save classification processing time and improve classification accuracy [3]. The problem of online feature selection (OFS) has also been investigated. The goal of OFS is to develop online classifiers that involve only a small, fixed number of features. The performance of the proposed algorithms for mining big data sets has been evaluated, in which case each of the datasets contained at least 100,000 instances [22].

4.2. Some Big Data Technologies in Stream Processing, Batch Processing and Micro-batch Processing

Data processing can be divided into three main approaches: batch processing, micro-batch processing, and stream processing. The analysis of large sets of static data that are collected over previous time periods is done with batch processing. Micro-batch treats stream data as a sequence of smaller data blocks. Stream processing analyses massive sequences of unlimited data that are continuously generated. Table 2 [27] lists some technologies of big data that are categorized according to specific processing paradigms. Spark is also regarded as a micro-batch processing method. Distributed stream processing platforms are a new class of real-time monitoring systems that analyse and extract knowledge from large continuous streams of data. These types of systems are crucial for providing the high throughput and low latency required by Big Data or Internet of Things (IoT) monitoring applications [28].

Table 2. Some Big Data Technologies Categorized according to Processing Paradigms

Paradigms	Technologies
Batch processing	Sqoop, Pig, Apache Mahout, MapReduce, Hadoop (HDFS, Hive, HBase)
Stream processing	Strom, Spark Streaming, Splunk, *S*4, SQLstream
Hybrid processing	SummingBird, Lambdoop

Apache Flink is a hybrid processing platform supporting both batch and stream processing. One of the most used technologies for stream analytics is Apache Storm. Other big data technologies for stream analytics like Apache Spark are also quickly becoming the analytics engine of choice for many organisations. Storm is an open-source framework for robust, distributed, and real-time computation on streams of data. One of the main advantages of Storm is that it can be used in manifold data gathering scenarios including stream processing and distributed remote procedure call (RPC**)** for solving computationally intensive functions on-the-fly and continuous computation applications. Storm uses an upstream backup and acknowledgments mechanism to ensure that tuples are re-processed after failure. Streaming computation is treated as a series of deterministic batch computation in small time intervals in Spark. When a stream enters Spark, it divides data into micro-batches which are the input data of the distributed resilient dataset (RDD) and the main class in Spark Engine stored in memory. Then the Spark Engine executes by generating jobs to process the micro-batches. Storm and Flink possibly lose messages despite using more complex recovery mechanisms. Storm loses more messages since it uses a subsystem called Zookeeper for nodes synchronization. Flink uses a checkpoint algorithm and has a lower message loss rate during the redistribution process after a failure. It has a smaller loss of messages during a fault compared with Storm [2,28,29].

4.3. The Encryption and Security Mechanism of Big Data in Network Systems

The security mechanism of big data is an important issue. Big data brings about challenges to data encryption because of its large scale and high diversity. The performance of previous encryption methods on small and medium-scale data could not meet the demands of big data; consequently, efficient big data cryptography approaches need to be developed. Effective schemes for access control, safety communications, and safety management need to be studied for structured, semi-structured, and unstructured data. Furthermore, availability, completeness, isolation, confidentiality, controllability, and traceability of tenants' data under the multi-tenant mode also need to be enabled in the premise of efficiency assurance [4]. Some information security issues (e.g., network, big data, Hadoop, MapReduce, and authentication) have been suggested in the past and are as follows [30]:

- File encryption: All the stored data should be encrypted. Distinct encryption keys should be used

on different machines and the key information should be stored centrally behind strong firewalls.
- Network encryption: All the network communication should be encrypted according to industry standards.
- Logging: All the MapReduce jobs that modify the data should be logged. Also, any information from users who are responsible for those jobs should be logged. These logs should be audited regularly to determine if any malicious operations are performed or if any malicious user is manipulating the data on the nodes.
- Software format and node maintenance: Nodes that run the software should be formatted regularly to eliminate any virus. All the application software and Hadoop software should be updated to make the system more secure.
- Nodes authentication: Whenever a node joins a cluster, it should be authenticated. Authentication techniques like Kerberos help validate the authorized nodes and avoid malicious ones.
- Rigorous system testing of MapReduce jobs: After a developer writes a MapReduce job, it needs to be thoroughly tested in a distributed environment instead of a single machine to ensure the robustness and stability of the job.
- Move security close to the data: Place controls as close as possible to the data store and to the data itself to create a more effective line of defense.
- Deploy a purpose-built security solution for Hadoop and big data: Zettaset Orchestrator provides an enterprise-class security solution for big data that is embedded in the data cluster itself. It is designed to meet the security requirements of the distributed architectures and improves the user authentication process through the fine-grained access control.

4.4. Some Applications of Big Data Technologies in Network Security

Although a single event does not reveal a user's intent, a collection of logged events may reveal the intent. This hidden intent or latent semantics in the terminology of topic modelling is modelled as a "topic", which is defined as a probability distribution over the vocabulary (the set of monitored events) and is called topic-word distribution. A hybrid approach to knowledge discovery from big data for intrusion detection was proposed using latent dirichlet allocation (LDA), in which the topics capture the "patterns" of both security incidents and normal activities by using "bag of words" and probability distribution. This new pattern representation has the flexibility to capture attacks in a wider range [18].

Analysing logs, network packets, and system events for intrusion detection and forensics has been a substantial issue. Traditional technologies cannot support long-term and large-scale analytics and Big Data technologies has the potential to deal with the problem for several reasons [31]:

1) Storing and retaining a large volume of data is not economically feasible. Hence, most event logs and other recorded computer activity are deleted after a fixed retention period (e.g., 60 days).
2) Performing analytics and complex queries on large and structured datasets is inefficient because traditional tools do not leverage Big Data technologies.
3) Traditional tools are not designed to analyze and manage unstructured data. So, traditional tools have rigid and defined schemas; however, Big Data tools can query data in flexible formats.
4) Big Data systems use cluster computing infrastructures. Therefore, the systems are more reliable and provide guarantees.

Big Data tools have the potential to provide a significant advance in actionable security intelligence by reducing the time for consolidating, correlating, and contextualizing diverse security event information; and for correlating long-term historical data for forensic purposes. Two examples of Big Data analytics for network security are interesting: first, a large-scale graph inference approach introduced to identify malware-infected hosts in an enterprise network and the malicious domains accessed by the enterprise's hosts; and second, analysing terabytes of DNS events consisting of billions of DNS requests and responses collected at an ISP. The goal is to use the rich source of DNS information to identify malicious domains, botnets, and other malicious activities in a network [31].

Big data are often unstructured and metadata (data about data) is important. Linked data provide some significant advantages in correlating different records to provide a view of a bigger picture. There are three rules for linked data [32]:

1) Linked data uses HTTP URIs — not just for documents as with "traditional" websites but for the subjects of the documents such as products, places, and events, etc.
2) Fetching data using the URI returns data in a standard format with useful information such as who is attending an event and where a person was born.
3) When the information is retrieved, relationships are defined. The relationships are also expressed using URIs; therefore, looking up people links them to their towns where they were born, their regions, and their countries, etc.

5. Conclusion

Network behaviour analysis is an effective method in intrusion detection and helps avoid more intrusion prevention system devices. WLAN IDS results are often more accurate than wired IDS results. Anomaly-based intrusion detection has a high false positive rate. Signature-based methods work well in detecting known attacks; however, they are not effective for new types of attacks. Data mining helps improve the process of intrusion detection. Feature selection helps save classification processing time and improve classification accuracy. Distributed stream processing platforms are a new class of real-time monitoring systems. Big data encryption is a challenge due to large scale and high diversity. Big Data systems are more reliable due to using cluster computing infrastructures. Big data are often

unstructured and metadata is important. Linked data helps correlate different records.

References

[1] Stouten F. Big data analytics attack detection for Critical Information Infrastructure Protection. Thesis, Department of Computer Science, Electrical and Space Engineering, dissertation, Luleå University of Technology, 2016.

[2] Oseku-Afful T. The use of Big Data Analytics to protect Critical Information Infrastructures from Cyber-attacks, 2016, 1-64.

[3] Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: a survey. Journal of Big Data. 2015 Feb 27; 2(1): 3.

[4] Chen M, Mao S, Liu Y. Big data: A survey. Mobile Networks and Applications. 2014 Apr 1;19(2):171-209.

[5] Mansmann F, Fischer F, Keim DA, North SC. Visual support for analyzing network traffic and intrusion detection events using TreeMap and graph representations. In Proceedings of the Symposium on Computer Human Interaction for the Management of Information Technology 2009 Nov 7 (p.3). ACM.

[6] Kabiri P, Ghorbani AA. Research on intrusion detection and response: A survey. IJ Network Security. 2005 Sep 1; 1(2): 84-102.

[7] Conteh NY, Schmick PJ. Cybersecurity: risks, vulnerabilities and countermeasures to prevent social engineering attacks. International Journal of Advanced Computer Research. 2016 Mar 1; 6(23): 31-38.

[8] Rothman M. Network-based Threat Detection, Technical Report, Securosis, LLC, June 19, 2015, 1-24.

[9] Beigh BM, Peer MA. Intrusion Detection and Prevention System: Classification and Quick Review, ARPN Journal of Science and Technology, 2(7), August 2012, 661-675.

[10] Youssef A, Emam A. Network intrusion detection using data mining and network behaviour analysis. International Journal of Computer Science & Information Technology. 2011 Dec 1; 3(6): 87-98.

[11] Harbola J, Vaisla KS, Harbola A. An Examination of Network Intrusion Detection System Tools and Algorithms: A Review. International Journal of Computer Applications. 2014 Jan 1; 95(6).

[12] Wu TM. Information Assurance Tools Report–Intrusion Detection Systems. Sixth Edition. Information Assurance Technology Analysis Center (IATAC), USA, 25-09-2009. Retrieved online from http://iac.dtic.mil/csiac/download/intrusion detection.pdf.

[13] Scarfone K, Mell P. Guide to intrusion detection and prevention systems (idps). NIST special publication. Publ. 800-94, 2007 Feb.

[14] Lappas T, Pelechrinis K. Data mining techniques for (network) intrusion detection systems. Department of Computer Science and Engineering UC Riverside, Riverside CA. Jan 2007, 1-13.

[15] Patel A, Taghavi M, Bakhtiyari K, JúNior JC. An intrusion detection and prevention system in cloud computing: A systematic review. Journal of network and computer applications. 2013 Jan 31; 36(1): 25-41.

[16] Marchal S, Jiang X, State R, Engel T. A big data architecture for large scale security monitoring. InBig data (BigData Congress), 2014 IEEE international congress on 2014 Jun 27 (pp. 56-63). IEEE.

[17] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM computing surveys (CSUR). 2009 Jul 1;41(3):15.

[18] Huang J, Kalbarczyk Z, Nicol DM. Knowledge discovery from big data for intrusion detection using LDA. InBig data (BigData Congress), 2014 IEEE international congress on 2014 Jun 27 (pp. 760-761). IEEE.

[19] Singh J, Nene MJ. A survey on machine learning techniques for intrusion detection systems. International Journal of Advanced Research in Computer and Communication Engineering. 2013 Nov; 2(11):4349-4355.

[20] Data B. Analytics: Seeking Foundations for Effective Privacy Guidance. A Discussion Document, February 2013.

[21] Li W. Automatic Log Analysis using Machine Learning. Department of Information Technology, Uppsala University. 2013 Nov.

[22] Hoi SC, Wang J, Zhao P, Jin R. Online feature selection for mining big data. InProceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications 2012 Aug 12 (pp. 93-100). ACM.

[23] Barlow M. Real-Time Big Data Analytics: Emerging Architecture. O'Reilly Media, Inc. 2013 Jun 24.

[24] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011 Jun 9.

[25] Ogino T. Evaluation of Machine Learning Method for Intrusion Detection System on Jubatus. International Journal of Machine Learning and Computing. 2015 Apr 1; 5(2): 137.

[26] Virvilis N, Serrano O, Dandurand L. Big Data analytics for sophisticated attack detection. ISACA Journal. 2014; 3: 22-25.

[27] Wang H, Xu Z, Fujita H, Liu S. Towards felicitous decision making: An overview on challenges and trends of Big Data. Information Sciences. 2016 Nov 1; 367: 747-765.

[28] Lopez MA, Lobato A, Duarte OC. A performance comparison of Open-Source stream processing platforms. InIEEE Global Communications Conference (Globecom), Washington, USA 2016 Dec.

[29] Curry E, Kikiras P, Freitas A. et al. Big Data Technical Working Groups, White Paper, BIG Consortium, 2012, 1-167.

[30] Chandrasekhar AM, Revapgol J, Pattanashetti V. Big Data Security Issues in Networking. International Journal of Scientific Research in Science, Engineering and Technology, 2(1), 2016, 118-122.

[31] Cárdenas AA, Manadhata PK, Rajan S. Big data analytics for security intelligence. University of Texas at Dallas@Cloud Security Alliance. 2013 Sep.

[32] Mitchell I, Wilson M. Linked Data: Connecting and exploiting big data. White paper. Fujitsu UK. 2012 Mar; 302-323.