Zuech *et al. Journal of Big Data* (2015) 2:3
DOI 10.1186/s40537-015-0013-4

Journal of Big Data
a SpringerOpen Journal

SURVEY PAPER **Open Access**

Intrusion detection and Big Heterogeneous Data: a Survey

Richard Zuech[*], Taghi M Khoshgoftaar and Randall Wald

[*]Correspondence: rzuech@fau.edu
Florida Atlantic University, 777
Glades Road, Boca Raton, FL, USA

Abstract

Intrusion Detection has been heavily studied in both industry and academia, but cybersecurity analysts still desire much more alert accuracy and overall threat analysis in order to secure their systems within cyberspace. Improvements to Intrusion Detection could be achieved by embracing a more comprehensive approach in monitoring security events from many different heterogeneous sources. Correlating security events from heterogeneous sources can grant a more holistic view and greater situational awareness of cyber threats. One problem with this approach is that currently, even a single event source (e.g., network traffic) can experience Big Data challenges when considered alone. Attempts to use more heterogeneous data sources pose an even greater Big Data challenge. Big Data technologies for Intrusion Detection can help solve these Big Heterogeneous Data challenges. In this paper, we review the scope of works considering the problem of heterogeneous data and in particular Big Heterogeneous Data. We discuss the specific issues of Data Fusion, Heterogeneous Intrusion Detection Architectures, and Security Information and Event Management (SIEM) systems, as well as presenting areas where more research opportunities exist. Overall, both cyber threat analysis and cyber intelligence could be enhanced by correlating security events across many diverse heterogeneous sources.

Keywords: Intrusion detection; Big data; Security; IDS; SIEM; Data fusion; Heterogeneous; Hadoop; Cloud; Feature selection; Situational awareness; Big Heterogeneous Data

Introduction

Cybersecurity is critical as society becomes increasingly dependent on computerized systems for its finances, industry, medicine, and other important aspects. One of the most important considerations in cybersecurity is Intrusion Detection. In order to mitigate or prevent attacks, awareness of an attack is essential to being able to react and defend against attackers. Cyber Defenses can be further improved by utilizing Security Analytics and Intrusion Detection data to look for hidden attack patterns and trends. Intrusion Detection is also important for forensic purposes in order to identify successful breaches even after they have occurred. For example, it is important to know afterwards if information such as credit card data has already been stolen, in order to take additional precautions or possibly take law enforcement or legal actions. Intrusion Detection can also be very helpful beyond detecting cyber-attacks in noticing abnormal system behavior to detect accidents or undesired conditions. For example, an Intrusion Detection System

(IDS) could report anomalies where a malfunction or human error is causing customer credit card numbers to be erroneously charged multiple times. Or perhaps an IDS could alert on something out of the ordinary and detect a gas leak, and help prevent an explosion which could harm or even kill humans. Intrusion Detection can be helpful in providing early warnings and minimizing damage.

This study evaluates some of the advancements in Intrusion Detection technology along with important considerations like monitoring a wide array of heterogeneous security event sources. As cyber-attacks have evolved and grown in sophistication, Intrusion Detection products have also become much more sophisticated, monitoring an ever increasing amount of diverse heterogeneous security event sources. IDSs were the first specialized products developed to detect and alert for potential cyber-attacks, and they can either employ misuse detection or anomaly detection. An IDS utilizing misuse detection evaluates data it is monitoring against a database of known attack signatures to determine attack matches. An IDS utilizing anomaly detection, on the other hand, evaluates data it is monitoring against a normal baseline, and can issue alerts based on abnormal behavior.

One traditional IDS product is a Network Intrusion Detection System (NIDS) which monitors for cyber threats at the network layer by evaluating network traffic. Another traditional IDS product is a Host-based Intrusion Detection System (HIDS) which monitors for cyber threats directly on the computer hosts by monitoring a computer host's system logs, system processes, files, or network interface. An IDS can monitor specific protocols like a web server's Hyper Text Transfer Protocol (HTTP); this type of IDS is called a Protocol-based Intrusion Detection System (PIDS). IDSs can also be specialized to monitor application-specific protocols like an Application Protocol-based Intrusion Detection System (APIDS). An example for this could be an APIDS that monitors a database's Structured Query Language (SQL) protocol. Similar to the heterogeneity of the security event sources such as network and diverse host types, the IDSs themselves can be heterogeneous in their type, how they operate, and in their diverse alert output formats.

Today's Information Technology (IT) security systems and personnel can be inundated with an overload of ambiguous information or false alarms, and the cybersecurity domain frequently encounters problems dealing with Big Data from currently implemented systems. Compounding the problem further, existing IT security systems seldom integrate across a wide spectrum of an organization's information systems. For example, an organization can typically have the following systems: Firewalls, IDSs, computer workstations, Anti-virus software, Databases, end-user Applications, and a variety of other systems. However with traditional IDSs there is rarely any integration among them in the context of monitoring for security breach attempts, and very seldom is there any sort of integrated security monitoring approach across a large proportion of an organization's information systems. A basic illustration of what this paper evaluates is given in Figure 1, where security events from most (if not all) of an organization's computing assets are being monitored. This diagram exhibits the heterogeneity of a typical enterprise's network where security events from different workstations, servers, NIDSs, HIDs, firewall events, etc. can all be very different. For example, an organization might use different NIDS solutions to increase detection accuracy, and increase the heterogeneity of a single function in the security system. To improve Intrusion Detection these security events should be correlated with each other in order to improve alerting

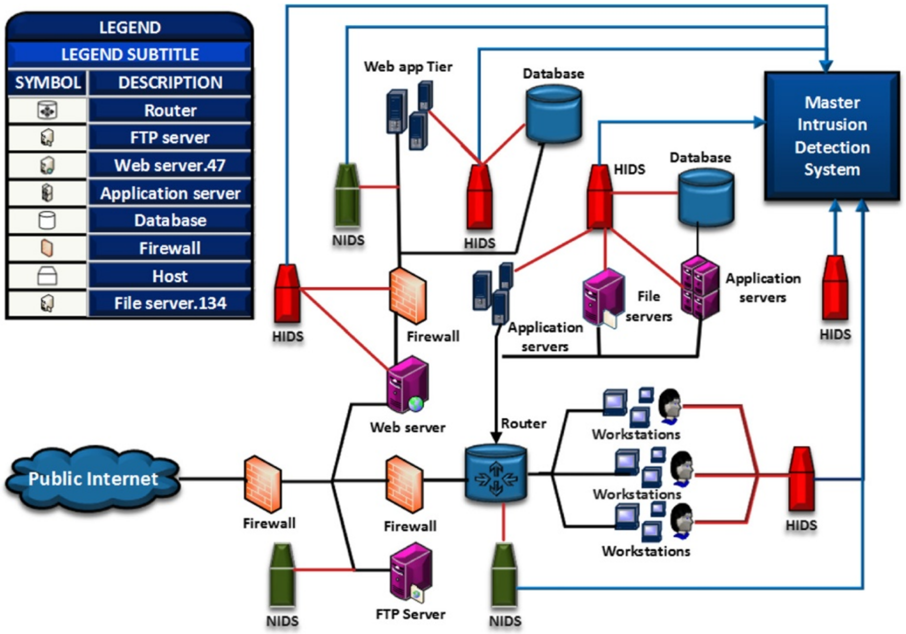

Figure 1 Illustration of Monitoring Heterogeneous Sources.

accuracy as well as give a more comprehensive overview of cyber threats from an overall perspective.

Intrusion Detection frequently involves analysis of Big Data, which is defined as research problems where mainstream computing technologies cannot handle the quantity of data. Even a single security event source such as network traffic data can cause Big Data challenges. According to Nassar et al. [1], merely 1Gbps of sustained network traffic can cause Big Data challenges for Intrusion Detection while using deep packet inspection. Another Big Data challenge that larger organizations can face is having an incredible amount of host log event data. The Cloud Security Alliance reported [2] that in 2013, it is estimated that an enterprise like HP can "generate 1 trillion events per day or roughly 12 million events per second". They report that such large volumes of data are "overwhelming" and they even struggle to simply store the data. Enterprises dealing with such Big Data issues at this scale cannot use existing analytical techniques effectively, and so false alarms are especially problematic. Additionally, it can be very difficult to correlate events over such large amounts of data, especially when that data can be stored in many different formats. Relational database technology can commonly become a bottleneck in Big Data challenges. For example, commercial SIEMs that use relational database technologies for their storage repositories will find the databases becoming bottlenecks in deployments at larger enterprises: storage and retrieval of data begins to take longer than is acceptable. Zions Bancorporation conducted a case study [3] where it would take their traditional SIEM systems between 20 minutes to an hour to query a month's worth of security data, however when using tools with Hadoop technology it would only take about one minute to achieve the same results. It is a clear sign that Intrusion Detection is facing Big Data challenges when a mainstream technology like relational databases becomes a bottleneck. Next generational Big Data storage technologies like Hadoop can help address these problems.

While traditional Intrusion Detection Systems (IDSs) are a critical component of Intrusion Detection, more focus should be placed on gathering security data from a wider variety of heterogeneous sources and correlating events across them to gain better situational awareness and holistic comprehension of cybersecurity. Analyzing security data across heterogeneous sources can be difficult for Intrusion Detection where homogeneous sources already face Big Data challenges. By analyzing additional heterogeneous sources, the problem can be compounded into a more significant Big Heterogeneous Data challenge as each source can potentially have Big Data. Improving situational awareness by correlating security events or alert data across heterogeneous sources where each can have Big Data challenges is a much more significant problem than performing Intrusion Detection independently on each homogeneous Big Data source, and this is the Big Heterogeneous Data challenge for Intrusion Detection.

A larger IT infrastructure can cause Big Heterogeneous Data challenges with its diversity of input event sources such as various hosts. Correlating among diverse sources like workstations, various application servers, and the network can be a significant problem when facing Big Data challenges. Compounding the problem further is that both the security alerting devices (e.g., IDSs, SIEMS, etc) as well as alert messages can be heterogeneous in nature. The typical enterprise can have a myriad of different security products which do not integrate well, and this heterogeneity causes difficulty for Intrusion Detection. Gartner Research Director Lawrence Pingree addresses this difficulty with a concept called "intelligence awareness" which is the capability of automated intelligence sharing and alerting across a myriad of security systems, and further explains that security systems must become "adaptable based on contextual awareness, situational awareness and controls themselves can inform each other and perform policy enforcement based on degrees or gradients of threat and trust levels" [4]. Ed Billis, CEO of Risk I/O further elaborates on this problem where security products are silo'ed from each other: "SIEMs weren't originally designed to consume much more than syslog or netflow information with a few exceptions around configuration or vulnerability assessment. Security analytics is more than just big data – it's also diverse data. This causes serious technical architectural limitations that aren't easy to overcome with just SIEM" [5].

In addition, industrial processes should also be monitored for Intrusion Detection as industrial systems are increasingly computerized. One example is the nation's electrical grid where most equipment has been computerized that is used to monitor the real-world physical sensors that measure electrical properties like power, voltage, and current. Being computerized, they should be monitored for Intrusion Detection as well. However, the overall Intrusion Detection system can also enhance its capabilities by considering abnormal operational electrical readings and even correlating those real-world events to security events in cyberspace, thus further enhancing situational awareness. Kezunovic et al. [6] discuss the role of Big Data in the electric power industry, and IBM [7] further describes the Big Data challenges faced in this industry along with the need for security monitoring. Clearly, all this Big Data must be monitored in the context of Intrusion Detection. Since real-world physical sensors from the electrical grid can generate Big Data separately, these sensors from the physical world constitute another heterogeneous data source beyond cyberspace and contribute another dimension of heterogeneity as an input to Big Heterogeneous Data. Other industrial applications and processes have increasingly been computerized, with their real-world physical sensors also having Big Data.

They can enhance their overall situational awareness by utilizing those physical sensors as inputs into their Intrusion Detection architecture. When doing so, organizations should be aware of the Big Heterogeneous Data challenges they will face.

Even though there have been other survey papers on the Intrusion Detection topic, our paper is unique compared to these prior surveys. We focus on improving intrusion detection from the perspective of aggregating security sensor data from systems and devices which exhibit a great deal of heterogeneity. At the same time, we consider the fundamental Big Data problems that are inherent with such forms of heterogeneous security data. One survey by Modi et al. [8] is especially relevant when considering our work as they focus on Intrusion Detection in the Cloud. Their work does a fantastic job of describing the great deal of heterogeneity of security data and systems encountered in the cloud, and this is increasingly relevant as cloud computing becomes more pervasive and presents more Big Data challenges. Another survey by Zhou et al. [9] is also relevant to ours as they consider heterogeneous architectures for IDSs which collaborate in teams to improve detection accuracy, but they do not consider the Big Data ramifications.

While this paper covers a large variety of issues, there are two main themes of this survey:

1. Cybersecurity Data across Heterogeneous Sources.
2. Big Heterogeneous Data for Intrusion Detection.

The remainder of this paper is presented as follows: The INTRUSION DETECTION AND BIG DATA BACKGROUND section presents a background on Intrusion Detection and some Big Data implications and challenges. The SECURITY DATA ACROSS HETEROGENEOUS SOURCES section covers Security Data across Heterogeneous Sources. The BIG HETEROGENEOUS DATA FOR INTRUSION DETECTION section discusses Big Heterogeneous Data for Intrusion Detection. The DISCUSSION section provides further discussion and insights about the issues covered. Finally, the CONCLUSION section concludes the work presented in this paper.

Intrusion detection and big data background

The purpose of this section is to briefly give a general background on Big Data, as well as insight into Big Data challenges facing Intrusion Detection. Some background information is also provided with regards to challenges in security learning such as utilizing publicly available data sets and feature selection. Finally, some examples are provided illustrating how Big Data technologies can be utilized to address Big Data challenges in Intrusion Detection.

Big Data is typically defined in terms of 3Vs, a designation originally developed by Gartner analyst Doug Laney [10] in 2001: Volume, Velocity, and Variety. Volume refers to the amount of data, and there certainly can be a Big Data challenge when large amounts of data pose challenges to processing with traditional computing or techniques (which is also referred to as “Big Volume”). Velocity refers to the speed at which data is processed, and there can be a Big Data challenge when the rate of data is moving too quickly to process with traditional computing or techniques (which is also referred to as “Big Velocity”). Variety refers to the complexity of the data, and there can be a Big Data challenge when the data includes complex problems such as high dimensionality, data from many sources, or data having many different data structures: all of these problems can cause difficulty in processing with traditional computing or techniques (which is also referred to as "Big Variety"). There are many other definitions of Big Data, such as the 5Vs defined by Zikopoulous [11] that adds Veracity and Value to the already existing 3Vs of Volume, Velocity, and Variety. Veracity accounts for the correctness of the data, and can include data quality problems such as noise or missing values (which is also referred to as "Big Veracity"). Value accounts for Big Data in the sense that if particular data does not provide significance (value), it is not relevant for Big Data analysis (which is also referred to as "Big Value"). However for simplicity, Big Data can just be summarized as any time current mainstream computing or techniques cannot process data effectively.

For Intrusion Detection, Big Data is currently a major challenge and has been a prevailing theme for quite some time. In 1994, a study by Frank [12] for Intrusion Detection focusing on data reduction and classification found: "a user typically generates between 3 - 35 Megabytes of data in an eight hour period and it can take several hours to analyze a single hour's worth of data". They further suggested that filtering, clustering, and feature selection on the data is "important if real-time detection is desired," which can improve detection accuracy. This example indicates that Intrusion Detection has been facing Big Data challenges long before the "Big Data" term was introduced.

While a more comprehensive security monitoring system across heterogeneous systems could improve security, it would further exacerbate the Big Data challenge for Intrusion Detection which is already present in isolated systems. Integrating across more security sensors would increase Big Data issues in terms of: Volume in having to store more information collectively, Velocity in that more information would be flowing collectively at a higher rate in and out of the monitoring system, and especially Variety in terms of many different types of information coming from very different sources and also collectively yielding higher dimensionality.

A more comprehensive approach for monitoring a myriad of diverse heterogeneous event sources for Intrusion Detection can yield a better situational awareness of the threats in cyberspace, and thus improve detection accuracy and minimize false alarms by correlating security events among these diverse sources. Experiments have indicated that embracing a more diverse heterogeneous approach to Intrusion Detection does yield better situational awareness and improve accuracy. However, Big Data challenges already exist in some of the individual sources, and when they are aggregated the existing Big Data problem is compounded into a more significant Big Heterogeneous Data problem. When Big Data challenges are already present in any of the underlying inputs or outputs for Intrusion Detection, the overall system will likely experience Big Data challenges as well unless the Big Data bottleneck is eliminated. One way to remove this Big Data challenge is by filtering out (removing) the Big Data from a subsystem. However this is not ideal if valuable information is lost. New techniques or Big Data technologies can alleviate the challenges and costs that Big Data impose for Intrusion Detection.

Big Heterogeneous Data definitions

When Big Data is present in heterogeneous forms, it can be considered Big Heterogeneous Data regardless of whether that data is input(s) or output(s) of the system. For example, this can arise due to the additive properties of Big Data. If one input is deemed

Big Data and is added to another input which is not Big Data, the result will still be Big Data. This can be shown in Equation 1 below:

$$BD(\text{“}BigData''') + NBD(\text{“}NotBigData''') = BD(\text{“}BigData''') \tag{1}$$

Similarly if some advanced data correlation (or data fusion which is presented in the SECURITY DATA ACROSS HETEROGENEOUS SOURCES section) for analysis is occurring and the Big Data is being combined with "Not Big Data" in a multiplicative manner, the result will still be Big Data. This can be shown in Equation 2 below (assuming "Not Big Data" is greater than one):

$$BD(\text{“}BigData''') \times NBD(\text{“}NotBigData''') = BD(\text{“}BigData''') \tag{2}$$

Therefore, when Big Data is being combined with other data that is not classified as Big Data, the result will still be Big Data.

Another important consideration is that Big Data Challenges can quickly escalate into a significantly larger Big Data problem when combining multiple heterogeneous sources for analysis where each of the sources can have Big Data challenges individually. An example of this would be if two or more heterogeneous sources which separately contain Big Data challenges individually were then analyzed with advanced data correlation techniques (or data fusion which is presented in the SECURITY DATA ACROSS HETEROGENEOUS SOURCES section) in order to give better accuracy through superior situational awareness. For complex systems such as Intrusion Detection where a large amount of heterogeneous sources are common and can contain Big Data challenges, the problem can quickly escalate into a more difficult Big Heterogeneous Data challenge. This can be shown in Equation 3 below (where n refers to the number of heterogeneous data sources that contain Big Data, and n >1):

$$BHD(\text{“}BigHeterogeneousData''') = \prod_{i=1}^{n} BHDSi(\text{“}BigHeterogenousDataSourcei''') \tag{3}$$

The above generalizations do not always apply and even if parts of the system (e.g., a subsystem) contains Big Data challenges, these do not always propagate throughout the rest of the system. Big Data can be effectively removed in one or more of the subsystems by filtering (removal), and then the Big Data would not necessarily propagate throughout the rest of the system. This is not always an ideal approach if the Big Data being filtered out contains value, but it is still necessary at times if retaining the Big Data is too costly. An example for this would be if netflow traffic was analyzed for a NIDS instead of deep packet inspection. The deep packet inspection will yield superior detection accuracy. However the cost may be prohibitive in doing so. Another example might be the time retention policy for very detailed forensic data, where costs can prevent this Big Data from being stored indefinitely. This is illustrated in Equation 4 below (where the subtraction operator is essentially filtering or removal of the Big Data):

$$BD(\text{“}BigData''') - BD(\text{“}BigData''') = NBD(\text{“}NotBigData''') \tag{4}$$

As the above scenario is a cost and benefit tradeoff, Big Data challenges can also be removed by some of the following "Big Data Handlers": Big Data technologies, natural technology evolution (e.g., Storage or Processing evolutions such as Moore's Law), or novel techniques (or new approaches). For example, if it is desired to retain forensic data longer and a "Big Data Handler" technology like Hadoop permits this to be performed in a cost permissible fashion, then the "Big Data Challenge" can be removed and the "Handled Big Data" can be retained in a manner that is within cost constraints. This is illustrated in Equation 5 below (where the addition operator is enabling the Big Data to be handled in a cost effective manner):

$$BDC(\text{``}BigDataChallenge''') + BDH(\text{``}BigDataHandler''') = HBD(\text{``}HandledBigData''') \tag{5}$$

Essentially, when dealing with Big Data challenges and heterogeneous inputs or outputs, the resulting data will still be Big Data if the Big Data Challenge is not eliminated in some way. For example, if there was a "Big Data Challenge" like a particular data source that had very high dimensionality and if a "Big Data Handler" like feature selection could be effectively used to create "Handled Big Data", then the "Big Data Challenge" will either remain or be eliminated depending on whether an effective "Big Data Handler" was used. In this case, if feature selection was effectively employed as a "Big Data Handler", then the "Big Data Challenge" would be removed and we would have "Handled Big Data". Likewise, if feature selection was not effective (or used), then the "Big Data Challenge" would remain and we were not able to effectively handle the particular challenges from Big Data.

Accordingly, when considering Big Data with heterogeneous sources or outputs, it can be better described as Big Heterogeneous Data so long as the Big Data Challenges are not eliminated in some way. The reason Big Heterogeneous Data is a more descriptive term is because typically the Big Data Challenges will be even more pronounced (i.e., magnified) when dealing with extreme heterogeneity in the input(s) or output(s) of Intrusion Detection Big Data within cyberspace. The BIG HETEROGENEOUS DATA FOR INTRUSION DETECTION section will further elaborate on Big Heterogeneous Data in terms of inputs and output categories now that the rationale behind the Big Heterogeneous Data terminology has been presented. This Big Heterogeneous Data challenge can become more pronounced while attempting to enhance Intrusion Detection through superior situational awareness by adopting more heterogeneity in the inputs, outputs, and architectural components as mentioned throughout this survey.

Important considerations for intrusion detection

With a more comprehensive security monitoring system, improvements to computer security do not need to be restricted to merely detecting security intrusions. Such a system could be extended to actually prevent security intrusions by integrating with technologies such as Intrusion Prevention Systems (IPSs), and embracing more of a "Defense in Depth" strategy [13]. Naturally, an IPS would require close to real-time detection. Note that this study is not limited to Intrusion Detection with the "real-time" distinction, and also includes offline forensic and security analytic capabilities.

This survey is not similar to previous Intrusion Detection surveys in that it evaluates the Intrusion Detection problem with an emphasis on aggregating security sensor data across many different systems and devices with the motivation of further improving security alerting accuracy. In addition, we consider the Big Data issues that arise when handling such forms of heterogeneous security data.

In the earlier days of computing, security monitoring was primarily performed by system administrators checking the log files of their servers. Then in the 1980s, the concept of the Intrusion Detection System was introduced, where a separate monitoring device would look for suspicious behavior at the network or computer host level. Denning [14] produced what many consider to be the first landmark research paper for Intrusion Detection System (IDS) research back in 1987. A good example of an IDS is the common and widely known open source IDS called Snort [15,16].

Intrusion Detection is a very active research area with important implications. The Center for Strategic and International Studies and McAfee conducted a study [17] and analyzed monetary losses from cybercrime and cyber espionage: “for the US, for example, our best guess is that losses may reach $100 billion annually.” They approximate these global losses to be about $300 billion annually. In 2012 through more than 250 client engagements, the Verizon RISK Team [18] found over 47,000 confirmed security incidents, with 92% of data breaches perpetuated by outsiders in their engagements with clients. In a study [19] by the Ponemon Institute, for the FY 2012 it was determined that the most expensive cybercrime category was “Detection” costing 26% (followed by: Recovery, Ex-post Response, Containment, Investigation, Incident management). These studies clearly demonstrate that cybersecurity (and specifically Intrusion Detection) have significant economic impact.

Julisch and Dacier [20] discuss how Intrusion Detection can have many false alarms: “IDSs can easily trigger thousands of alarms per day, up to 99% of which are false positives.” It is not uncommon for security analysts to grow numb to a flood of meaningless false alarms. Xu and Ning [21] state that in terms of detection rate, IDSs are typically not completely accurate, and can have an unacceptable number of False Negatives (“may miss some attacks”).

Intrusion Detection is inherently a Big Data problem according to Suthaharan and Panchagnula [22]: “However the biggest challenge is the ‘Big-Data’ problem associated with the large amount of network traffic data collected dynamically in the intrusion detection dataset”. Bhatti et al. [23] discuss how even current technologies cannot cope well with the Big Data challenges of Intrusion Detection: “Security analytics in a big data environment presents a unique set of challenges, not properly addressed by the existing security incident and event monitoring (or SIEM) systems that typically work with a limited set of traditional data sources (firewall, IDS, etc.) in an enterprise network”. From a study by Enterprise Strategy Group at the end of 2012, Olsten [24] discusses that: “44% of enterprise organizations consider their security analytics ‘big data’ today, while another 44% believe that their security analytics requirements will be regarded as ‘big data’ within the next two years”. Clearly, Intrusion Detection can be a Big Data challenge.

Challenges of machine learning in cybersecurity

This section addresses some issues found with Intrusion Detection data set challenges and feature selection. This background is very relevant for Intrusion Detection in general especially considering the widespread criticism of the publicly available data sets, and yet the majority of the research uses these criticized data sets. Accordingly, it is important that the reader understands that many of the experiments considered in this paper suffer the same criticisms when using an experiment with data sets (unless otherwise noted). Some researchers openly admit that the underlying data sets themselves have inherent flaws. Nonetheless, researchers continue to use them because essentially it is the best they have to work with.

A brief background is also given on feature selection and its application to Intrusion Detection data sets. This too can be important for considering experiments with data sets for Intrusion Detection, and especially more so when utilizing data sets from a multitude of diverse heterogeneous sources. The reason for this is the landscape for cybersecurity can change extremely rapidly, and accordingly can undermine the stability of the feature selection sets in general and especially so for real-time Intrusion Detection. These issues will also be evaluated from the perspective of Big Data.

Intrusion detection data set challenges

Many authors discuss the problems with existing public Intrusion Detection data sets: for example, Sommer and Paxson [25] give an excellent summary regarding some of the underlying reasons why this is a significant problem. The cybersecurity landscape has changed significantly over the last decade, and many don't even consider experiments that use older data sets to even be relevant today according to Sommer and Paxson. Unfortunately, organizations can be reluctant or even legally constrained from divulging sensitive data that these types of data sets can contain, and Coull et al. [26] note that attempts to anonymize sensitive data are not always effective. In addition, lab simulation of real-world network traffic to generate data sets is often not very realistic. As shown in a survey by Azad and Jha [27], the two most popular data sets used for research in Intrusion Detection are DARPA and KDD Cup where out of the 75 studies discussed, 46 used one of these data sets while only 29 chose a different one. This is disappointing because the DARPA and KDD Cup data sets are over a decade old, and even recent studies still frequently use them. The cybersecurity landscape has changed significantly over the last decade, and many don't consider experiments that use those data sets to be relevant today [25]. To make matters even worse, it became widely known shortly after the initial release of those data sets that they contained inherent flaws, as discussed by McHugh [28] and Mahoney and Chan [29]. Nonetheless, the DARPA and KDD data sets are still widely used even today. Some of the more commonly used data sets can be seen in Table 1.

These flawed public data sets lack Veracity from the perspective of Big Data, and so they would not be relevant as a consequence of having poor quality. Due to this low Veracity, these data sets would also lack Value as well, further reducing their relevancy.

Table 1 Summary of popular datasets in the intrusion detection domain [30]

Data source	Dataset name	Abbreviation
Network Traffic	DARPA 1998 TCPDump Files	DARPA98
	DARPA 1999 TCPDump Files	DARPA99
	KDD99 Dataset	KDD99
	10% KDD99 Dataset	KDD99-10
	Internet Exploration Shootout Dataset	IES
User behavior	Unix User Dataset	UNIXDS
System call sequences	DARPA 1998 BSM Files	BSM 98
	DARPA 1998 BSM Files	BSM 99
	University of New Mexico Dataset	UNM

A few other data sets of note are: ISCX [31], MAWI [32], NSA Data Capture [33], and the Internet Storm Center [34] (which also hosts the dshield.org data set). However, these more recent data sets are not used as frequently as the DARPA and KDD data sets, even in recent studies.

Song et al. took an interesting approach [35] in building their own data set by using honeypot data. A honeypot is a system that is not completely patched in order to draw attention from attackers. They also placed a machine in the network to generate normal traffic, and so any activity related to that machine was labeled as normal (it did not receive much attack traffic) while all traffic related to the honeypots were labeled as an attack. Overall there were approximately 93,000,000 total sessions generated, with about 50,000,000 being normal sessions and the remainder being attack sessions. Also of interest, was that their IDSs and anti-virus did not successfully classify about 426,000 of the sessions as attacks, even though they were able to more correctly classify them as attacks upon deeper inspection of the shellcodes. Because this data did not come from an actual client or pertain to ongoing business efforts, Song et al. did not need to worry about the sensitivity of the data. Also, their data set is roughly balanced between the classes of normal and attack.

While there are some drawbacks to this approach (for example, the normal class could be considered too "simulated"), it shows good promise for future work. More researchers could take this or a similar approach instead of continuing to use datasets that are not very relevant from over a decade ago. Another option for researchers to generate more adequate data sets might be for them to actually launch attacks in honeypot networks with simulated normal traffic, or possibly even do so in a real-world environment if they can properly sanitize sensitive data or ensure the absence of sensitive data in the first place. Generating data sets at even larger scales with honeypots could also lead to Big Data challenges in terms of Volume, Variety, and even Velocity in having to accommodate such large amounts, speed, and variety of Intrusion Detection data.

Intrusion detection and feature selection opportunities

Feature selection is an important technique in addressing Big Data challenges posed by Intrusion Detection, and when applied properly it can significantly improve classification processing times. In some cases, it can even improve classification accuracy by removing misleading noise. However, one should take caution in how feature selection is applied and especially with regards to research studies versus real-world application in terms of both relevancy and efficiency.

Many studies (even recent studies) are essentially using static data sets (i.e., DARPA, KDD, etc.) in the sense that the labeled instances might not anticipate newer real-world attacks such as "zero-day exploits", and this is especially important in the domain of cybersecurity because it is inherently a dynamically changing landscape. Newer attacks can be significantly more diverse than old attacks in terms of both technical implementation as well as the underlying methods themselves in the ongoing arms race between attackers and defenders. So in terms of feature selection in Intrusion Detection, yesterday's selected features from a static data set might not be relevant for tomorrow's dynamically different data set. A new attack class can make different features important, and different feature sets may or may not be relevant even at the millisecond scale. Thus, it is important to rethink the relevancy of feature sets from data sets that are older, lack diversity, or are very static.

A smaller number of relevant features will improve classification processing times from an efficiency standpoint. However, the process of performing feature selection in some cases can take a considerable amount of computation time. Applying feature selection to a data set where its attributes can change both rapidly and diversely might not be able to generate feature sets in close to real-time. In certain scenarios where generating feature sets takes too long for the effectiveness of the system, perhaps generating feature selection sets could be delegated to an offline process similar to what Bass proposed for offline Data Mining [36]. Those feature selection sets could be applied by the Intrusion Detection templates which are then used at the various sensors. In this manner, some static and stable feature sets could be used for Intrusion Detection. However, it cannot be assumed that all feature selection sets will be stable, especially when they are built from a myriad of heterogeneous sources in a constantly evolving and hostile environment where the diversity in attributes of data sets can vary considerably.

Wang et al. [37] conducted an experiment employing feature selection, specifically to address the so-called "dimensional disaster" problem which often prevents multi-sensor fusion from being applied to Network Security. In their experiment the original number of features was 84, which took 2.66 seconds of CPU time to process the test set. When they reduced the number of features to 37, it only took 1.54 seconds to process the test set. While they only assessed network data (some classification errors were higher or lower depending on the attack type), they asserted that fusing from other heterogeneous sources such as a host log could be beneficial. Perhaps different feature selection techniques could have further reduced the number of features without significantly sacrificing classification accuracy.

Tsang et al. use a MOGFIDS (fuzzy-logic based) feature selection technique on KDD-Cup99 [38] and achieved the best overall feature selection results as compared to eleven other techniques in terms of classification accuracy. Chebrulu et al. [39] present an ensemble approach of feature selection and are able to achieve higher classification accuracies with the combination of feature selection techniques versus using each technique independently, and in their case they did improve overall Intrusion Detection accuracy for all attack categories and the normal class of the DARPA dataset by using an ensemble of Bayesian Networks and Classification and Regression Tress. Chen et al. [40] show that classification times can be sometimes be reduced in half when using SVM and C4.5 feature selection techniques on the KDD datasets, and they also evaluate Random Forest (RF) as a feature selection technique but they do not provide classification times for all the features of RF to compare its performance of classification times. Elngar et al. [41] use a Particle Swarm Optimization (PSO-Discretize-HNB) technique which uses feature selection to reduce the feature set size from 41 to 11 features, and with the smaller feature set Detection Accuracy improved from 97.7

Clearly feature selection can be beneficial with Intrusion Detection. However care must be taken in its application, as the nature of attack threats changes, so can the data. Correspondingly, the feature sets also change. Also similar to other domains, feature selection can be used to address Big Data challenges. Feature Selection can help reduce the dimensionality of the data being processed with Intrusion Detection, and it has the potential to mitigate Big Variety challenges simply through reduction of certain features.

However, feature selection can also introduce additional Big Variety challenges when feature sets are highly unstable and a large Variety of different feature sets need to be utilized. Any time feature selection is used it can help address Big Volume challenges simply by collecting less data from the removal of certain features. Similarly, feature selection can be especially helpful in reducing Big Velocity challenges by increasing processing speeds. For example, Elngar et al. [41] found they could reduce processing times by a factor of ten simply by reducing 41 features to 11 features. Feature selection shows good promise for addressing Big Data challenges found within Intrusion Detection.

Using Hadoop to ddress big data challenges for intrusion detection

Traditional computing storage platforms like relational databases do not scale effectively against the onslaught of Big Data challenges posed by Intrusion Detection. Hadoop, an open-source distributed storage platform that can run on commodity hardware, has been utilized to better accommodate the Big Data storage requirements of massive Volume and fast Velocity along with potentially very diverse heterogeneous data structures. Collectively, Hadoop can refer to several technologies such as HDFS, Hive, MapReduce, Pig, etc. HDFS is the Hadoop Distributed File System, Hive is a data warehouse implementation for Hadoop, MapReduce is a programming model in Hadoop, and Pig is a querying language for Hadoop which has similarities to the SQL language for relational databases. Refer to [42] for further details on Hadoop.

Suthaharan [43] proposes the use of Big Data technologies like Hadoop, Hive, and the Cloud. He argues that before Big Data technologies should be employed to address Intrusion Detection, it should first be apparent that there are Big Data challenges present so as to not unnecessarily deploy Big Data technologies. Suthaharan argues that the current 3Vs of Volume, Variety, and Velocity cannot adequately provide for the early detection of Big Data, and so he proposes 3Cs of Cardinality, Continuity, and Complexity to more easily develop metrics with mathematical and statistical tools. A brief summary of the definitions for the proposed 3Cs follows:

Cardinality - number of records at an instant

Continuity - (1) continuous functions represent data; (2) continuous growth with respect to time

Complexity - (1) data type variety is large; (2) high dimensionality; (3) high speed data processing

Suthaharan proposes a Big Data model to deal with Intrusion Detection as shown in Figure 2. The User Interaction and Learning System (UILS) performs the learning on the data, permits users to interact with the system, and can control the storage requirements. The Network Traffic Recording System (NTRS) simply captures the network traffic and either stores it locally in the Hadoop Distributed File System (HDFS) or the Cloud Computing Storage System (CCSS). If data is needed immediately it is stored locally in the HDFS, otherwise it can be stored in the CCSS and can be processed later. Also, for Machine Learning in Intrusion Detection and Big Data, Suthaharan recommends the following should receive more attention: multi-domain representation-learning, cross-domain representation-learning, and machine lifelong learning.

Whitworth and Suthaharan [44] address the security challenges introduced with a model that can utilize the public Internet and the Cloud. Even though storage in

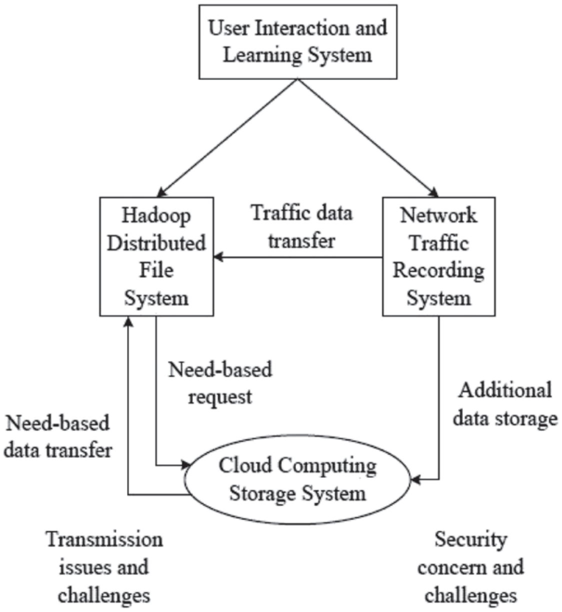
User Interaction and Learning System
Hadoop Distributed File System
Traffic data transfer
Network Traffic Recording System
Need-based request
Additional data storage
Need-based data transfer
Cloud Computing Storage System
Transmission issues and challenges
Security concern and challenges

Figure 2 Suggested network topology for Big Data analytics [43].

the Cloud can incur a significant communication cost, higher latency, and additional security challenges, the authors contend that the Cloud can extend storage beyond a local network's capacity in an elastic and "cost effective and efficient manner" using Infrastructure as a Service (IaaS). Trust levels are proposed to assess varying levels of encryption requirements based on weighted values of cloud provider "risk level" and the sensitivity of the data. A Data Key Store (DKS) is also proposed to manage security and efficiently provide for data retrievability (ensuring the data is unchanged and available).

Jeong et al. [45] give an overview of issues encountered with Intrusion Detection and Big Data and how various Hadoop technologies can address these challenges, specifically focusing on anomaly-based (misuse) IDSs. They describe various techniques and issues found with Intrusion Detection, as well as what some of the main issues are in applying Hadoop technologies for Intrusion Detection. Their study provides a good introduction for readers not already familiar with Hadoop technologies and how they can be applied to Big Data challenges found with Intrusion Detection.

Lee and Lee [46] conducted an experiment with Hadoop technologies (e.g., HDFS, MapReduce, and Hive) to measure and analyze Internet traffic for a DDOS Detector. In their experiment they were able to achieve throughput speeds of up to 14 Gbps in some scenarios, and some of their slower results were close to 6 Gbps for some analysis types while using 30 or more nodes in a cluster. Several options were tested in the experiment, such as varying the number of cluster nodes (specifically, there were either 30 more powerful nodes or 300 less powerful nodes), and they also varied the file size of the playback file from 1 TB to 5 TB while performing 5 different types of analysis. Their study only considered previously recorded traffic data from files and not real-time traffic monitoring. However, they indicated that they plan to support real-time traffic monitoring with future work. Hadoop and its related technologies show good feasibility as an Intrusion Detection tool as they were able to achieve up to 14 Gbps for a DDOS detector, and this is only a preliminary experiment with future improvements planned.

Cheon and Choe [47] propose a distributed IDS architecture based on Snort and Hadoop technologies. They performed an experiment to see if additional Hadoop-based nodes for analysis could increase processing efficiency. Their methodology was to use replay files rather than real-time data, and then to evaluate the efficiency in terms of total processing time of the replay files while varying the number of Hadoop-based analysis nodes from zero to eight. A total of nine computers were used in the experiment with one acting as the "master node". They discovered that the performance efficiency increased (it took less time to process the dataset) as they increased the number of Hadoop-based nodes. However, processing efficiency actually decreased with only one Hadoop-based analysis node. When all eight nodes were used, they saw an increase of 424% in performance as compared to using the stand-alone machine without any distributed nodes of Hadoop analysis slaves. It would be interesting to repeat the experiment with significantly more Hadoop-based nodes in order to see how far this methodology can scale out and if a certain threshold would offer diminishing returns.

Veetil and Gao [48] conducted an experiment and created Hadoop clusters to implement the Naïve Bayes algorithm in a distributed fashion. With a 6 node "homogeneous" Hadoop-based cluster where the nodes had similar hardware, they were able perform classification 37% quicker than a stand-alone machine could. While the experiment was successful as a proof of concept to use a distributed Hadoop-based cluster to implement Naïve Bayes classifier, it could only classify an average of 434 packets per minute. Much more research and experimentation can be done to implement Hadoop technologies to improve Intrusion Detection efficiency and classification accuracy.

Security data across heterogeneous sources

The purpose of this section is to describe various techniques and architectures to accommodate diverse heterogeneous sources for Intrusion Detection. In order to not deviate from that focus since it is a central theme for this study, the Big Data implications of these systems will only be partially addressed within this section. Overall, Intrusion Detection systems need to consider more diverse heterogeneous sources to provide better situational awareness within cyberspace. This can yield significant improvements to cybersecurity as Intrusion Detection is one of the core pillars of any cyber defense system. The first section gives a background on how data fusion can be used to improve situational awareness as has been done in other domains like Military applications. In the second section for illustrative purposes, a small sampling from academic studies of Intrusion Detection architectures with heterogeneous sources will be presented to give a brief overview and background of these systems. In the third section, several studies regarding SIEM systems will be presented, as well as some of the issues surrounding their deployment in the commercial sector. SIEM technology is not simply just another type of heterogeneous IDS architecture, but rather is a completely different architecture in its own right with an approach to heterogeneous data for Intrusion Detection which also provides for security analytics and forensic capabilities.

Enhancing situational awareness in cyberspace with data fusion

In 2000, Bass [36] made a major contribution to Intrusion Detection research by suggesting data fusion as a technique to aggregate Intrusion Detection data from many different heterogeneous sources such as "numerous distributed packet sniffers, system log files, SNMP traps and queries, user profile databases, system messages, and operator commands". Essentially, data fusion is a technique to make overall sense of data from different sources which commonly have different data structures. Bass also elaborated extensively on using data fusion online (near real-time) in conjunction with data mining offline in order to process the enormous amount of cybersecurity data more effectively so that it could be useful for Intrusion Detection purposes. The purpose of the data mining is to discover previously undetected intrusions based on past data, and use these to build Intrusion Detection templates. This is not performed in real-time because the data mining operations cannot always be performed quickly enough to perform near real-time reactions for Intrusion Detection (which also suggests that Big Velocity was causing problems for real-time Intrusion Detection back in 2000). These Intrusion Detection templates are applied to the online (near real-time) data fusion operations in order to better assess possible threats.

Bass [36] described that he borrowed some concepts directly from military applications such as multisensor data fusion, where on the battlefield or in military theaters a widely diverse array of heterogeneous sources can be employed. He also described using a methodology discovered by the military's concept of Observe, Orient, Decide, and Act (OODA) to gain an overall higher cyberspace situational awareness by using data fusion for Intrusion Detection, and that data fusion can provide varying levels of inference from being merely aware of an intrusion attempt up to being able to analyze the threat and vulnerability. In "Multisensor data fusion for next generation distributed intrusion detection systems" [49], Bass elaborated further on his proposed model and provides further details on data fusion. Bass's approach of analyzing Intrusion Detection data across many different types of devices and systems concurrently is an excellent example of utilizing many diverse heterogeneous sources, helping researchers gain enhanced insight into cybersecurity (particularly in the context of Big Data challenges).

Similar to Bass's approach, Lan et al. [50] utilized data fusion across diverse heterogeneous sources with the explicit goal of improving Intrusion Detection through superior situational awareness. They warned that traditional deployments of security products such as Firewalls, IDSs, and security scanners rarely work together and only possess very minimal knowledge of the network assets they are protecting. In order to bolster cyber defense through a superior situational awareness, the authors proposed using a form of data fusion known as Dempster-Shafer (D-S) evidence theory in order to make good sense out of the heterogeneous sources. D-S evidence theory is a fairly common data fusion technique utilized by researchers using data fusion within the Intrusion Detection domain, which applies probabilistic techniques to the current observations of the system. Providing details for D-S evidence theory is beyond the scope of this study, so refer to [50] for more details.

A prevailing theme encountered by Lan et al. [50] was the Big Data challenges encountered when combining events from heterogeneous sources (e.g., IDS, firewall, host log files, netflow, etc.) to achieve better situational awareness. They discuss how Big Velocity problems can make it hard "to obtain the security state of the whole network precisely when facing too much warning information". Big Volume issues exist in collecting, fusing, and analyzing "a great deal of information". With diverse heterogeneous sources, the Big Variety challenges were very clear: "the complexities and diversities of security alert data on modern networks make such analysis extremely difficult". They conducted an experiment with the DARPA2000 data set, and by using data fusion were able to simplify alert messages from a total count of 64,481 to 6,164, which is an order of magnitude improvement. While this experiment shows data fusion can be an effective technique, further experiments using more modern and robust data sets would likely be of greater interest.

It is important to caution that just arbitrarily adding a multitude of sensors and fusing them all together does not necessarily improve accuracy. This is a phenomenon described by Mitchell [51] as "catastrophic fusion" where often the performance of an entire data fusion system is worse than that of the individual sensors. Careful design and consideration must be given to properly construct a data fusion system. Further background information regarding data fusion especially with regards to Intrusion Detection can be found in [52] by Hall.

This section described the importance of enhancing cyber defense through improving situational awareness. Just like data fusion is used in other domains for improving situational awareness, it can also be applied to Intrusion Detection. Research applying data fusion to Intrusion Detection shows good potential for improving the state of the art; however, researchers should carefully consider Big Data challenges that can exist within Intrusion Detection when applying data fusion.

A sampling of various heterogeneous intrusion detection architectures

The studies presented in this section give a brief conceptual overview of the various Intrusion Detection architectures found in academic studies when dealing with heterogeneous sources. Given that the previous section illustrated the importance of considering heterogeneous sources to improve cybersecurity, the purpose of this section is to explore the architectural issues of these systems identified by researchers. Following are five different examples of architectures proposed by researchers to accommodate diverse heterogeneous event sources.

In one study, Fessi et al. [53] consider Intrusion Detection across heterogeneous sources. A good illustration for this is given in Figure 3 where multiple distributed "Observers" harvest data from various heterogeneous sources (e.g., both network and various host-based monitoring) and a "Global Analyzer" makes the ultimate decision for whether security events originating from the "Observers" are security incidents. In making its final decision, this "Global Analyzer" will perform data fusion across the various "Analyzers" to gain a better situational awareness across the multiple analyzers especially in the case of distributed attacks. One of the interesting aspects of this model is that the "Analyzers" themselves can be heterogeneous, and different types of "Analyzers" such as misuse detection or anomaly detection could simultaneously be used for the same events from observers. So essentially, each observer could be associated to one or more "Analyzers" for the motivation of detecting different classes of attacks. This model could scale well in the face of Big Data challenges given some of its distributed characteristics, which enables "Observers" and "Analyzers" to be added for scalability. However if there is only one centralized "Global Analyzer", it could become a bottleneck in the face of very

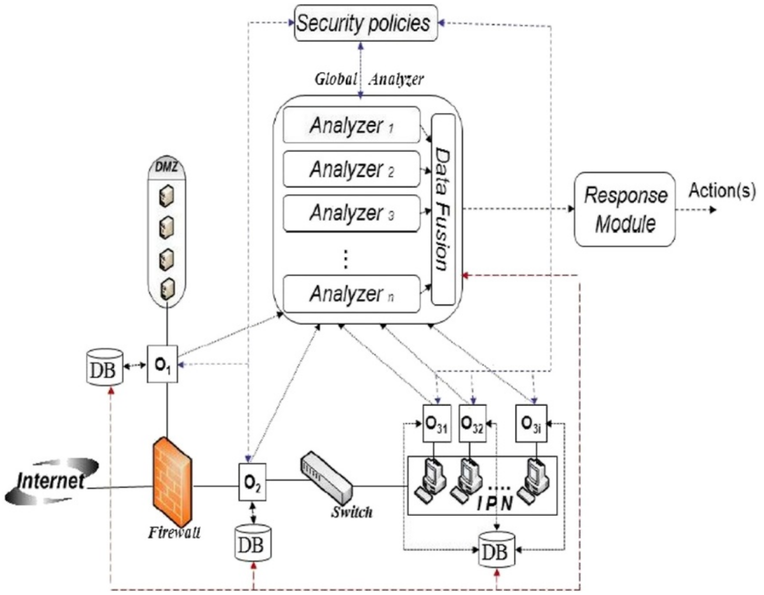

Figure 3 System Architecture [53].

Big Data or it could also be problematic if it was successfully attacked or had reliability faults.

To take more of a global view of Intrusion Detection, Ganame et al. [54] extend upon their earlier work with a centralized Security Operation Center (SOC) called a SOCBox, and develop an enhanced version called Distributed Security Operation Center (DSOC). Their architecture allows an organization to scale the system across the Internet to provide even better correlation across geographical boundaries and provide enhanced defense resiliency if one site comes under attack. Obviously, this architecture could even scale beyond multiple organizations.

One of the reasons Ganame et al. [54] extended the original centralized SOC architecture to DSOC was that the SOC architecture could be compromised by an attacker flooding the network on one site, and the Centralized SOC wouldn't be able to receive all the security events, allowing attackers to evade detection. They presented several examples of being able to compromise the SOC with "flood" attacks, demonstrating that the original SOC architecture was suspect to "flood" attacks and faced a Big Volume problem. The DSOC was able to overcome this problem by using a Local Analyzer (LA) at each site to assess intrusion detection by collecting, analyzing, and correlating alert security events locally. Each LA would then transmit a smaller and more intelligible payload of alerts to a Global Analyzer (GA) which would then perform further aggregation and analysis of alerts sent from all LAs in order to build better global awareness for Intrusion Detection (and the GA can be mirrored for redundancy and fault tolerance).

Ganame et al. [54] also describe the benefit of using diverse heterogeneous sources to correlate events across multiple sources in order to successfully detect attacks, and give an example where most homogeneous NIDS systems would be unable to detect certain multi-step attacks. Importantly, they were able to experimentally show that utilizing heterogeneous sources yielded superior Intrusion Detection capabilities over what most homogeneous approaches such as NIDS are capable of with more advanced attacks. The

DSOC system utilizes diverse heterogeneous sources and accordingly monitors all network components such as "IDS, IPS, firewall, router, work-station, etc." to yield a more comprehensive situational awareness. Refer to Figure 4 for an illustration of examples that a Local Analyzer could use as diverse heterogeneous sources. The system also employs Protocol Agents and Application Agents to better facilitate harvesting the information from the source events in an understandable format as well as in a redundant fashion and with encrypted transmission. One other interesting aspect they discussed was the need for common message formats among different devices and protocols like the Intrusion Detection Message Exchange Format (IDMEF). However, they found that the XML bus used for IDMEF was "too heavy and resource consuming," especially for event correlation. The authors implemented a separate translation process to overcome this Big Velocity challenge.

This study demonstrates a couple of Big Volume challenges in that their original SOC architecture was prone to "flood" attacks, and that they could not directly use standard IDMEF formatting due to poor event correlation performance. Also, the use of heterogeneous data sources gave superior detection accuracy over homogeneous sources in some cases.

A Collaborative Intrusion Detection System (CIDS) is presented by Bye et al. in [55], where multiple "participants" (e.g., IDSs) form teams to work together to better assess Intrusion Detection collectively. As IDS technology has proliferated, the deployment of multiple IDSs within one environment has become more prevalent. A CIDS is a way for the multiple (and even different) IDSs to work together in teams. This allows a "Bigger Picture" to be realized through collaboration. The authors present a framework which

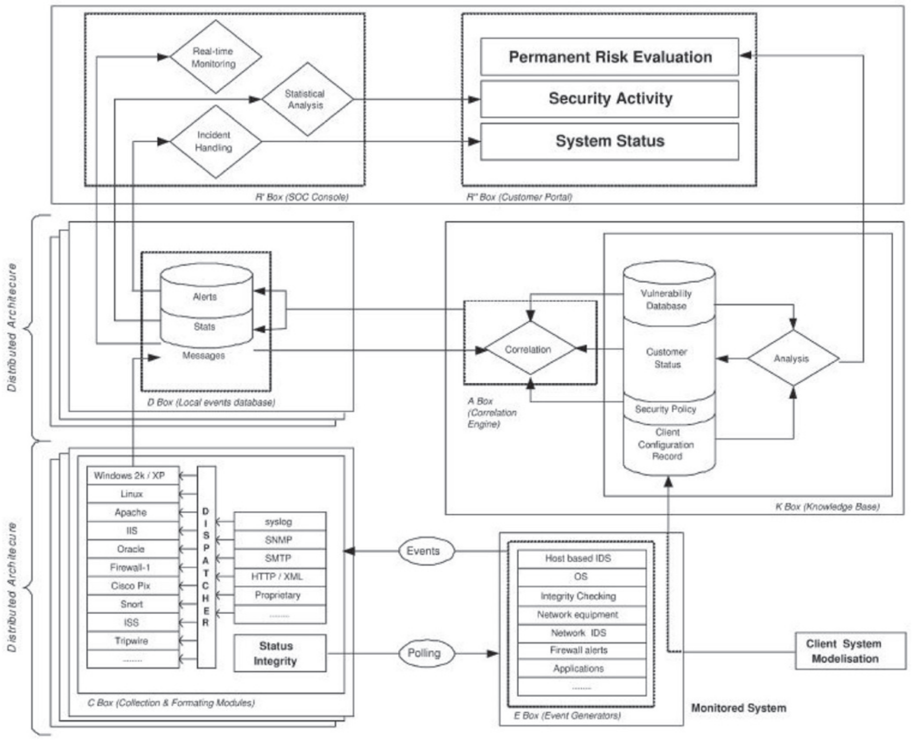

Figure 4 The internal architecture of a local analyzer [54].

can work across many different heterogeneous sources called Collaborative Intrusion Detection Framework (CIDF), and a set of mechanisms is used for a given detection or correlation algorithm to enable the Collaboration among IDSs. An "agent" is a participant in the CIDF which is also a member of a "detection group," possibly including other agents. These groups/subsets of agents have the same objective (such as anomaly detection), while another group/subset of agents may have a different objective (such as misuse detection). This study is relevant in that the authors are formally defining a framework for how CIDSs operate in general as well as how to cope with more complex issues such as security (and other issues) while collaborating. The authors also give examples of heterogeneous sources being used such as DSHIELD. Another interesting aspect is the overall heterogeneity of the framework, beyond just heterogeneous event sources. Agents within a group can themselves be comprised of heterogeneous "agents" (for example, by having different IDSs), and even the "detection groups" can be tasked with heterogeneous Intrusion Detection roles.

A Distributed Intrusion Detection System (DIDS) model is proposed by Bartos and Rehak [56] to overcome one major shortcoming of traditional IDSs: operating in isolation. Their main motivation is to increase overall accuracy and detect more threats. Importantly, their proposed DIDS can also accommodate heterogeneous sources, and their study gives examples of different event sources. The distributed IDS nodes are referred to as "sensors", and they have the capability to conduct data fusion to correlate different event types (i.e., if they are in different formats). Overall, every "sensor" IDS can communicate with every other sensor in the network with the motivation of redundancy as well as extra resiliency against attack. Each IDS "sensor" can tune itself to specialize its detection capabilities in order to improve accuracy for that specific attack class, and rely on other IDS "sensors" to evaluate other attack classes. Also, the IDS "sensors" can send requests for assistance when suspect behavior is encountered. Bartos and Rehak conducted an experiment for the proposed architecture and found that they could improve detection accuracy while keeping the false alarms constant. Their study is interesting in that data fusion across heterogeneous sources can help detection accuracy, but it is also interesting that it could not reduce false alarms especially considering that their architecture has a more global view. It would be interesting to evaluate the performance of this approach on a larger scale; however, it is a fairly novel exploration into utilizing distributed IDSs along with heterogeneous sources.

In evaluating DIDSs, Cai and Wu [57] discuss the software agent based approach for host-based systems where the agent monitors all relevant information of the host "including file system, logs and the kernel". While they also discuss the NIDS components, this is yet another example where more diverse heterogeneous sources are being monitored, enabling analysts "to get a broader view of what is occurring on their network as a whole". Cai and Wu also discuss the benefits of correlating IDS alerts across the Internet, similar to what Ganame et al. refer to in [54], and Bartos and Rehak also share this global view for Intrusion Detection in [56]. Other studies such as [58,59] show alert correlation across geographical boundaries to be an important cyber defense strategy for the enterprise as a whole. In these studies, a prevailing theme is that more diverse heterogeneous sources will enhance Intrusion Detection capabilities through event correlation and a better comprehension of situational awareness of cyber threats.

Security information and event management systems

SIEM systems are architecturally different than typical IDS solutions, and are the result of computer security vendors in the commercial sector seeking to profit by solving problems that enterprises were experiencing. The SIEM term was first coined in 2005 by Gartner Analysts Mark Nicolett and Amrit Williams [60] to describe how the industry was converging Security Information Management (SIM) and Security Event Management (SEM) technologies. SEM primarily dealt with real-time analysis for the purposes of incident response, and SIM mostly dealt with the long-term storage for the purposes of historical and trend analysis as well as providing forensic capabilities. Anuar et al. [61] discuss additional background information on SIEM technology, specifically in terms of comparing SIEM products to more traditional IDS and IPS products. SIEM systems take a more comprehensive approach beyond traditional IDSs with the motivation of giving a better holistic view of an organization's IT security, and a good definition is given by Rouse [62] in that a SIEM gives the ability to see trends and patterns of security data from a single point of view even though the security data can originate from diverse heterogeneous sources such as the network, end-user devices, servers, firewalls, antivirus systems, and intrusion prevention systems.

According to Gartner [63] SIEM software sales was $976.4 million in 2012 with 27.5 percent growth, and for comparison the overall security software market grew from $17.7 billion in 2011 to $19.1 billion in 2012 as tracked by Gartner (with SIEM software comprising about 5% of the total market share in security software). Per Mosaic Security Research [64], there are currently 65 SIEM products as of this writing with 6 of them being classified as freeware. As SIEM technology is relatively new as compared to IDS technology, there are still many academic research opportunities especially considering the widespread commercial growth of SIEM technology. Following is a brief overview of SIEM technology. SIEM products can differ from each other in how they operate and in terms of features they provide, and one particular SIEM definition might not universally apply to all SIEM products. Aguirre and Alonso [65] generalize the major SIEM functionalities as the following: "aggregates data from many sources, continuously monitors incidents, correlates events, and issues alert notifications". In their study, they also contend it is important for organizations to aggregate SIEM information across their multiple domains and they propose a federation of SIEMs to accomplish that goal.

Similarly, after analyzing SIEM systems, Kotenko et al. [66] contend the four main SIEM components are the following: "event filtering, aggregation, abstraction, and correlation; reasoning and visualization; decision support reaction and counter measures; attack modeling and security evaluation".

Either definition is sufficient for SIEMs, especially since Kotenko et al. drew their definition from the most advanced SIEMs as defined by Gartner Analysts Nicolett and Kavanagh in [67], while Aguirre and Alonso apply their generalization to a broader range of SIEM products.

Kotenko et al. [66] also discuss the various standards used by SIEMs to represent security events and incidents in standardized formats: SCAP [68], Common Base Event (CBE) [69], and Common Information Model (CIM) [70]. One of the main design motivations of SIEM technology is that vendors will typically try to ensure all the security data sensor sources have as common a format as possible in order to minimize the amount of "data fusion" analysis that needs to be performed. They also discuss the inadequacy found in all the major and widespread SIEM systems in that they use a traditional relational database for the storage and querying of the security data, and that the relational database model that the SIEMs are using is often overloaded. To overcome this limitation, they recommend using a hybrid database approach for the SIEM repository where a traditional RDBMS is used in conjunction with both XML-based databases and a triplet store ("A triple store is a purpose-built database for the storage and retrieval of RDF metadata" [71] and the triple is based on the "subject-predicate-object" methodology). It would be interesting to determine whether Hadoop technologies could provide any storage repository benefits as that was not mentioned by Kotenko et al.

The fact that traditional RDBMSs are a performance bottleneck for SIEM systems demonstrates that they face major Big Data challenges. This makes sense as they are analyzing security data across a myriad of diverse heterogeneous sources, whereas this survey points out numerous times that Big Data challenges can be encountered at only single sources for Intrusion Detection. Kotenko and Chechulin [72] propose an interesting attack modeling framework for SIEM systems in [72] called Attack Modeling and Security Evaluation Component (AMSEC) to address both known and unknown (zero day) vulnerabilities.

Metzger et al. [59] conducted a study in Higher Education Institutes (HEIs) regarding Intrusion Detection and how it can be applied with SIEM technology in conjunction with formalized Incident Management techniques. The overall system can react to security events either automatically or manually through a Computer Security Incident Response Team (CSIRT). HEIs can be highly targeted among botnets, email spammers, and others for their high bandwidth capabilities among other reasons. The authors propose a framework where in addition to the traditional SIEM approach, a couple of other non-traditional sources for the SIEM system are considered: Manual Reporting and the "DFN-CERT service". Manual Reporting allows outside organizations or individuals, internal Administrators or Support staff, and Help Desk tickets to report security incidents and information directly to the system for automated processing. This extra source can benefit Intrusion Detection for the SIEM system with increased detection accuracy as it broadens the scope of events being monitored. The SIEM can also correlate its other events with this new source for increased benefit. The "DFN-CERT service" is a worldwide service to automatically report malicious behavior and metadata to the local SIEM system, with the similar benefit of enhanced detection and correlation capabilities for the SIEM. These two other methods are used in conjunction with the traditional SIEM monitoring, correlation, and analysis functionalities. Additionally, their model includes having the SIEM either take automatic responses to events or to notify appropriate Administrators for action based on configurable policies and/or threshold measurements of events. With their model, they were able to automatically react (at least partially) to more than 85% of all abuse cases in their HEI study. This is important in that it shows more heterogeneous sources can enhance detection, correlation, and reaction capabilities of SIEM systems, especially with regards to reporting more diverse security events and metadata to the local SIEM system. Benefit can also be gained with the local SIEM receiving cybersecurity intelligence from a worldwide network. Also, it is important to note that heterogeneous sources need not be limited to cyberspace as shown in this study, and that reports from humans can enhance the situational awareness as well. A major motivation for the system implementation was to automate everything as much as possible in a formalized manner.

In a study evaluating systems to monitor security for cloud computing, Diego et al. [73] conclude that no single solution can currently cover all existing security threats for Infrastructure as a Service (IaaS) platforms. To enhance the overall infrastructure security it is recommended to use more diverse and heterogeneous solutions. Diego et al. give an example that two different types of SIEMs could detect more threats than just one. In addition, this study proposes a Quality of Protection (QoP) in terms of both better fault tolerance and enhanced security for the security system itself by using systematic redundancy (i.e., if one part of the system fails or comes under attack then a redundant piece can still function). An experiment was carried out with the commercial ArcSight SIEM product to test the throughput when using redundant SIEMs in a Byzantine fault tolerant architecture. A total of 4 ArcSight SIEM "replicas" were used with one being allowed to be faulty, and over 250,000 events per second could be processed. They determined the system was bound by resource exhaustion, and additional resources could further increase event throughput. This study did not elaborate on the methodology of storage and querying of the events into the archival repository with regards to forensic purposes and how the archival repository would scale out as additional SIEMs were added to the system. This study is interesting from a Big Data standpoint in that it shows good experimental results in the scalability of SIEM technology, but there is no clear indication in whether SIEM technology would face a scalability threshold with relational databases still being the prevalent storage engine. However, recently some vendors such as Splunk [74] have adopted relational database technologies in order to better address Big Data challenges.

In an effort to make SIEM technology more effective in defending against rapidly evolving cyber threats, Li et al. [75] recommend an Enterprise Security Monitoring (ESM) solution. Large enterprises are facing increasingly challenging attacks such as Advanced Persistent Threats (APTs), and one major problem these large enterprises can have is their various security teams might be fragmented into different organizational silos which can cause difficulties in sharing security intelligence information across these boundaries to better correlate events, especially against more advanced attacks. Their proposal to advance cyber defense to face these challenges is to use SIEM technology as a core component, and use this in conjunction with Enterprise Security Intelligence (ESI) to enhance overall next generation cyber defense architectures. Essentially, ESI will extend the overall security intelligence of the SIEM capabilities similar to how Business Intelligence (BI) is traditionally applied, and would allow more advanced security intelligence analytics to be developed and utilized in order to adapt to more advanced threats. For example, to provide improved situational awareness with ESI, business context information specific to the organization could be combined with alerts generated from the SIEM as well as various intelligence sources (i.e., those reported by humans or systems).

As originally put forth by Gartner in [76], Li et al. explain the six design principals of the next generation ESM shown in Table 2 (Refer to Figure 5 for a visual illustration).

The creation of a "fusion center" for the enterprise is recommended by the authors where the ESM is collaboratively utilized across organizational boundaries in a variety of ways especially regarding any aspects that could be fragmented (e.g., planning, risk assessment, data sources, intelligence analysis, etc.). They also contend that different

Table 2 Six design principals of the next generation ESM

1. "Comprehensive Enterprise Coverage"	The entire production IT stack (e.g., "networks, hosts, applications, databases, identities") for the enterprise must be monitored by the ESM regardless of environment (i.e., onsite or in the cloud).
2. "Information Interaction and Correlation"	All meaningful events, logs, and similar from input sources in #1 must be capable of being collected for correlation.
3. "Technology Interaction and Correlation"	The SIEM will serve as the foundation of the correlation engine, however it should also integrate with other important security technologies such as: Firewalls, IDSs/IPSs, DLPs, Vulnerability Management, and Anti-Malware.
4. "Business Interaction and Correlation"	The ESM must be aware and tuned to the specifics of the organization's business context to better assess an attacker's motivation and yield better correlation and intelligence.
5. "Cross-Boundary Intelligence for Better Decision Making"	The ESM solution must span organizational boundaries across the entire enterprise in a cohesive and collaborative manner, and not permit fragmentation with regards to its overall cyber defense.
6. "Visualized Output for Dynamic and Real-time Defense"	The output of the system must be easily visualized and understandable by end user analysts in an effective manner.

organizations could even benefit by collaborating and sharing security information with each other. However they emphasize the great difficulty posed by this because of competitive, technical, legal, and possibly embarrassing reasons (i.e., disclosing certain breaches could harm their image). In order to better cope with the Big Data challenges of processing and storing "massive amounts of data", Li et al. suggest that technologies like Hadoop could be leveraged. They also recommend cloud-based Enterprise Security Monitoring vendors as a "natural solution" for Big Data and scalability issues of enterprises.

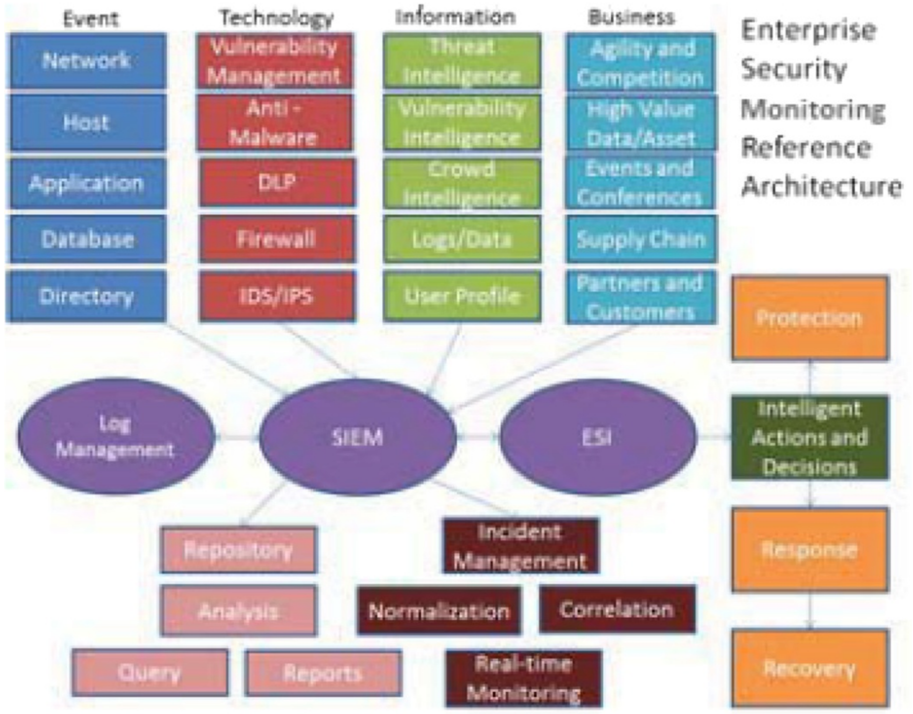

Figure 5 An enterprise security monitoring reference architecture [75].

Big heterogeneous data for intrusion detection

The purpose of this section is to describe Big Heterogeneous Data in terms of different categories to illustrate the various underlying levels of heterogeneity for Big Data within Intrusion Detection. At a high level, Big Heterogeneous Data can be described in terms of being input or output data. Big Heterogeneous Input Data can be further categorized into traditional Big Cyberspace Data and Big Industrial Data (i.e., data from industrial processes in the real physical world). Big Heterogeneous Output Data will be presented in the categories of Big Archival Security Data (which considers the long term storage aspects) and Big Alert Data (which will present Big Data issues surrounding alert data).

Big Heterogeneous input data

Big Heterogeneous Input Data is essentially just the types of input data in the spirit of Big Heterogeneous Data from the previous section (security data across heterogeneous sources). It can be considered simply as just heterogeneous input Big Data. The following sections each discuss one particular type of heterogeneous input Big Data grouped by higher-level categories. It is important to consider that a great deal of heterogeneity among the sources can be present within these categories. First, the traditional cyberspace input Big Data is presented. Then, Big Heterogeneous Industrial Data beyond cyberspace is discussed, and this section gives examples of Big Data from the physical world outside of cyberspace (e.g., industrial process data) which can further improve situational awareness even in cyberspace.

Big Heterogeneous cyberspace data

Big Heterogeneous Cyberspace Data are the traditional input types of data which are commonly considered in Intrusion Detection literature, but here they are presented in the context of Big Data. Both network layer and host layer event sources are considered. The network layer coverage is essentially just the network traffic that traditional approaches like NIDSs (e.g., Snort) monitor with a focus on Big Data. The host layer coverage focuses on Big Data challenges with different host sources, and is equivalent to the traditional HIDS approaches where computer servers, workstations, devices, etc. are being monitored. Again, it is important to consider that a great deal of diverse heterogeneity can occur among event sources in this category.

Nassar et al. [1] contend that outsourcing flow-based network monitoring and Intrusion Detection to cloud providers can be cost effective if done so in a secure manner. They give an example of Big Data with a university network that produces an average load of 650 Mbps and peaks up to 1.0 Gbps, and assert that because of such Big Network Data that “many monitoring systems have already shifted from the deep packet inspection to the aggregated flow data level”. In other words, because of such Big Velocity at the network level, more accurate techniques of intrusion detection such as deep packet inspection are being abandoned in favor of less computationally intensive techniques such as monitoring at the network flow data level. Nassar et al. discuss privacy and anonymization issues in being able to securely outsource network monitoring, and that a university network with an average load of 650 Mbps posing Big Velocity challenges is of particular interest.

In order to evaluate Big Network Data, Sitaram et al. [77] consider network-based IDS challenges faced by large cloud providers or those with fat network pipes such as OC-192 and OC-768 links. They consider such data as a “clear representation of big data streams in its most raw form (which is hundreds of thousands of TCP/IP packets per second)". Sitaram et al. envision building a NIDS capable of handling Big Data network streams such as these by utilizing Big Data tools such as Hadoop and a network monitoring tool called PacketPig [78]. According to the authors, PacketPig is capable of Deep Packet Inspection, deep network analysis, and even full packet capture when using it with Hadoop. In this study, they mainly consider the effectiveness of clustering algorithms for analyzing packet classification. Their experiment with the KDD data set found the K-means clustering algorithm to generally outperform the Expectation-Maximization and DBSCAN Clustering algorithms. However, their future work sounds especially interesting if it can successfully operate in terms of such Big Velocity.

Beyond the network, host-based event log data has traditionally been one of the main sources for Intrusion Detection monitoring. An organization can have a multitude of computing hosts both in quantity as well as diversity in terms of the different types of log files being generated. All of this log data can quickly add up to Big Host Event Log Data in that it can be very high in Volume, Velocity, and Variety.

The hosts that produce these logs can have Variety such as end-user computer workstations, computer infrastructure servers, devices, appliances, virtualization hosts, or even cloud-based hosts. The types of logs being generated are heterogeneous and can vary from Operating System events to a wide variety of application events such as antivirus software, firewall logs, honeypot activity, web server logs, ftp server logs, email server logs, domain controller logs, web proxy logs, VPN logs, DHCP server logs, etc. While this is not a comprehensive listing of the various types of logs one can encounter in the typical organization, it illustrates that the various types of logs can pose Big Variety challenges in having to correlate security events across a wide range of heterogeneous log types.

To better cope with Big Data challenges organizations can face with their log data, Yen et al. [79] developed a system called Beehive which performs "large-scale log analysis for detecting suspicious activity in enterprise networks". They report that organizations are facing Big Volume challenges in terms of the logs being "very large in volume", and implemented their system at a large enterprise, EMC, for two weeks. At EMC, they describe their major challenges as the "Big Data problem" where 1.4 billion log messages are generated on average per day (about 1 terabyte). This also suggests Big Velocity challenges in dealing with such a high data rate as well. They also discuss the problem of organizations implementing a variety of different security products which generate logs whose formats vary widely, and this suggests a Big Variety challenge. Also, they note that these logs from various security products may have problems such as incompleteness or even inconsistency, and so they describe logs with these challenges as being "dirty".

Beehive monitors the communication of dedicated hosts (e.g., workstations) with other host targets. This is accomplished by monitoring logs from a wide range of network devices such as web proxies, DHCP servers, VPN servers, windows domain controllers, and antivirus software. The logs are ultimately stored in a commercial SIEM system. They reported all the log information being stored in the SIEM as "a big data problem", and that "efficient data-reduction algorithms and techniques" are required to cope with such Big Data logging challenges. Another major challenge they encountered with the logs stored in the SIEM was that the information which was actually being stored proved difficult to correlate against other events because of the underlying quality of the data. As an example, logs could have the incorrect time stamp because of being in a different time zone. Another log format difficulty encountered was that typically only IP addresses were stored, and it was difficult to associate events to specific hosts given that IP addresses could change via DHCP. So, correlation against additional logs was necessary to properly identify hosts.

In terms of detecting security incidents, the Beehive project proved fairly successful versus the enterprise's state of the art security tools. Over a period of two weeks, 784 security incidents were discovered by Beehive whereas the enterprise's existing security tools only detected 8 of these incidents. So, a large number of security incidents were in fact unknown with the existing security tools. One source which was especially beneficial was the web proxy logs, where suspicious traffic activity or questionable destinations could be discovered. Yen et al. [79] were able to effectively reduce the amount of messages inspected by 74% simply by filtering out whitelisted target hosts. This is an interesting study in the effectiveness of the approach, especially considering the Big Data challenges encountered. It would be interesting to see if future similar work could perhaps extend beyond workstations to server log behavior.

According to Myers et al. [80], event correlation is frequently not performed with log analysis due to "difficulties and inadequacies with current technologies". One reason they indicate that organizations have difficulties analyzing security logs is because of "the sheer volume of data to collect, process and store". This suggests that log analysis with event correlation for Intrusion Detection is a Big Data challenge in the contexts of both volume and velocity. Myers et al. conducted an experiment on web server logs to evaluate the effectiveness of applying event correlation in a distributed fashion, and their results showed this technique could effectively detect many common web application attacks while maintaining a low false positive rate. Their distributed approach also showed a reduction in network traffic of syslog messages by 99.88%. This distributed approach illustrates good potential for addressing Big Velocity found in security network traffic by reducing that amount of traffic.

Big Heterogeneous industrial data

Cyber threats can damage and even destroy real-world physical targets beyond cyberspace. Industrial and Utility operations are especially prone to this exposure given their evolution of integrating and automating their physical operations with Information Technology from cyberspace. Even when these systems are "air gapped" and physically disconnected from the public Internet and other networks, these cyber threats can still be catastrophic in nature to real-world objects. An example of a successful attack occurred against Iran's nuclear program with the Stuxnet virus, and some of Iran's nuclear centrifuges were destroyed in the attack. Further details of this incident are given by Langner [81].

Therefore, it can also be important to include heterogeneous sources from the physical world to better improve overall situational awareness for security. A good conceptual illustration for how to extend monitoring beyond cyberspace is given in Figure 6, and this shows different Host, Network, and Device IDSs harvesting information into a centralized SIEM system with the goal of improving Intrusion Detection by also analyzing data from Process Control System sensors. This illustration is from a study performed by Valdes and Cheung [82] with the explicit goal of gaining better situational awareness in process control systems. The motivation was to extend the existing functionality of

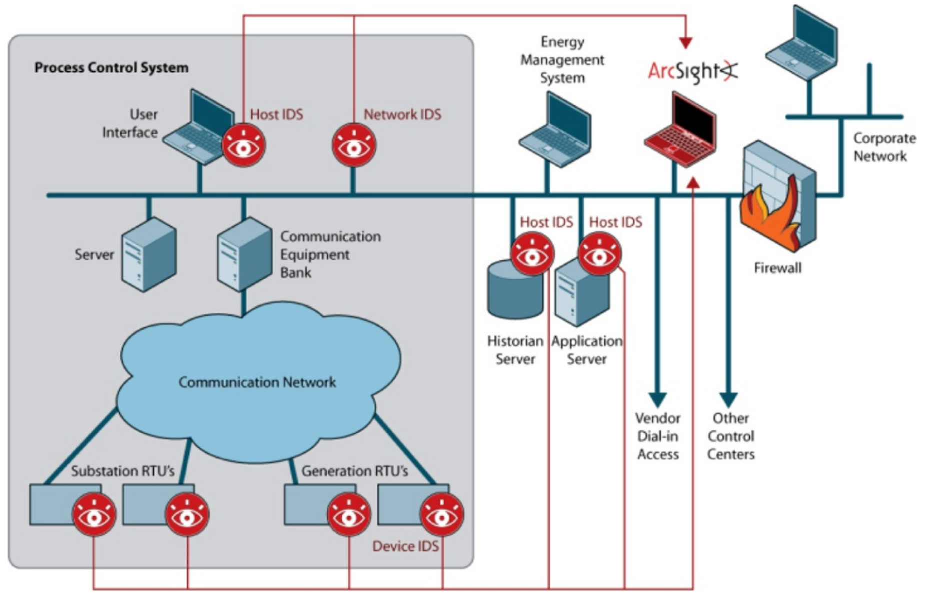

Figure 6 Control Center Level View of the Detection and Event Correlation Frame-work [82].

a SIEM product (ArcSight) to correlate control/process system alarms with IDS events from the SIEM, and thus extend situational awareness beyond cyberspace to also include industrial physical process control systems by correlating IDS data with measurements from underlying physical process data such as electrical current, pressure, flow rate, and similar industrial measurements. This is an interesting concept in that cyberspace situational awareness can be improved by correlating data from heterogeneous sources in the physical world beyond cyberspace, and that Intrusion Detection need not be merely limited to cyberspace sources. The authors indicate that important industries such as refining, pipelines, and electric power can benefit from this approach of utilizing more diverse heterogeneous sources, while cautioning that the stakes are especially high for detecting cyber-attacks against those platforms, as damage can also be physically harmful or even deadly, such as releasing hazardous materials into the environment. Refer to [83] for further background information and also the final report [84].

According to Xiao-bo et al. in [85], Data from Industrial and Utility operations certainly can have Big Data. They design "a big data acquisition engine based on rule engine" to better handle Big Data acquisition flow problems when industrial processes face such challenges. Their study describes how supervisory control and data acquisition (SCADA) systems are evolving and increasingly using Ethernet-based networks rather than traditional serial port connections. SCADA systems are essentially Industrial Control Systems which are based on computers, and they control and monitor industrial processes in the physical world. Xiao-bo et al. propose a rule engine implemented in Java Expert System Shell (JESS), and seek to improve performance and quality from industrial Big Data acquisition flows. Their design accommodates the real-time demands of SCADA systems, and can be used for data analysis, alarming, and forecasting. It is interesting that they chose to address Big Data challenges as a cornerstone of their design in dealing with data flows from SCADA systems.

A study conducted by Datta Ray et al. [86] evaluates Security and Big Data challenges that the electric power grid is facing. They assert that a paradigm shift is required to properly address the cybersecurity demands of the smart grid (i.e., electric power grid). Even more diverse heterogeneous event sources will need to be monitored, and this will further exacerbate the "Big Data onslaught" the utilities are currently facing. Additionally, a Formal Risk Management system should make "the analysis of such Big Data manageable, scalable, and effective." A more formal Return On Investment (ROI) analysis should rigorously address an enterprise's multitude of security and risk contexts.

When it comes to cybersecurity, Datta Ray et al. contend that "the crux of the problem is that organizations have taken a piecemeal approach to security". Various security products such as firewalls and antivirus software do not communicate systematically with each other to yield "holistic intelligence", and typically by the time meaningful patterns are found it is too late and the damage has already occurred. They further assert "these point or perimeter solutions applied to host computers, networks, or applications often work with little knowledge of each other's functions and capabilities". In order to be successful in achieving a holistic risk management system, a major design consideration is interoperability between a diverse myriad of devices such as "meters, synchrophasors, IEDs, firewalls, field devices, etc." This interoperability is important, and should even consider both structured and unstructured data from the sources. By interoperating among "existing point, perimeter, and defense-in-depth security solutions with actionable insights", a more systematic and superior cyber defense can be realized.

In this spirit of considering more diverse heterogeneous security event sources, Datta Ray et al. provide great detail on enhancing overall security by integrating security event sources beyond cyberspace and the Big Data challenges of doing so. The model that they use to illustrate this is by categorizing traditional cyberspace event sources as Utility Business Information Technology (IT), and by categorizing event sources beyond cyberspace as Power System Operation Technologies (OT). For example, IT event sources could be typical cyberspace components such as firewalls, and OT event sources could be a meter that measures electrical quantities such as power, voltage, and current.

By combining and correlating security events across both IT and OT sources, an improved situational awareness can be realized. However, both the IT or OT sources can face Big Data challenges, and Datta Ray et al. propose a model shown in Figure 7 on how to cope with the Big Data onslaught in a systemic manner to improve risk management. In addition to diverse heterogeneous sources, the model also considers "exogenous and endogenous sources of intelligence and asynchronous and real time interactions among its various components". The model indicates how the system can provide feedback in near real time to react to specific events, and that the intelligence of the system is based on an aggregate of Big Data IT and OT inputs.

By unifying the IT and OT domains and correlating events across the entire spectrum, the overall quality of the entire smart grid can be enhanced even while processing a Big Velocity of information from a Big Variety of diverse sources which can lead to valuable insights from previously unknown correlations and hidden patterns. The smart grid faces all of the "3Vs" of Big Data criteria: "huge volume, diverse sources and types and the varied speed of the incoming data". The amount of data the smart grid faces is so large that it has "surpassed the ability of traditional relational and scan/sort systems to process the data".

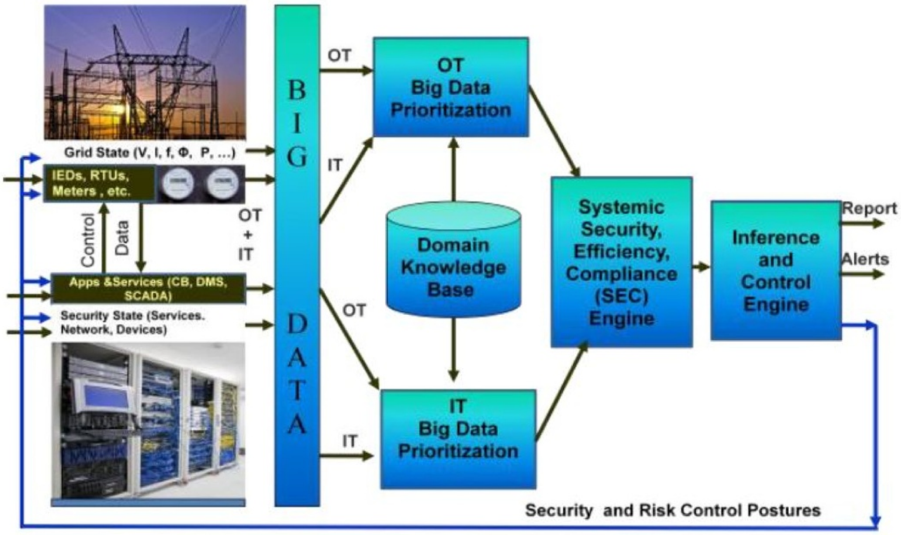
Grid State (V, I, f, Φ, P, ...)
IEDs, RTUs, Meters , etc.
Control
Data
Apps &Services (CB, DMS, SCADA)
Security State (Services. Network, Devices)
OT + IT
BIG DATA
OT
IT
OT
IT
OT Big Data Prioritization
Domain Knowledge Base
IT Big Data Prioritization
Systemic Security, Efficiency, Compliance (SEC) Engine
Inference and Control Engine
Report
Alerts
Security and Risk Control Postures

Figure 7 Big Data and OT, IT Systemic Risk Management [86].

Formal Risk Management and ROI analyses are very important considerations in how utility companies process Big Data, as the risks are high. While financial risks such as electrical theft or outages are obvious, the loss of human life is possible in extreme cases like a gas explosion from a leak, or if a hospital (or home) that uses life-support systems were to be maliciously denied electricity. Instead of merely “drowning in data”, Big Data can be used as an advantage to improve risk analysis and combined with ROI analysis can more effectively prioritize information assets in terms of effectiveness, scalability, and protection.

Big heterogeneous output data

Big Heterogeneous Data can be output data as well, and this is classified as Big Heterogeneous Output Data. This section addresses the heterogeneity of output Big Data for Intrusion Detection in two main categories: Big Archival Data and Big Alert Data. Big Archival Security Data is output data which is being archived either for the purpose of forensics or Security Analytics, while Big Alert Data is output data either for further alerting analysis or for notifying an administrator or system component to take action. Both of these Big Heterogeneous Outputs can have very pronounced Big Data attributes in terms of Volume, Velocity, and Variety.

Big archival security data

A very important aspect for Intrusion Detection is long-term storage of certain security data. Essentially, there are two main goals for the archival of security data. The first goal is to improve Intrusion Detection capabilities even in real-time with offline data mining operations and Security Analytics. This offline data mining operation on security data can further try to identify previously unknown cyber threats, and then update the real-time detection capabilities with additional new signatures or behavior traits. The second goal is to provide forensic capabilities with this data so that in the event of a security breach, forensic evidence is available to assist the investigation. This data can also be used as evidence in legal proceedings if properly maintained. Typically not every single piece of computing data will be kept in the offline repository store, and care must be taken to properly filter out what is not necessary.

In an experiment using log data generated from anti-malware software, Hoppe et al. [87] use data mining to search for patterns among malware infections in the archived storage repository of a SIEM. They described difficulties in performing the data mining in dealing with such a large amount of data, and that a critical success factor was utilizing the formal CRISP-DM data mining process. In addition, they also described that in IS infrastructures: "the amount of data is enormous". Big Data challenges were a factor in their study especially in the context of Big Volume, given that their data mining operations were performed offline and not for real-time Intrusion Detection. In their study they found that the age and gender of workstation users could infer high or low risk of malware infection. However it was also found that users with or without administrator rights on their workstations did not influence malware infection. Hoppe et al. contend that in specific scenarios, performing data mining on data collected by SIEM systems can enhance the quality of Information Security Infrastructure for companies. This is an illustration that feedback from the offline archive store of a SIEM can be useful for better real-time inference of events which might even possibly yield better situational awareness.

A model proposed by Hunt et al. [88] seeks to both enhance real-time adaptive security while improving long-term forensic capabilities. They point out that security data "arriving too fast to store" and process can now be better addressed with new terabyte storage devices, parallel processing, clustered computers, and even super computers. To give a point of reference, a network with a 10Gbps flow over one hour of incoming traffic requires 5 terabytes of storage. They assert that such infrastructure is currently out of reach for medium sized organizations. Given the processing and storage problems and the application of super computers, there are clear Big Data challenges, especially in the context of Volume and Velocity.

According to the Hunt et al., most IDS/IPS and firewall systems even when reporting information to SIEM systems frequently do not capture sufficient information for robust forensic capabilities as they do not create "digital evidence bags". These systems usually do not sufficiently automatically react in real-time or provide sufficient "traceback" functionality. "Traceback" functionality is the ability to correlate already identified malicious sources (e.g., source IP, port, ISP, etc.) with other real-time components of the network as well as for future forensic purposes. In their model, they also suggest that honeypots, honeynets, and sinkholes are important components for the overall system. Sinkholes can be summarized as a way to draw these packets into a "sinkhole" and allow the malicious packets in so that they can be recorded for forensics as well as later correlation of suspicious behavior, versus having a firewall merely dropping packets with malicious characteristics. These malicious packets are ultimately drowned and not permitted to propagate past the sinkhole, as another firewall explicitly blocks these packets.

Hunt et al. also assert that Data Loss Prevention (DLP) systems are mature in some regards. However their capability to link back to the original events pertaining to breaches with forensics is "largely lacking". They emphasize that DLPs should integrate better with SIEM systems to improve the "very serious" situation of needing better forensic capabilities. Importantly they realize it is "inevitable" that for both real-time security and forensics the focus needs to shift from network protection to data centric protection (i.e., data leakage from the database). Intuitively, this is a compelling argument given that gaining access to such data is a major motivation for attackers.

While their model gives many interesting details, a few examples will be briefly mentioned. One of the recommendations is to use encryption for the log devices from all sources, and to use a "digital evidence bag" for the purposes of forensically sound data (including the possibility of using a kernel security module to mitigate interception attacks). Also emphasized is that a proper chain of custody must be employed in order to be able to properly prosecute perpetrators, and a couple of helpful items for this would be cryptographic hashes and key management for all evidence.

Their proposed model seeks to improve upon existing forensic capabilities while also enhancing real-time Intrusion Detection with additional correlation sources. The authors emphasize that not all systems have these weaknesses, although many do. They also indicate that the extent to which correlated data can adapt firewall rules in the real-time is an open research question.

Big alert data

Intrusion Detection Systems and other security systems produce alerts to notify administrators of suspicious activity. Even an individual IDS can trigger many alerts, and the problem becomes even more prominent when dealing with heterogeneous sources such as a wide array of sensors or multiple IDSs. The basic problem is that a single security inspection event can trigger many alarms even if it is a single incident, or many false alarms can even be raised with normal traffic.

A common technique which is used to stop a flood of alerts is called alert correlation. The basic concept of alert correlation is that when the same characteristic is causing the same alarm, the system should filter and aggregate multiple alarms into one alarm so that a flood of alarms of the same type does not occur (instead just a count of those same alarm types could be reported). An illustrative example of alert correlation is given in Figure 8

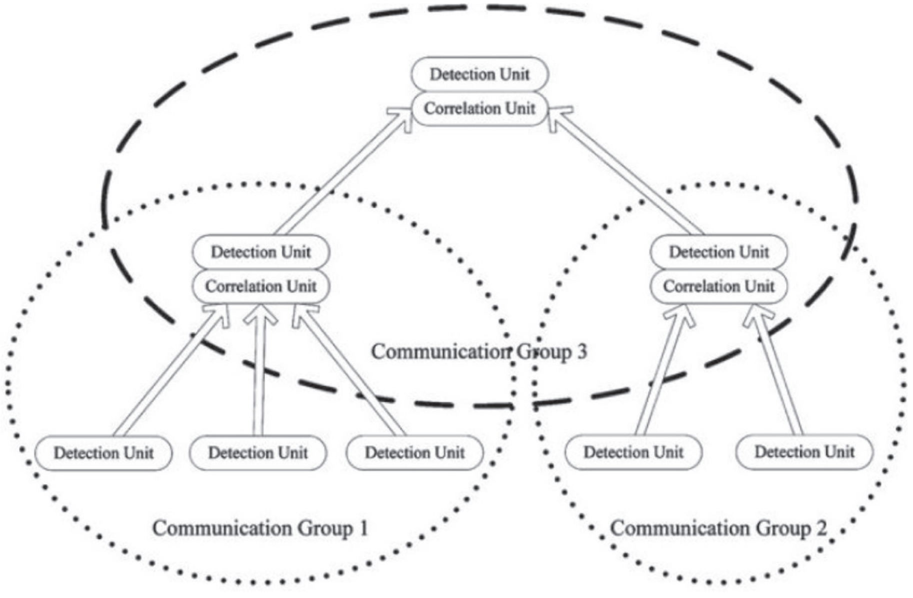

Figure 8 Hierarchical CIDS architecture [9].

where alerts are initially correlated locally in a hierarchical fashion. They are subsequently correlated again at a more global level.

The process of generating alerts certainly can involve Big Data challenges in terms of Volume, Velocity, and Variety. Big Volume and Big Velocity challenges for alert generation can involve correlation with other alerts, events, rules, or knowledge bases. These correlation activities can involve massive processing power, storage requirements, and network traffic. Big Variety challenges for alert generation can involve correlation among alert generators such as IDSs that can have many different formats for their alert messages or event data. It is common for organizations to have security products with many different proprietary alert formats, even though efforts are still being made to standardize. Semantically, alerts can either be considered inputs or outputs as they can also serve as inputs for alert correlation purposes. Alerts always operate at least once in an output capacity, but alerts do not always operate in an input capacity. Since alerts are typically considered outputs conceptually for notification purposes as well as for archiving and forensic purposes, they will be categorized as outputs for this study's organizational purposes.

The actual alerts themselves can indeed pose Big Data challenges, and a study by Sundaramurthy et al. [89] gives an example where they analyzed a data set with over 35 billion alerts from the HP TippingPoint IPS product. This data was collected over a 5 year period in over 1,000 customer networks worldwide. The data mining analytics posed the most significant Big Data challenges given the sheer volume and processing requirements which included correlating the alert data with the filter metadata (which contains further details about the actual alert). One interesting insight learned from the data was that with the Denial of Service (DOS) or Distributed Denial of Service (DDOS) attacks, the majority of these attacks were actually attacking the application layer instead of the network layer. Traditionally, DOS attacks are at the network layer where a bunch of packets "chokes" a device. However, newer attacks can more effectively attack the application layer by sending a specially crafted malicious packet which will cause the application to consume significant resources. They found 78.65% of the DOS alerts were from application-layer attacks, and only 21.35% of the DOS alerts were from network-layer attacks. DOS attacks can present Big Data challenges in themselves. If alerts from these DOS attacks are not properly handled via alert correlation or similar techniques, then the alerts themselves can actually compound the Big Data problems from the DOS attacks with alert floods. These DOS attacks pose a Big Velocity problem, and their corresponding alerts could as well if not properly handled.

For Collaborative Intrusion Detection Systems (CIDSs), Zhou et al. [58] found that a decentralized approach for alert correlation versus a centralized approach could significantly reduce the processing time and number of alarms while not significantly sacrificing detection accuracy. This is especially important given that one of their major challenges was a Big Data challenge with the alert messages themselves being bottlenecked at a centralized alert correlation server: "the scalability problem addresses the challenge of how to cope with the huge volume of raw alerts that can be generated by each participating IDS in the system". They were able to solve this bottleneck problem by utilizing a hierarchical alert correlation architecture as shown in Figure 8. For a given stealthy scan scenario, they were able to significantly increase detection accuracy with their probabilistic threshold approach versus a naïve scheme that used the same threshold. This architecture was able to accommodate multi-dimensional data sources by restricting their analysis to only four specific features (Source IP Address, Source Port, Destination Port, and Protocol), and Zhou et al. applied these four features to evaluate eight different combinations containing them. Admittedly their approach had drawbacks such as not being able to always properly detect if the attacker spread their attack over a longer period of time than the threshold could handle. Future work for their architecture should also address security countermeasures if distributed node(s) are compromised, which is quite possible considering that this architecture is a distributed system that could span across the public Internet. Their study is interesting from a Big Data perspective in that they could significantly improve overall detection time by actually benefiting from the increased message routing and processing time in a distributed architecture. However, their study obviously only considered network traffic features and was not a comprehensive evaluation across heterogeneous sources beyond network traffic features. More work in this area should evaluate just that.

The processing for alert correlation is clearly a Big Data challenge. Roschke et al. conducted an experiment [90] to evaluate the efficiency of alert correlation with three different database architectures: traditional row-based DBMS (e.g., MySQL), memory-based column database (e.g., MonetDB), and a custom built in-memory database. In all cases, the in-memory database yielded far superior processing for alert correlation with the authors concluding that the row-based database showed “poor performance” for clustering and correlation. While the column database performed better than the row-based database in terms of alert clustering and correlation analysis, the column database performed extremely poor in terms of record insertion, being able to only insert 63 records per second, and so the authors concluded that particular database was “unfeasible” for general IDS utility and only useful for analytic purposes. The authors do point out that the utility of the in-memory database is obviously limited by the amount of memory available which will give an upper limit to the amount of alerts that can be correlated (or clustered), whereas traditional databases do not have this limitation. They propose limiting the time window in which alerts can be correlated or clustered to avoid this problem. This experiment shows that alert correlation is a Big Velocity challenge, although this experiment was limited to only evaluating network-based alerts via Snort. Big Data challenges for alert correlation would be further exacerbated if additional diverse heterogeneous sources were utilized as the number of alerts could increase, in addition to the broader need to perform data fusion across the different sources.

Discussion

When it comes to Intrusion Detection for Big Heterogeneous Data, a great deal of future work can be done with research. If it can be presumed that the annual cost in the USA for cybercrime is $100 Billion annually [17] and that organizations spend about 26% on Detection [19], obviously it cannot be inferred that Intrusion Detection is costing $26 Billion annually (as the total national cost for Intrusion Detection cannot be directly extrapolated from amounts spent by organizations for Intrusion Detection and national Cybercrime costs). However, even if these numbers are only estimates it can be gauged in terms of the order of magnitude that the national costs for Intrusion Detection are very large. Analyzing much more heterogeneous security data can yield significant improvements to the Intrusion Detection domain, and employing Big Data technologies and techniques will allow more of this Big Heterogeneous Data to be utilized.

Promising future research in this spirit will be presented in the following main areas: Data Sets, Feature Selection, Data Fusion, SIEMs, Database Issues, and Other Architectural Considerations.

Data sets

Clearly, the relative lack of high-quality Intrusion Detection data sets is a problem for the academic research community. Some researchers have been building their own data sets due to the lack of existing relevant ones. However one difficulty with this approach is how to accurately label Intrusion Detection data as either attack or normal. Honeypots, honeynets, and sinkholes as well as intentionally cyber-attacking for the purposes of generating data sets might be able to help with labeling data as an attack, but there are still challenges with these approaches. Also, other traffic cannot always be assumed to be normal as it could also be contaminated with attack data. Synthesizing both attack and normal data might not build robust enough datasets in that they are not similar enough to the real world. Another issue is that data sets should constantly be updated as time progresses with new instances containing new normal traffic (i.e., new technologies, applications, and users) and attacks (i.e., new techniques or exploits) to keep research relevant as well as to better train learning systems for Intrusion Detection in an iterative fashion as technology and cyber-attacks evolve.

High quality, robust, and heterogeneously diverse public data sets are so fundamental to Intrusion Detection experimentation and research that the availability of such data sets is fundamentally a significant research problem in its own right. Actual research into the production of these data sets could allow researchers to make better overall progress in the Intrusion Detection arms race. Various techniques and issues such as those mentioned in the preceding paragraph and in the Intrusion Detection Data Set Challenges section should be considered when developing such data sets. Another important consideration is that Intrusion Detection data sets should also accommodate many diverse heterogeneous security sources to adequately address the issue of improving situational awareness. In addition, perhaps Intrusion Detection may need to use more than binary classification in some scenarios and use multiple classifications (i.e., “attack”, “normal”, “suspicious”, “unknown”, etc.).

Feature selection

The application of feature selection can be helpful for real-time intrusion detection in terms of reducing detection classification times and even sometimes improving classification accuracy, but care must be taken in the application of feature selection. It is important to understand where and how feature selection can be applied effectively, especially with regards to unstable Intrusion Detection data sets. More specifically, how can feature selection play a role in security event data that is constantly evolving with new characteristics due to issues such as new technologies as well as new cyber-attacks? Thus, the stability of feature sets is very significant in the cybersecurity domain where the landscape is extremely dynamic and not as static as some other domains (e.g., bioinformatics) where feature selection is applied. Feature selection could play a very prominent role where many diverse heterogeneous sources are being analyzed and correlated as feature reduction could significantly reduce storage Volume, processing Velocity, and complex Variety.

Data fusion

Data Fusion has not been widely adopted within the Intrusion Detection domain as compared to other domains like military applications that have a multitude of diverse heterogeneous sensors. A considerable amount of research has been conducted for alert correlation with Intrusion Detection, but very few works consider event fusion or other types of data fusion. Many more experiments with different data fusion techniques should be performed, especially in the context of many more diverse heterogeneous sources which contain Big Data. The question is therefore raised: can significant improvements be realized for Intrusion Detection by doing this, or would the costs be too high?

SIEMs

SIEM technology is proprietary and it is difficult to speculate on some of the internals of different product offerings, but it is apparent that SIEMs do attempt to normalize the data sources into as common formats as possible (e.g., SCAP, CBE, and CEE). While standardizing the formats of security events is good in the sense of reducing some of the more challenging data fusion aspects and to increase performance, in some cases it could possibly be detrimental if valuable information for a particular source is being lost or minimized. However, standardized message formats are not a reality in the real-world in many situations due to a wide variety of security products simply still using different formats. More experimentation into how SIEMs compare to traditional data fusion approaches might be beneficial. Obviously there are benefits to formatting all security event sources into as common formats as possible. However what are the drawbacks with the common format approach from actual experiments?

NIDSs and HIDSs still play a prominent role in the commercial sector. However SIEMs seem to have recently taken the security industry by storm for larger organizations with their ability to both detect cyber-attacks in the real-time, and provide offline security analytics and forensics. But the research community still seems too narrowly focused on the Intrusion Detection problem by conducting much more research on traditional IDSs than a more heterogeneous approach such as SIEMs. SIEMs should be evaluated much more by the research community, especially with regards to experimentation. Furthermore, the research community could further evaluate how to extend SIEM functionality in novel ways as SIEMs have not been Intrusion Detection's "silver bullet".

Database issues

When Big Data poses challenges for Intrusion Detection, whether it is in standalone subsystem event sources or major architectural components for a heterogeneous system, how can technologies like Hadoop and similar technologies improve upon the state of the art and solve pressing problems? More experimentation is needed in this regard. Specifically for SIEMs (or their research-based data fusion equivalents), how can we improve the efficiency of the offline repository stores (which are used for both forensics as well as offline security analytics which provide feedback to the online system?) Many SIEMs use traditional RDBMS technology for this purpose, and more experiments could be conducted to see how other approaches such as columnar databases, xml databases, Hadoop technologies, or hybrids thereof could make these repository stores more efficient and effective. Research into using better storage platforms effectively is needed for the enormous Volume, fast Velocity, and complex Variety processing requirements for Intrusion Detection.

Other architectural considerations

From a general conceptual framework, more experimentation is needed into what Bass [36] proposed, which explored how data mining from offline repositories can give useful feedback to effectively and efficiently benefit real-time Intrusion Detection. Bass's concept was fairly sophisticated and very highly cited by the research community. However there has not been significant experimental research with this model. A good deal of experimentation is still needed to explore the effectiveness of Bass's model with concepts such as real-time feedback and utilizing different Intrusion Detection "feature templates" based on the current situation. Similarly, SIEMs do afford archive repository storage for security analytics and forensics, and researchers could evaluate how to further improve the real-time Intrusion Detection component with this repository store.

In general, more experimentation is needed to illustrate the effectiveness of the cloud in solving Big Heterogeneous Data challenges for Intrusion Detection. More studies and experiments to illustrate the costs involved with utilizing the cloud platform would be beneficial in terms of network bandwidth, processing, and storage requirements. An important consideration is to determine what data gets filtered out, which is an important consequence in terms of establishing good ground truth for Intrusion Detection as well as regarding forensic capabilities.

Some good preliminary research has been conducted regarding architectural topology in terms of Centralized, Distributed, or Hybrid approaches for Sensors, IDSs, and Analyzers. Typically SIEMs will use a Centralized approach for the decision unit which possesses advantages and disadvantages. More experiments could be conducted to further refine these. Also, more research into how SIEM systems can scale out with multiple SIEMs would be beneficial as well. Some benefits have been realized with distributed or hierarchical architectures for IDSs and analyzers, but much more potential experimentation is yet to be realized especially in terms of utilizing more data fusion.

Heterogeneity among the actual Sensors, IDSs, Analyzers, or even SIEMs can be beneficial for Intrusion Detection where detection accuracy can be improved. For example, different IDSs working in teams to evaluate the same security events can improve detection accuracy, and the same can be accomplished with using more than one type of SIEM to analyze the same security events. So far, this type of Detector heterogeneity has showed good benefits. Even more research should be conducted into this area especially in terms of being sensitive to the additional costs versus the benefits.

Greater geographical and organizational heterogeneity should be employed for Intrusion Detection. Alert and event correlation beyond geographical and organizational boundaries could further improve situational awareness. Projects such as the Internet Storm Center [34] for honeypots and collecting malicious Intrusion Detection metadata. However organizations could participate much more actively. Will a natural evolution be to outsource the Intrusion Detection function to Managed Service Providers (MSPs)? Will these MSPs be able to more cost effectively manage Intrusion Detection through scaling out with a more comprehensive Intrusion Detection knowledge base and better core competency? More research should evaluate how significantly sharing of Intrusion Detection events, alerts, analysis, and knowledge across many organizations could enhance the state of the art. Are there ways for even competing organizations to share cyber-threat security event data for their collective good while still protecting their competitive interests? A prevalent consensus among the research community is that greater sharing for Intrusion Detection is necessary, and more research should be conducted to evaluate this on a significantly larger scale of sharing.

Conclusion

Historically most of the academic research for Intrusion Detection has focused too narrowly on the network layer with NIDs and to some extent at the host level with HIDSs. The academic research community should actively embrace a more diverse heterogeneous-based event source approach and follow the lead of the commercial sector where the rapid proliferation of SIEM technology has blossomed into a billion dollar industry (despite the term "SIEMS" having only been coined in 2005). This proliferation of SIEM technology throughout industry is an important consideration, given that one of the inherent features in this technology is to correlate security events from a wide array of diverse heterogeneous sources, and its successes in the commercial sector should give credence to this approach.

Both cybersecurity and physical security for organizations such as those in the utility and the industrial sector can even be enhanced by correlating traditional IT security events with those beyond cyberspace such as sensor devices measuring anomalous real-world quantities like gas leaks, electrical power/voltage/current, temperature, fire alarms, or many other sensors. Correlating security events from physical world sensors with cyberspace is becoming significantly more important as the utility and industrial sectors are becoming increasingly computerized for automation, and thus exposing their physical infrastructures to new cyber threats such as malicious attackers or "cyber accidents".

More diverse heterogeneous sources can provide for improved situational awareness within the Intrusion Detection domain similar to the military's use of diverse heterogeneous sources in its doctrines, strategies, tactics, and engagements. The onslaught of all this Big Input Data drives the engine for an onslaught of Big Output Data. For Intrusion Detection, great heterogeneous diversity in both its input and output data poses significant Big Heterogeneous Data challenges.

While Intrusion Detection does not always face Big Data challenges, it does face Big Data challenges more often as time progresses and especially more so for larger private and government organizations. This trend of Big Data challenges will continue as a multitude of more heterogeneous sources are analyzed. Even medium and smaller organizations will need to assess whether their Intrusion Detection architecture or Security Analytics merit the deployment costs of Big Data technologies. Big Data challenges for Intrusion Detection already exist for the nation's electric grid given its tight integration with computers, and will become even more pronounced as many more diverse cyber and non-cyber heterogeneous sources are brought online to enhance overall cyber defense and improve situational awareness for critical infrastructure.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RZ performed the primary literature review and analysis for this work, and also drafted the manuscript. RW worked with RZ to develop the article's framework and focus. TMK introduced this topic to RZ and RW, and coordinated the other authors to complete and finalize this work. All authors read and approved the final manuscript.

Acknowledgements
We thank the reviewers for their constructive comments.

Received: 6 August 2014 Accepted: 28 January 2015
Published online: 27 February 2015

References

1. Nassar M, al Bouna B, Malluhi Q (2013) Secure outsourcing of network flow data analysis. In: Big Data (BigData Congress), 2013 IEEE International Congress On. IEEE, Santa Clara, CA, USA. pp 431–432
2. Group BDW (2013) Big Data Analytics for Security Intelligence. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf. Accessed 2015-1-10
3. Chickowski E (2012) A Case Study In Security Big Data Analysis. http://www.darkreading.com/analytics/security-monitoring/a-case-study-in-security-big-data-analysis/d/d-id/1137299?. Accessed 2015-1-10
4. Chickowski E (2013) Moving Beyond SIEM For Strong Security Analytics. http://www.darkreading.com/moving-beyond-siem-for-strong-security-analytics/d/d-id/1141069?. Accessed 2015-1-10
5. Marko K (2014) Big Data: Cyber Security's Silver Bullet? Intel Makes the Case. http://www.forbes.com/sites/kurtmarko/2014/11/09/big-data-cyber-security/. Accessed 2015-1-10
6. Kezunovic M, Xie L, Grijalva S (2013) The role of big data in improving power system operation and protection. In: Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium. IEEE, Rethymno, Greece. pp 1–9
7. Software I (2013) Managing big data for smart grids and smart meters. http://www-935.ibm.com/services/multimedia/Managing_big_data_for_smart_grids_and_smart_meters.pdf. Accessed 2015-1-10
8. Modi C, Patel D, Borisaniya B, Patel H, Patel A, Rajarajan M (2013) A survey of intrusion detection techniques in cloud. J Netw Comput Appl 36(1):42–57
9. Zhou CV, Leckie C, Karunasekera S (2010) A survey of coordinated attacks and collaborative intrusion detection. Comput Secur 29(1):124–140
10. Laney D (2001) 3d data management: Controlling data volume, velocity and variety. Technical Report 949, META Group (now Gartner). http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
11. Zikopoulos P, Parasuraman K, Deutsch T, Giles J, Corrigan D (2012) Harness the power of big data The IBM big data platform. McGraw Hill Professional, New York, NY. http://books.google.com/books?id=HhSON0xOCQ0C
12. Frank J (1994) Artificial intelligence and intrusion detection: current and future directions. In: Proceedings of the 17th national computer security conference. Vol. 10. Citeseer, Baltimore, MD, USA. pp 1–12
13. Information Assurance Solutions Group (2015) Defense in depth. Technical report, National Security Agency. http://www.nsa.gov/ia/_files/support/defenseindepth.pdf. Accessed 2015-1-10
14. Denning DE (1987) An intrusion-detection model. Softw Eng IEEE Trans SE-13(2):222–232. doi:10.1109/TSE.1987.232894
15. Sourcefire (2015) Snort, Home Page. http://www.snort.org/. Accessed 2015-1-10
16. Roesch M (1999) Snort: Lightweight intrusion detection for networks. In: LISA. Vol. 99. USENIX, Seattle, WA, USA. pp 229–238
17. Center for Strategic and International Studies (2013) The economic impact of cybercrime and cyber espionage. Technical report. McAfee http://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime.pdf
18. Verizon RISK Team (2013) 2013 data breach investigations report. Technical report. Verizon http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf
19. Ponemon Institute LLC (2012) 2012 cost of cyber crime study: United states. Technical report. Ponemon Institute http://www.ponemon.org/local/upload/file/2012_US_Cost_of_Cyber_Crime_Study_FINAL6%20.pdf
20. Julisch K, Dacier M (2002) Mining intrusion detection alarms for actionable knowledge. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Edmonton, Alberta, Canada. pp 366–375
21. Xu D, Ning P (2008) Correlation analysis of intrusion alerts. Intrusion Detect Syst 38:65–92
22. Suthaharan S, Panchagnula T (2012) Relevance feature selection with data cleaning for intrusion detection system. In: Southeastcon, 2012 Proceedings of IEEE. IEEE, Orlando, FL, USA. pp 1–6
23. Bhatti R, LaSalle R, Bird R, Grance T, Bertino E (2012) Emerging trends around big data analytics and security: Panel. In: Proceedings of the 17th ACM Symposium on Access Control Models and Technologies. SACMAT '12. ACM, New York, NY, USA. pp 67–68. doi:10.1145/2295136.2295148. http://doi.acm.org/10.1145/2295136.2295148
24. Oltsik J (2013) Defining Big Data Security Analytics. Networking Nuggets and Security Snippets (Blog). http://www.networkworld.com/community/blog/defining-big-data-security-analytics. Accessed 2014-5-23
25. Sommer R, Paxson V (2010) Outside the closed world: On using machine learning for network intrusion detection. In: Security and Privacy (SP), 2010 IEEE Symposium On. IEEE, Oakland, CA, USA. pp 305–316
26. Coull SE, Wright CV, Monrose F, Collins MP, Reiter MK (2007) Playing devil's advocate: Inferring sensitive information from anonymized network traces. In: NDSS. Vol. 7. Internet Society, San Diego, CA, USA. pp 35–47
27. Azad C, Jha VK (2013) Data mining in intrusion detection: a comparative study of methods, types and data sets. Int J Inf Technol Comput Sci 5(8):75–90
28. McHugh J (2000) Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans Inf Syst Secur 3(4):262–294
29. Mahoney MV, Chan PK (2003) An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In: Recent advances in intrusion detection. Springer, Berlin Heidelberg. pp 220–237
30. Wu SX, Banzhaf W (2010) The use of computational intelligence in intrusion detection systems: A review. Appl Soft Comput 10(1):1–35
31. Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA (2012) Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput Secur 31(3):357–374. doi:10.1016/j.cose.2011.12.012

32. Fontugne R, Borgnat P, Abry P, Fukuda K (2010) Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In: Proceedings of the 6th International COnference. Co-NEXT '10. ACM, New York, NY, USA. pp 8–1812. doi:10.1145/1921168.1921179. http://doi.acm.org/10.1145/1921168.1921179
33. United States Marine Academy – West Point (2015) Cyber Research Center – DataSets. http://www.usma.edu/crc/SitePages/DataSets.aspx. Accessed 2015-1-10
34. Internet Storm Center (2015) Reports – Internet Security | SANS ISC. https://isc.sans.edu/reports.html. Accessed 2015-1-10
35. Song J, Takakura H, Okabe Y, Eto M, Inoue D, Nakao K (2011) Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation. In: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security. ACM, Salzburg, Austria. pp 29–36
36. Bass T (2000) Intrusion detection systems and multisensor data fusion. Commun ACM 43(4):99–105
37. Wang H, Liu X, Lai J, Liang Y (2007) Network security situation awareness based on heterogeneous multi-sensor data fusion and neural network. In: Computer and Computational Sciences, 2007. IMSCCS 2007. Second International Multi-Symposiums On. IEEE, Iowa City, IA, USA. pp 352–359
38. Tsang C-H, Kwong S, Wang H (2007) Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognit 40(9):2373–2391
39. Chebrolu S, Abraham A, Thomas JP (2005) Feature deduction and ensemble design of intrusion detection systems. Comput Secur 24(4):295–307
40. Chen Y, Li Y, Cheng X-Q, Guo L (2006) Survey and taxonomy of feature selection algorithms in intrusion detection system. In: Lipmaa H, Yung M, Lin D (eds). Information Security and Cryptology. Lecture Notes in Computer Science. Vol. 4318. Springer, Berlin Heidelberg. pp 153–167
41. Elngar A, Mohamed D, Ghaleb F (2013) A real-time anomaly network intrusion detection system with high accuracy. Inf Sci Lett Int J 2(2):49–56
42. The Apache Software Foundation (2015) Welcome to Apache Hadoop!. http://hadoop.apache.org/. Accessed 2015-1-10
43. Suthaharan S (2013) Big data classification: problems and challenges in network intrusion prediction with machine learning. In: Big Data Analytics Workshop, in Conjunction with ACM Sigmetrics. ACM, Pittsburgh, PA, USA
44. Whitworth J, Suthaharan S (2013) Security problems and challenges in a machine learning-based hybrid big data processing network systems. In: ACM Sigmetrics 2013 (Big Data Analytics Workshop). ACM, Pittsburgh, PA, USA
45. Jeong H, Hyun W, Lim J, You I (2012) Anomaly teletraffic intrusion detection systems on hadoop-based platforms: A survey of some problems and solutions. In: Network-Based Information Systems (NBiS), 2012 15th international conference on. IEEE, Melbourne, Australia. pp 766–770
46. Lee Y, Lee Y (2013) Toward scalable internet traffic measurement and analysis with hadoop. ACM SIGCOMM Comput Commun Rev 43(1):5–13
47. Cheon J, Choe T-Y (2013) Distributed processing of snort alert log using hadoop. Int J Eng Technol(0975-4024) 5(3):2685–2690
48. VeetiL S, Gao Q (2013) A real-time intrusion detection system by integrating hadoop and naive bayes classification. In: Dalhousie Computer Science In-house Conference (DCSI). Dalhousie University, Halifax, Canada
49. Bass T (1999) Multisensor data fusion for next generation distributed intrusion detection systems. In: IRIS National Symposium. IRIS National Symposium, Laurel, MD, USA
50. Lan F, Chunlei W, Guoqing M (2010) A framework for network security situation awareness based on knowledge discovery. In: Computer Engineering and Technology (ICCET), 2010 2nd international conference on. Vol. 1. IEEE, Chengdu, China. pp 1–226
51. Mitchell HB (2012) Data fusion: concepts and ideas. Springer, New York, NY. http://books.google.com/books?id=ZyPYGh-WAgYC
52. Hall DL, Llinas J (1997) An introduction to multisensor data fusion. Proc IEEE 85(1):6–23
53. Fessi B, Benabdallah S, Hamdi M, Rekhis S, Boudriga N (2010) Data collection for information security system. In: Engineering Systems Management and Its Applications (ICESMA), 2010 second international conference on. IEEE, Sharjah, United Arab Emirates. pp 1–8
54. Karim Ganame A, Bourgeois J, Bidou R, Spies F (2008) A global security architecture for intrusion detection on computer networks. Comput Secur 27(1):30–47
55. Bye R, Camtepe SA, Albayrak S (2010) Collaborative intrusion detection framework: characteristics, adversarial opportunities and countermeasures. In: Proceedings of CollSec: Usenix Workshop on Collaborative Methods for security and privacy. USENIX, Washington, DC, USA
56. Bartos K, Rehak M (2012) Self-organized mechanism for distributed setup of multiple heterogeneous intrusion detection systems. In: Self-Adaptive and Self-Organizing Systems Workshops (SASOW), 2012 IEEE sixth international conference on. IEEE, Lyon, France. pp 31–38
57. Cai H, Wu N (2010) Design and implementation of a dids. In: 2010 IEEE International Conference on Wireless Communications, Networking and Information Security. IEEE, Beijing, China. pp 340–342
58. Vincent Zhou C, Leckie C, Karunasekera S (2009) Decentralized multi-dimensional alert correlation for collaborative intrusion detection. J Netw Comput Appl 32(5):1106–1123
59. Metzger S, Hommel W, Reiser H (2011) Integrated security incident management–concepts and real-world experiences. In: IT Security Incident Management and IT Forensics (IMF), 2011 Sixth International Conference On. IEEE, Stuttgart, Germany. pp 107–121
60. Williams A (2007) The Future of SIEM – The market will begin to diverge. http://techbuddha.wordpress.com/2007/01/01/the-future-of-siem-%E2%80%93-the-market-will-begin-to-diverge/
61. Anuar NB, Papadaki M, Furnell S, Clarke N (2010) An investigation and survey of response options for intrusion response systems (irss). In: Information Security for South Africa (ISSA), 2010. IEEE, Johannesburg, South Africa. pp 1–8
62. Rouse M (2012) security information and event management (SIEM). http://searchsecurity.techtarget.com/definition/security-information-and-event-management-SIEM

63. Messmer E (2013) Gartner security report: McAfee up, Trend Micro down. http://www.networkworld.com/news/2013/053013-gartner-security-survey-270297.html
64. Mosaic Security Research Log Management & Security Information and Event Management (SIEM) Software Guide | Mosaic Security Research. http://mosaicsecurity.com/categories/85-log-management-security-information-and-event-management. Accessed 2014-5-23
65. Aguirre I, Alonso S (2012) Improving the automation of security information management: A collaborative approach. Secur Privacy IEEE 10(1):55–59
66. Kotenko I, Polubelova O, Saenko I (2012) The ontological approach for siem data repository implementation. In: Green Computing and Communications (GreenCom), 2012 IEEE international conference on. IEEE, Besancon, France. pp 761–766
67. Nicolett M, Kavanagh KM (2011) Critical capabilities for security information and event management technology. Gartner Report
68. Radack S, Kuhn R (2011) Managing security: the security content automation protocol. In: IT Professional. IEEE 9(13):9–11
69. Ogle D, Kreger H, Salahshour A, Cornpropst J, Labadie E, Chessell M, Horn B, Gerken J, Schoech J, Wamboldt M (2002) Canonical situation data format: the common base event v1.1.1. IBM Corporation. http://xml.coverpages.org/IBMCommonBaseEventV111.pdf. Accessed 2015-1-10
70. Distributed Management Task Force Inc (2014) Common Information Model (CIM). http://dmtf.org/standards/cim. Accessed 2014-5-23
71. Revelytix Inc. (2010) Triple store evaluation analysis report. Technical report, Revelytix. http://www.algebraixdata.com/wp-content/uploads/2014/02/Revelytix-Triplestore-Evaluation-Analysis-Results.pdf
72. Kotenko I, Chechulin A (2012) Common framework for attack modeling and security evaluation in siem systems. In: Green Computing and Communications (GreenCom), 2012 IEEE international conference on. IEEE, Besancon, France. pp 94–101
73. Kreutz D, Casimiro A, Pasin M (2012) A trustworthy and resilient event broker for monitoring cloud infrastructures. In: Distributed applications and interoperable systems. Springer, Berlin Heidelberg. pp 87–95
74. Splunk Inc. Operational Intelligence, Log Management, Application Management, Enterprise Security and Compliance | Splunk. http://www.splunk.com/. Accessed 2014-5-23.
75. Li Y, Liu Y, Zhang H (2012) Cross-boundary enterprise security monitoring. In: Computational Problem-Solving (ICCP), 2012 international conference on. IEEE, Leshan, China. pp 127–136
76. Blum D, Schacter P, Maiwald E, Krikken R, Henry T, de Boer M, Chuvakin A (2011) 2012 planning guide: Security and risk management. Technical Report G00224667 Gartner, Inc.
77. Sitaram D, Sharma M, Zain M, Sastry A, Todi R (2013) Intrusion detection system for high volume and high velocity packet streams: A clustering approach. Int J Innovation Manag Technol 4(5):480–485
78. Kaszuba G (2013) packetloop/packetpig. GitHub.0 https://github.com/packetloop/packetpig
79. Yen T-F, Oprea A, Onarlioglu K, Leetham T, Robertson W, Juels A, Kirda E (2013) Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks. In: Proceedings of the 29th Annual Computer Security Applications Conference. ACM, New Orleans, LA, USA. pp 199–208
80. Myers J, Grimaila MR, Mills RF (2011) Log-based distributed security event detection using simple event correlator. In: System Sciences (HICSS), 2011 44th Hawaii International Conference on. IEEE, Kauai, HI, USA. pp 1–7
81. Langner R (2011) Stuxnet: Dissecting a cyberwarfare weapon. Secur Privacy IEEE 9(3):49–51
82. Valdes A, Cheung S (2009) Intrusion monitoring in process control systems. In: System Sciences, 2009. HICSS'09. 42nd Hawaii international conference on. IEEE, Waikoloa, Big Island, HI, USA. pp 1–7
83. SRI International (2014) Detection and Analysis of Threats to the Energy Sector (DATES). http://www.csl.sri.com/projects/dates/. Accessed 2014-5-23
84. Valdes A (2010) Detection and analysis of threats to the energy sector: Dates. Technical report, SRI International
85. XU X-b, YANG Z-q, XIU J-p, LIU C (2013) A big data acquisition engine based on rule engine. J China Universities Posts Telecommunications 20:45–49
86. Ray PD, Reed C, Gray J, Agarwal A, Seth S (2012) Improving roi on big data through formal security and efficiency risk management for interoperating ot and it systems. In: Grid-Interop Forum 2012, Irving, Texas, USA
87. Gabriel R, Hoppe T, Pastwa A, Sowa S (2009) Analyzing malware log data to support security information and event management: Some research results. In: Advances in databases, knowledge, and data applications, 2009. DBKDA'09. First international conference on. IEEE, Cancun, Mexico. pp 108–113
88. Hunt R, Slay J (2010) The design of real-time adaptive forensically sound secure critical infrastructure. In: Network and System Security (NSS), 2010 4th International conference on. IEEE, Melbourne, Australia. pp 328–333
89. Sundaramurthy SC, Bhatt S, Eisenbarth MR (2012) Examining intrusion prevention system events from worldwide networks. In: Proceedings of the 2012 ACM workshop on building analysis datasets and gathering experience returns for security. ACM, Raleigh, NC, USA. pp 5–12
90. Roschke S, Cheng F, Meinel C (2010) A flexible and efficient alert correlation platform for distributed ids. In: Network and System Security (NSS), 2010 4th international conference on. IEEE, Melbourne, Australia. pp 24–31