Security Analytics: Big Data Analytics for Cybersecurity

A Review of Trends, Techniques and Tools

Dr. Tariq Mahmood
College of Computing & Information Sciences
Karachi Institute of Economics & Technology
Karachi, Pakistan
mahmood@pafkiet.edu.pk

Uzma Afzal
Department of Computer Science
Federal Urdu University of Arts, Science and Technology
Karachi, Pakistan
uzma.afzal@fuuast.edu.pk

***Abstract*—The rapid growth of the Internet has brought with it an exponential increase in the type and frequency of cyber attacks. Many well-known cybersecurity solutions are in place to counteract these attacks. However, the generation of Big Data over computer networks is rapidly rendering these traditional solutions obsolete. To cater for this problem, corporate research is now focusing on Security Analytics, i.e., the application of Big Data Analytics techniques to cybersecurity. Analytics can assist network managers particularly in the monitoring and surveillance of real-time network streams and real-time detection of both malicious and suspicious (outlying) patterns. Such a behavior is envisioned to encompass and enhance all traditional security techniques. This paper presents a comprehensive survey on the state of the art of Security Analytics, i.e., its description, technology, trends, and tools. It hence aims to convince the reader of the imminent application of analytics as an unparalleled cybersecurity solution in the near future.**

Keywords—cybersecurity; big data analytics; security analytics; survey; real-time; suspicion; fraud; outliers; network streams

I. BACKGROUND AND MOTIVATION

The internet is a global network of interconnected systems which serves billions of users worldwide. Its popularity and rapid growth have come at an expensive cost, i.e., loss of information and resources due to cyber threats and attacks. The first cyber crime was reported in 2000 and infected almost 45 million internet users [1]. Over the few past years cyber-crimes have increased rapidly with cyber criminals continuously exploring new ways to circumvent security solutions to get illegal access to computer systems and networks. Some important cyber attacks are as follows:

- **Spamming:** Spamming is sending unsolicited bulk messages to multiple recipients [2]. By 2015, the spam volume is forecasted to be 95% of all email traffic [3]. Munging, access filtering and content filtering are important anti-spam techniques. Munging makes email addresses unusable for spammers, e.g., abc@gmail.com munges as “abc at gmail dot com”. Access filtering detects spam based on IP and email addresses while content filtering recognizes pre-defined text patters in emails to detect spam.
- **Search Poisoning:** Search poisoning is the dishonest use of Search Engine Optimization techniques to falsely improving the ranking of a webpage [4]. Typically, frequent search keywords are used to illegally direct users towards short-lived websites. The first poisoning case was reported in 2007 [5], followed by many others.
- **Botnets:** Botnets are networks of malware-infected compromised computers managed by an adversary [6]. Attackers use bot software equipped with integrated command and control system to control these zombies (bots) and group them into a network called the *bot net* [7].
- **Denial of Service (DoS):** A DoS attack makes a system or any other network resource inaccessible to its intended users. It is launched by a large number of distributed hosts, e.g., bot net. Many defensive techniques such as intrusion detection systems, puzzle solution, firewalls etc. have been developed to prevent DoS attacks [8].
- **Phishing:** Phishing fraudulently acquires confidential user data by mimicking e-communication [9], mainly through email and web spoofing [10]. In email spoofing, fraudulent emails direct users to fraudulent web pages which lure to enter confidential data. In web spoofing, fraudulent websites imitate legitimate web pages to deceive users into entering data. Many anti-phishing solutions are in corporate use to counteract this threat.
- **Malware**: Malware is software programmed to perform and propagate malicious activities, e.g., viruses, worms and Trojans. Viruses require human intervention for propagation, worms are self-propagating, while Trojans are non-self-replicating. Damage from malware includes corruption of data or operating system, installation of spyware, stealing personal credentials or hard disk space etc.
- **Website Threats:** Website threats refer to attackers exploiting vulnerabilities in legitimate website, infecting them and indirectly attacking visitors of

978-1-4799-1288-9/13/$31.00 ©2013 IEEE

these sites. SQL injections, malicious ads, search result redirection are the few techniques which are used to infect the legitimate sites [11].

The extensive damage caused by these cyber attacks has lead to the design and implementation of *cybersecurity* systems. Cybersecurity refers to the techniques, processes and methodologies concerned with thwarting illegal or dishonest cyber attacks in order to protect one or more computers on any type of network from any type of damage [12]. The important goals of cybersecurity are: 1) securely obtain and share information for accurate decision-making, 2) find and deal with vulnerabilities within applications, 3) prevent unauthorized access and 4) protect confidential information. Some of the well-known cybersecurity solutions are being provided by Accenture, HP, Invensys, IBM, EADS, CISCO, Unisys etc.

More recently, the focus of cybersecurity has shifted to monitoring network and Internet traffic for the detection of bad *actions* as compared to the traditional approach of the detection of bad *signatures*. Specifically, traditional cybersecurity is focused on catching malware by scanning incoming traffic against malware signatures which only detect limited-scope threats that have been already encountered in the past. Also, the development of signatures lags far behind the development of cyber attack techniques. Thus, techniques like intrusion detection systems, firewalls and anti-virus softwares can be easily rendered ineffective by hackers. This scenario has become more crucial in the presence of *big data* within computer networks – petabytes and exabytes of information being transferred daily between nodes make it very easy for hackers to enter any network, hide their presence effectively and cause severe damage efficiently. These big data problems are stressed in the following points:

- Corporations are now extending their data networks to allow partners and customers to access data in different ways to facilitate collaboration, hence making networks more vulnerable to cyber attacks. The advent and extensive use of cloud and mobile computing have also generated novel cyber attack methods.
- The advent of big data has seen a corresponding increase in the hacking skills of cyber attackers, and evading traditional security measures such as signature-based tools is now a thing of the past.
- Due to big data, it is possible only to collect a relatively small slice of security information, e.g., network logs, Security Information and Event Management (SIEM) alerts, access records etc. Hence, damage done by novel hacking methods could be realized only after an attack.
- Big data also prevents most security data from being analyzed due to its complexity, e.g., data could be coming from different sources, it could be stored in different formats on different machines, or could be generating too quickly to make any type of analysis feasible through traditional techniques, computer hardware and software architectures.

Security Analytics addresses these issues by reinventing the wheel of cybersecurity. It employs techniques from Big Data Analytics (BDA) to derive useful information for thwarting cyber attacks [13]. It is now possible due to the progress in hardware and software technology that can cope up with big data (see Section II). It provides the following unique features:

- A more agile decision-making approach for networks managers with surveillance and monitoring of real-time network streams,
- Dynamic detection of both known and previously-unknown suspicious/malicious behavior, usage, access pattern, transaction or network traffic flow, applicable to all types of cyber threats
- Effective detection of suspicious and malicious behavior (least possible false positive rate),
- Ability to deal with suspicious and malicious behavior in real-time,
- Appropriate dashboard-based visualization techniques to provide full visibility (360^0 view) of network progress and problems in real-time.
- Appropriate big data hardware and software to cope up with the aforementioned requirements.

The need for the applications of BDA to cybersecurity is just starting to be realized even in developed countries like USA. In this paper, we present a survey on the BDA cybersecurity applications, focusing mainly on the BDA solutions available and its importance according to the network landscape of the future. This paper is structured as follows: In Section II, we present the basics of BDA and in Section III, we detail the BDA cybersecurity issues and applications. We conclude the paper and present future work in Section IV.

II. BASICS OF BIG DATA ANALYTICS (BDA)

Big data is data whose complexity hinders it from being managed, queried and analyzed through traditional data storage architectures, algorithms, and query mechanisms. The "complexity" of big data is defined through 3V's: 1) *volume* – referring to terabytes, petabytes, or even exabytes (1000^6 bytes) of stored information, 2) *variety* – referring to the co-existence of unstructured, semi-structured and structured data, and 3) *velocity* – referring to the rapid pace at which big data is being generated. . Some researchers have added the 4^{th} V, i.e., *veracity* to stress the importance of maintaining quality data within an organization. Some primary sources of big data transactions are data from computer networks, telecommunication networks, finance, healthcare, social media networks, bio-informatics, E-Commerce, surveillance etc. The domain of Big Data Analytics (BDA) is concerned with the extraction of *value* from big data, i.e., *insights* which are non-trivial, previously unknown, implicit and potentially useful. These insights have a direct impact on deciding or manipulating the current business strategy and drives what is being called "From Data to Decision" initiative [14]. The assumption is that *patterns* of usage, occurrences or behaviors exist in big data. BDA attempts to fit mathematical models on these patterns through different data mining techniques such as

Predictive Analytics, Cluster Analysis, Association Rule Mining, and Prescriptive Analytics [13]. Insights from these techniques are typically represented on interactive dashboards and help corporations maintain the competitive edge, increase profits, and enhance their CRM.

It is important to note that the term "big" in big data is relative; even gigabytes of data can be "big" if it is not being managed or queried efficiently. In such a situation, BDA is largely supported by Apache's Hadoop framework, which is an open-source, completely fault-tolerant and a highly scalable distributed computing paradigm. Hadoop is able to distribute BDA tasks across commodity hardware nodes through the MapReduce algorithm (of Google). At an abstract level, data is first "mapped" in a domain-specific format and then processed at different nodes; results from each node are then "reduced" to produce the final result. Hadoop is being used by big companies like Yahoo!, Facebook, Twitter, eBay, and Amazon, and some notable Hadoop-based BDA solutions are currently being offered by IBM, Microsoft, Oracle, Talend, Cloudera, Greenplum, Hortonworks and Datameer. Along with this, notable names providing big data hardware architectures are Teradata, HP (Vertica), Infobright, Aster Data and ParAccel.

The basic stages of BDA process are shown in Fig. 1. Initially, data to be analyzed is selected from real-time streams of big data and pre-processed (cleaned). This is called ETL (Extract Transform Load) typically and is a strenuous activity that can take up to 60% of the effort of BDA, e.g., catering for inconsistent, incomplete and missing values, normalizing, discretizing and reducing data, ensuring statistical quality of data through boxplots, cluster analysis, normality testing etc., and understanding data through descriptive statistics (correlations, hypothesis testing, histograms etc.). Once data is cleaned, it is stored in BDA databases (cloud, mobile, network servers etc.) and analyzed with analytics with a possible use of Hadoop if required. The results are then shown in interactive dashboards using computer visualization techniques. It is important to note that BDA is a "trial-and-error" activity driven by feedback to any of previous stage in order to refine the outputs through "tuning" of analytical approaches.

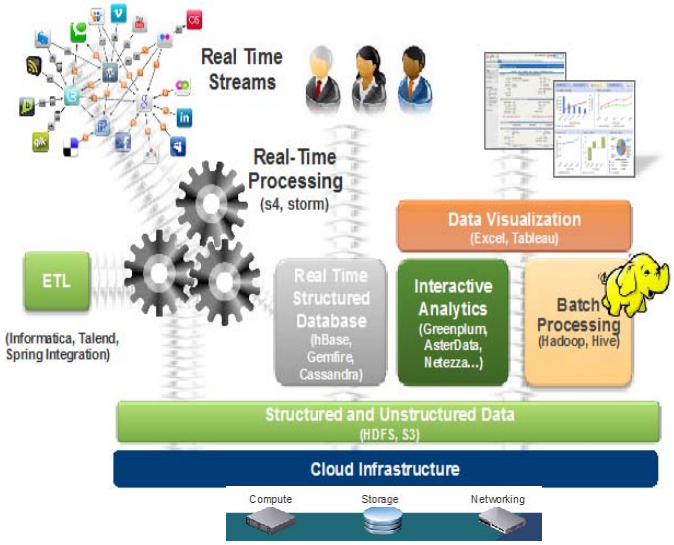

Fig. 1 Basic BDA process

Fig. 1 also shows well-known corporate solutions (in parentheses) that are available for different BDA stages, e.g., Informatica for ETL, Netezza for analytics etc.[1]

III. SECURITY ANALYTICS WITH BDA

Compared to traditional approaches, security analytics provides a "richer" cybersecurity context by separating what is "normal" from what is "abnormal", i.e., separating the patterns generated by legitimate users from those generated by suspicious or malicious users.

A. Big Data Sources for Security Analytics

The concept of "data" for security analytics is expansive, and can be categorized into passive and active sources [15]. Passive data sources can include:

- Computer-based data, e.g., geographical IP location, computer security health certificates, keyboard typing and clickstream patterns, WAP data.
- Mobile-based data, e.g., GPS location, network location, WAP data.
- Physical data of user, e.g., time and location of physical access of network.
- Human Resource data, e.g., organizational role and privilege of the user.
- Travel data, e.g., travel patterns, destinations, and itineraries.
- SIEM data, e.g., network logs, threat database, application access data.
- Data from external sources, e.g., rogue IPs, external threats.

Active (relating to real-time) data sources can include:

- Credential data, e.g., user name and password
- One-time passwords, e.g., for online access
- Digital Certificates
- Knowledge-based questions, e.g., "what is your typical activity on Saturdays from 3 pm to 6 pm?"
- Biometric identification data, e.g., fingerprint, facial recognition, voice recognition, handwriting recognition
- Social media data, e.g., Twitter, Facebook, internal office network etc.

Analytics applied on passive and active sources collectively will provide a 360^0 view of network traffic, e.g., by singling out an abnormal behavior in the access pattern of a given user. Appropriate prevention techniques can then be applied, e.g., lock accounts, quarantine, modify network settings, multiple authentications, alert on an on-going fraud etc.

[1] Description of these solutions can be obtained from their respective websites.

B. Security Analytics Model

A security analytics model has the following primary characteristics [15]:

- **Diverse Data Sources:** Data can be acquired from many sources (as mentioned in (A)). The frequency and complexity of data sources should have no effect on the final BDA outputs.
- **Enterprise-level Data Warehouse:** All network security data should be stored in an enterprise warehouse, with network performance measures stored in fact tables, and query data in dimension tables (time, location, customer etc.). This is important as data mining (BDA) solutions are now strongly coupled with a warehouse, e.g., Oracle, SQL Server, DB2 etc. Different managers (requiring different analytics) can generate their own security analytics warehousing process.
- **ETL Tools:** These tools can help perform ETL at two levels, i.e., at a generic level over a complete data source, and at specific level over selected data in order to apply one or more BDA techniques on it. Some common names are Talend, Informatica, Pentaho etc.
- **BDA Engines:** These engines combine Hadoop, sophisticated hardware and analytics softwares to process real-time network streams. As mentioned above, these engines are typically coupled with a warehouse allowing network managers to employ warehouse resources, e.g., ETL, dashboards, data maintenance etc. for the BDA tasks as well.
- **Monitoring Systems:** These system monitor network traffic and compare it with behavior and risk models discovered by BDA, or those given by security experts based on their experience.
- **Advanced Security Controls:** These are used to impart appropriate security measures in real-time, e.g., additional user authentication, blocking highly suspicious transactions, calling customers while the transaction is in progress etc.
- **Interactive Dashboards:** These apply Business Intelligence techniques to display the real-time cybersecurity status using interactive tools, e.g., graphs, charts, tables etc. A lot of research work is being carried out to develop novel visualization techniques which are at par with BDA outputs [13].
- **Robust Security Infrastructure:** This should be able to facilitate efficient communications between different locations or regions along with efficient execution of queries using BDA architectures.
- **Big Data Integration Infrastructure:** This will facilitate integration of data from diverse number of data sources, which are quite common as mentioned in (A).

C. Security Analytics for Threat Detection

The largest application of security analytics is in threat monitoring and incident investigations, which is of major concern to both financial and defense institutions. The focus is on discovering and learning both known and unknown cyber attack patterns, which is expected to remarkably influence the efficiency of identifying hidden threats faster, track down attackers and predict future attacks with increasing accuracy (minimum false positive rate). Some examples of how BDA can help with respect to different security dimensions are as follows:

- **Network Traffic:** Detecting and predicting suspicious sources and destinations, along with abnormal traffic patterns.
- **Web Transactions:** Detecting and predicting abnormal user access patterns, particularly in the usage of critical resources or activities.
- **Network Servers:** Detection and predicting abnormal patterns related to server manipulation, e.g., abnormal or sudden configuration changes, non-compliance with pre-defined policy etc.
- **Network Source:** Detecting and predicting abnormal usage patterns of any machine, e.g., related to the type of data the source transmits, processes and receives.
- **User Credentials:** Detecting anomalies with respect to a user, or a group of user, *not* complying with its inherent access behavior, e.g., abnormal access time or transaction amount.

These activities are bringing a revolutionary change in the domains of security management, identity and access management, fraud prevention and governance, risk and compliance, e.g., through centralization of threat data and alert management system, correlating hundred thousands of network events per second, real-time continuous assessment of risk, distinction between legitimate and abnormal activities, along with appropriate prioritization of risks.

D. Steps for Implementing A Security Analytics Solution

BDA is a relatively novel field. The corporate sector in Pakistan is typically more tilted towards the use of traditional (simpler) analytical techniques as compared to advanced ones[2]. Hence, it could be some time before the Pakistani network companies embrace the BDA technology whole heartedly. A major reason for this trend is the uncertainty with respect to the usefulness of BDA outputs, as acquiring useful outputs is a strenuous recursive activity which requires appropriate BDA expertise (data scientists). There is also a lack of in-depth technical knowledge regarding basic BDA concepts, e.g., Hadoop, Predictive Analytics, Cluster Analysis etc. Finally, there are not many proprietary BDA cybersecurity solutions that are well-known in the cybersecurity world. With these

[2] This statement is derived from the first author's experience of training and consultancy with various Pakistani companies

limitations in mind, we propose the following steps to the corporations for adopting a security analytics solution [15]:

- **Develop Security Analytics Business Strategy:** The first step is to create a business strategy for implementing the analytics platform. For this, CIOs and CTOs need to initially build the domain knowledge by looking at some of the successful analytics-based security solutions (detailed in next section) to determine their feasibility, impact and value for their own organizations.
- **Participate in Analytics Trainings and Workshops:** The real impact of threat detection can only be realized by C-level executives if they understand the technical details of BDA. Hence, attending BDA workshops and training sessions is particularly recommended. Especially because several open-source analytical tools, e.g., Rapid Miner, allow users to experiment with a host of BDA techniques through simple drag-and-drop functionality. Hands-on experiments with these tools on security data sets can provide in-depth knowledge of the expected BDA outputs and aid in developing the analytics strategy.
- **Implement a Centralized Data Management Infrastructure:** This should allow seamless integration of network streams from various sources, along with tools to support ETL, stream processing, data warehousing, data storage and query execution mechanisms. The infrastructure should be flexible to extension as well as to modification, e.g., a change of ETL tools.
- **Implement an Analytics Platform:** This should support experimentation with a diverse number of BDA tools, techniques and algorithms. An efficient communication medium should be provided with the data repository to acquire the data and to store BDA outputs.
- **Hire Data Scientist as Consultant:** As acquiring BDA outputs is typically lengthy task requiring background knowledge of many domains, it is advised to hire a data scientist as a consultant who can guide the execution of the BDA process at different stages.
- **Implement a "Network Monitoring" Layer:** This layer monitors the network streams at run time by employing the set of mathematical models that have been mined by BDA. It can also employ any type of security knowledge given by the system designers based on their experience, or knowledge from some governmental database regarding new types of external threats. This layer should always be "live" as network streams are monitored 24/7. It intimidates the "Suspicion Alert" later if any suspicious behavior is detected.
- **Implement a "Suspicion Alert" Layer:** This layer implements all the measures that can be taken to ensure cybersecurity, e.g., alert authorities of an ongoing suspicious transaction, lock access, cross-check user identity etc. It must be ensured that this module is always "live", even if the analytics platform is not being used at a given time.
- **How to Streamline Analytics with Current Workflow?** This is the most important question facing C-level executives. Due to the uncertainty which accompanies BDA, it is advisable to complete the aforementioned steps within a minimum level of resources initially. These steps should also be implemented independently of the current workflow, with collaboration occurring only where needed. For instance, a team of 4-5 people along with the data scientist should be given 5-6 months to implement the analytics solution, as an organizational project. Once the managers are satisfied with the outputs, steps can be taken to infuse the analytics platform gradually in the current workflow.

E. Corporate Security Analytics Solutions

In this section, we list some of the important corporate cybersecurity solutions. The solution by Solera Networks [16] provides the following features:

- **Root Cause Analysis:** Ability to go back in time from any security event to discover the root cause, to determine that something happened, what exactly happened, and how it entered the network.
- **Pathway Analysis:** Ability to go forward in time from a security event to determine the full purport of this event in the future, e.g., expected path, damage etc.
- **Application Discovery:** Ability to discover suspicious *applications* on the network, allowing security systems to block applications based on an application ID.
- **Data Leakage Discovery:** Ability to provide the reasons behind data loss based on the context of the lost information
- **Insider Threat Analysis:** Ability to keep track of the activity of employees to foresee and detect reasons of possible attacks carried out by them.

Fortscale [17] is a high-tech startup by a group of Israeli entrepreneurs whose security analytics solution has the following features:

- Discovery of targeted and high-level cyber attacks
- Detection of malicious employees who take advantage of their legitimate permissions to leak information
- Real-time Analytics of access to sensitive data
- Verification of security alerts
- Identify compromised machines by sophisticated malware.

IBM has also developed a solution called IBM Security Intelligence [18] which has the following features:

- Real-time abnormality detection of diverse security data

- High-speed querying of security intelligence data
- Flexible BDA across all types of data sources
- Interactive visualization dashboards for exploring BDA outputs
- Network forensics (the measurement of network performance variables)

Finally, Actian has implemented the Actian DataRush [19] solution which has the following features:

- Monitoring up to 1 million events per second
- Dynamic scaling of big data architectures
- Detection of abnormal patterns, event correlations, intrusion detection, threat and vulnerability assessment, discovery of security events

The aforementioned features are novel and are also being offered by other names include Juniper [20], RSA [21], LogRythm [22] and Blue Coat [23].

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have highlighted the state of the art issues facing the cybersecurity domain in the face of big data. Traditional security solutions are not capable anymore of encompassing the real-time big data network streams using traditions tools and techniques. We have shown how Security Analytics – the application of Big Data Analytics techniques to derive actionable intelligence and insights from streams in real-time – is rapidly becoming a strong need for cybersecurity setups. Although the current adoption of analytical solutions is by no means revolutionary (as shown by this study by Teradata [24]), the awareness of adoption is increasing rapidly. To support this cause, in the paper, we specifically mention the different types of big data sources for such an analytics solution, the primary components of a security analytics model, some examples of security analytics outputs, steps for implementing a security analytics solution, and finally, the corporates which are offering analytics solutions for cybersecurity along with their features. We are currently applying security analytics techniques on network-related datasets of our institute in order to try and decrease the false positive rate of a predictive model for fraudulent transactions.

REFERENCES

[1] MessageLabs Intelligence: 2010 Annual Security Report, Symantec

[2] M.T. Banday and J.A. Qadri, "SPAM – Technological and Legal Aspects", Kashmir University Law Review, Vol. 8, No. 8, 2006,

[3] B. Whitworth, "Spam and the social technical gap," IEEE Computer, vol. 37, no. 10, pp. 38-45, Oct. 2004.

[4] L. Lu, R. Perdisci, and W. Lee, "SURF: Detecting and Measuring Search Poisoning", CCS'11, October 2011, Illinois, USA.

[5] L. Vaas, "Malware poisoning results for innocent searches", 27th November, 2007, http://www.eweek.com/c/a/Security/Malware-Poisoning-Results-for-Innocent-Searches

[6] B. Stone-Grass, M. Cova, L. Cavallaro, B. Gilbert and M. Szydlowski, "Your Botnet is My Botnet: Analysis of a Botnet Takeover", CCS'09, November 2009, Illinois, USA

[7] M. Bailey, E. Cooke, F. Jahanian, Y. Xu and M. Karir, "A Survey of Botnet Technology and Defenses", CATCH'09, Cybersecurity Applications and Technology, March 2009, Washingtin DC, USA

[8] Q. Gu and P. Liu, "Denial of Service Attacks", Technical Report, http://s2.ist.psu.edu/paper/DDoS-Chap-Gu-June-07.pdf

[9] M. Jakobsson and S. Myers, "Phishing and Countermeasures: Uunderstanding the Increasing Problem of Electronic Identity Theft". John Wiley & Sons, Inc., 2007.

[10] J. Shi and S. Saleem, "Phishing", Technical Report, http://www.cs.arizona.edu/~collberg/Teaching/466-566/2012/Resources/presentations/2012/topic5-final/report.pdf

[11] Symantec, http://www.symantec.com/index.jsp

[12] Computer Security, Wikipedia, http://en.wikipedia.org/wiki/Computer_security, Last updated on 2nd December, 2013.

[13] A. Sathi, Big Data Analytics: Disruptive Technologies for Changing the Game, Mc Press, 1st Edition, February 5, 2013

[14] Data to Decisions, 26th September 2013, http://data-to-decisions.com/

[15] S. Curry, E. Kirda, E. Schwartz, W. H. Stewart, and A. Yoran, "Big Data Fuels Intelligence-Driven Security", RSA Security Brief, January, 2013

[16] Solera Networks, http://www.soleranetworks.com/about/security-analytics/

[17] Fortscale, www.fortscale.com

[18] IBM Security Intelligence with Big Data, http://www-03.ibm.com/security/solution/intelligence-big-data/

[19] Cyber Security Analytics. Prevent Intrusion Before Its Too Late: http://bigdata.pervasive.com/Solutions/Cyber-Security.aspx

[20] Secure Analytics, http://www.juniper.net/us/en/products-services/security/secure-analytics/

[21] RSA Security Analytics, http://www.emc.com/security/security-analytics/security-analytics.htm

[22] LogRythm Security Analytics: http://logrhythm.com/siem-2.0/the-platform-for-security-analytics/logrhythm-security-analytics.aspx

[23] Security Analytics Platform: Analyze Actualize, http://www.bluecoat.com/products/security-analytics-platform

[24] Teradata and Ponemon Institute, Big Data Analytics in Cyber Defense, February 2013, http://www.ponemon.org/local/upload/file/Big_Data_Analytics_in_Cyber_Defense_V12.pdf