Chapter 2
Artificial Intelligence and Big Data Analytics in Support of Cyber Defense

Louise Leenen
University of the Western Cape, South Africa & CAIR, South Africa

Thomas Meyer
University of Cape Town, South Africa & CAIR, South Africa

ABSTRACT

Cybersecurity analysts rely on vast volumes of security event data to predict, identify, characterize, and deal with security threats. These analysts must understand and make sense of these huge datasets in order to discover patterns which lead to intelligent decision making and advance warnings of possible threats, and this ability requires automation. Big data analytics and artificial intelligence can improve cyber defense. Big data analytics methods are applied to large data sets that contain different data types. The purpose is to detect patterns, correlations, trends, and other useful information. Artificial intelligence provides algorithms that can reason or learn and improve their behavior, and includes semantic technologies. A large number of automated systems are currently based on syntactic rules which are generally not sophisticated enough to deal with the level of complexity in this domain. An overview of artificial intelligence and big data technologies in cyber defense is provided, and important areas for future research are identified and discussed.

DOI: 10.4018/978-1-5225-8304-2.ch002

INTRODUCTION

Governments, military forces and other organisations responsible for cyber defence deal with vast amounts of data that has to be understood in order to lead to intelligent decision making. The rapid increase in the number and variety of cyber threats, and in the volume of information that has to be processed, integrated and understood to provide efficient counter-measures provide challenges to the defence community. Integration of information requires an encoded common vocabulary and shared understanding of the domain whilst the vast amounts of information pertinent to cybersecurity requires automation for processing and decision making. It is widely recognised that cyber defence requires the capabilities of artificial intelligence (AI) and big data processing (Vasudevan, 2018), (Graham, 2018) (Masimbuka, Grobler, & Watson, 2018). Big data provides the huge sets of data AI algorithms require to train data and to learn, i.e. to determine what normal behaviour is and thus to be able to detect abnormal events. These technologies are used for intrusion detection, malware classification and attribution, attack prediction and other applications. Artificial intelligence has made a resurgence in the past decade due to an underlying component, semantic technology. Semantic technologies represents a number of different technologies aiming to derive meaning from information. The combination of AI with big data capabilities go hand in hand to manage different data sets, to gain interoperability and insights and to make predictions[1]. One example of a limitation of current cybersecurity systems is that they tend to produce large numbers of false positives. Semantic technologies in conjunction with big data can improve this limitation due, in part, to recent advances in the scalability of techniques for managing semantic technologies.

Big data processing refers to data processing that is different from traditional processing technologies with respect to the volume of data, the rate at which data is data generated and rate at which data is transmitted, in addition to the fact that it includes both structured and unstructured data. Big data refers to volumes of data that are too large to handle by traditional data base systems. Big data analytics refers to advanced analytic techniques such as machine learning, predictive analysis, and other intelligent processing and mining techniques applied to big data sets. Big data analytics is required to combine different sources of information in order to recognise patterns for the detection of network attacks and other cyber threats. This must take place fast enough so that counter measures can be put in place. According to Saurabh, the CEO of a cybersecurity platform provider, cybersecurity in mature organisations depend on big data and in most cases it is big data that eliminates vulnerabilities and halt attacks (Saurabh, 2017).

Semantic technologies is a knowledge representation paradigm where the meaning of data is encoded separately from the data itself. The use of semantic technologies

such as logic-based systems to support decision making and an ability to process large sets of data have become essential. Hernandez-Ardieta & Tapiador (2013) state that it is virtually impossible for any organisation to manage cyber threats without collaboration with partners and allies. Collaboration includes sharing of threat related and cybersecurity information on a near real-time basis and this requirement necessitates the development of infrastructure and mechanisms to facilitate the information sharing, specifically through standardisation of data formats and exchange protocols. It is not merely *how* to share information but also *what*, with *whom* and *when* to share, as well as reasoning about the repercussions of sharing sensitive data. This level of collaboration will be impossible without attaching meaning to data and the ability to reason over formal structures. The use of ontologies is the underlying semantic technology driving the Semantic Web initiative (Berners-Lee, Hendler, & Lassila, 2001). Blockchain technology is another emerging technology in the cyber defence domain.

Issues that arise from the use of AI and big data are the protection of privacy and the ethical use of these technologies. It should also be kept in mind that attackers can also use these technologies to their advantage.

Section 2 covers background on semantic technologies, including ontologies, and blockchain technology. Section 3 discusses current applications of AI and big data analysis in cyber defence. Section 4 focuses on emerging trends in the AI and big data communities that are relevant in the cyber domain. The cyber defence community should take note of the necessity to perform research in these identified areas. The paper is concluded in Section 5.

BACKGROUND

Semantic Technologies

Semantic technologies is a term that represents a number of different technologies aiming to derive meaning from information. Some examples of such technologies are natural language processing, data mining, semantic search technologies, and ontologies. It should be noted that semantic technologies are not the same as Semantic Web technologies; the latter is a subset of the former. Semantic Web technologies are technology standards from the World Wide Web Consortium (WC3) that are aimed at the representation of data on the Web. Examples of Semantic Web technologies are RFD (Resource Description Framework) and OWL (Web Ontology Language). The Cambridge Semantics group (Bio, n.d.) defines semantic technologies as "… algorithms and solutions that bring structure and meaning to information" and Semantic Web technologies as "…those that adhere to a specific set of WC3 open

technology standards that are designed to simplify the implementation of not only semantic technology solutions but other kind of solutions as well".

The use of semantic technologies such as logic-based systems to support decision making and an ability to process large sets of data have become essential. The use of ontologies is the underlying semantic technology driving the Semantic Web initiative (Berners-Lee, Hendler, & Lassila, 2001) and the next section provides an overview of ontologies.

Ontologies

An ontology consists of a shared domain vocabulary and a set of assumptions about the meaning of terms in the vocabulary. A formal definition of an ontology is given by Gruber (1993): a "formal, explicit specification of a shared conceptualisation". An ontology is a technology that enables a formal, shared representation of the key concepts of a specific domain and it provides a way to attach meaning to the terms and relations used in describing the domain.

The main benefits of ontologies are the ability to perform semantic search, provision of a common shared vocabulary and sharing of domain knowledge, and the facilitation of semantic integration and interoperability between heterogeneous knowledge sources. Any satisfactory solution to search and integration problems will have to involve ways of making information machine-processable, a task that is only possible if machines have better access to the semantics of the information.

The information in an ontology is expressed in an ontology language (which are frequently logic-based languages), and then progressively refined. The construction and maintenance of ontologies greatly depend on the availability of ontology languages equipped with a well-defined semantics and powerful reasoning tools. Fortunately there already exists a class of logics, called description logics, which provide for both, and are therefore ideal candidates for ontology languages. The web ontology language, OWL 2.0, which was accorded the status of a W3C (World Wide Web Consortium) recommendation in 2009, is the official Semantic Web ontology language. OWL was designed to provide a common way to process the content of web information instead of just displaying it. There are a number of tools and environments available for building ontologies.

Blockchain Technology

A blockchain is a decentralised public ledger of every transition that has taken place and it cannot be changed retrospectively. It is a collection of records that is overseen and validated by a network of people. It is basically a database that is owned by a wider community rather than one central authority. Each block records a number of

transactions and blocks are chained together by containing hashes of the previous block. Blockchain technology is the backbone of cryptocurrencies but there are many other applications. Every user can add information to the chain and the data in a chain is secured through cryptography. One block contains a hash of the data in the previous block. (Martindale, 2018)

CURRENT APPLICATIONS OF AI AND BIG DATA IN CYBER DEFENCE

Cybersecurity experts are struggling to keep up with the pace and variety of attacks, and the massive volume of data on cyber events. AI is flexible, adaptive and provides a learning capability to deal with the increasing complexity of cybercrime (Masimbuka, Grobler, & Watson, 2018). Existing literature discuss the numerous methods AI provides for cyber defence: computational intelligence, neural networks, intelligent agents, artificial immune systems, genetic algorithms, machine learning, data mining, pattern recognition, fuzzy logic, heuristics, constraint solving, intelligent search, etc. (Dilek, Cakir, & Aydin, 2015) (Tyugu, 2011). The subsections below discuss applications of these methods.

Big data analytics in cyber defence focuses on the ability to gather massive amounts of digital information to process, analyse, visualise and interpret results in order to predict and stop cyber-attacks. Advance warning of attacks and threat intelligence are becoming essential in security technologies. According to the Gartner report released in 2014 (Litan, 2014), big data analytics play a crucial role in detecting crime and security incidents. The vice president of Gartner, Avivah Litan, said that big data analytics gives companies faster access to their own data than ever before. It also enables them to integrate different data sources to get an overall picture of threats against their institutions (Riviera, 2014). A Trend Micro white paper on big data security challenges stated that (Addressing Big Data Security Challenges: The Right Tools for Smart Protection., 2012) "Successful protection relies on the right combination of methodologies, human insight, an expert understanding of the threat landscape, and the efficient processing of big data to create actionable intelligence".

In 2013, Teradata sponsored the Ponemon Institute (Big Data Analytics in Cyber Defense, 2013) to perform an investigation on organisations' cybersecurity defences and their use of big data analytics in their study "Big Data Analytics in Cyber Defense". The report covered a wide range of pertinent issues and gave an overview of the awareness and the application of new data management and big data analytics of organisations in the fight against network attacks and other cyber threats at that time. Although 61% of the respondents agreed that launching a strong defence against hackers and other cyber criminals requires their organisations to be

able to detect and quickly contain anomalous and potentially malicious traffic in networks, at the time of the investigation, only 35% of respondents' organisations employed these tools. A positive result was that 51% of the organisations had knowledge of big data products that were available and regarded these tools as necessary in the fight against cyber threats. Only 23% of companies regularly applied big data analysis to counter threats but many organisations had plans to incorporate these tools in the future. The Ponemon Institute and Cloudera held a webinar on the state of cybersecurity big data analytics on 11 October 2016 (Ponemon, 2016). They found that organisations are 2.5 times more likely to identify a security risk within hours or minutes when they employ big data analytics. 81% of respondents indicated that demand for cybersecurity analytics had significantly increased over the year before this webinar took place and that heavy users of these technologies had a higher level of confidence in their ability to detect cyberincidents than light users. In the Ponemon Study of 2018 importance (Study on Global Megatrends in Cybersecurity, 2018), respondents were asked to rate the relevaive importance of specific technologies currently and in three years' time. AI in cyber defence was given a rating 31% currently but 71% in the future, while big analytics was given an importance rating of 33% now and 59% in the future. AI, analytics and threat intelligence feeds were the three technologies that, consistent with the previous Ponemon studies, been increasing in terms of perceived importance.

In April 2018, Kilpatrick Townsend teamed up with Ponemon to publish the Second Annual Study on Cybersecurity Risk to Knowledge Assets. This study found a dramatic increase in threats and awareness of threats to knowledge assets by oganisations. (Neditz, Dial, Jones, Grundman, & Bush, 2018).

The Cloud Security Alliance (Big Data Analytics for Security Intelligence, 2013) published a report on how the incorporation of big data is changing security analytics by providing new tools for leveraging data from both structured and unstructured sources. In this report, it is mentioned that people now create 2.5 quintillion bytes of data per day. The rate at which data is currently generated is creating a need for new technologies to analyse huge data sets. Big data analytics can be leveraged to correlate different sources in order to get a big picture. For example, financial transactions, log files and network traffic can be analysed to identify suspicious activities. The report pointed out that the urgency for collaborative research on big data topics were emphasised by the US Federal government's $200 million funding for big data research in 2012 (Lohr, 2012). The report points out that big data analytics can, for instance, advance security intelligence produced by Security Information and Event Management (SIEM) alerts by "reducing the time for correlating, consolidating, and contextualising diverse security event information, and also for correlating long-term historical data for forensic purposes". According to the report, big data tools provide an advantage in:

- More economical storage of large data sets;
- Much faster analysis;
- Being able to analyse and manage unstructured data; and
- Providing cluster computing infrastructures which are more reliable and available.

There are many examples of the use of big data analytics in cyber defence systems but we only mention two:

- **Intel's Threat Intelligence Exchange (TIE) system (Marko, 2014):** Multiple systems all share security information detected on one device or system in a centralised big data repository which then informs other devices and systems. Each security system then adapts their policies and controls to block a newly detected threat.
- **IBM's QRadar Security Intelligence platform and IBM Big Data Platform (IBM Security Intelligence with Big Data, n.d.)**: Provide threat and risk detection via an integrated approach that combines real-time correlation analytics across structured and unstructured data, and forensic capabilities for evidence. With this approach it is possible to address advanced persistent threats as well as fraud and insider threats. A wide range of data is analysed over years of activity.

A few specific applications of these technologies and methods are discussed in the following subsections but it is not intended to be a comprehensive overview.

Intrusion Detection Systems, Attack Classification and Prediction

A quick reaction to a network attack is one of the most essential requirements in cyber defence. When a system can identify an ongoing attack and classify the attack, efficient counter-measures can be taken. Balepin et al. (Balepin, Maltsev, Rowe, & Levitt, 2003) highlighted the need for quick responses with the increase in the speed of computer attacks. Various researchers have developed applications to identify and classify network attacks. Dilek, Cakir and Aydin (2015) give an overview AI-based Intrusion Detection systems. Alrajeh and Lloret ((2013) used AI for an Intrusion detection system in wireless networks. McAlwee et al. (MecElwee, Heaton, Fraley, & Cannady, 2017) employed deep learning to prioritise and respond to intrusion detection alerts.

A number of researchers implemented ontology-based systems for Intrusion Detection Systems (IDS) and network attack prediction. Bhandari and Guiral (2014) developed an ontology to perceive the security status of a network. van Heerden, Leenen and Irwin (2013) developed a network attack ontology to support the automated classification of attacks.

Malware Classification

The classification of malware is a very complex discipline due to the fact that there does not exist clear boundaries for the different groups of malware; characteristics are often shared by different types of malware. Many attributes and state changes have to be considered to detect a piece of malware; this complexity also results in problems with the naming and the classification of malware. Good classification and naming schemes support the sharing of information across organisations, facilitate the detection of new threats, and assist with risk assessment in quarantine and clean-up (Bailey, et al., 2007). Bailey et al. also highlight that the complexity of modern malware makes the classification process increasingly difficult, especially in terms of consistency and completeness. Another problem is the rapid increase in the number and diversity of Internet malware. There are a number malware naming schemes, for example the CARO scheme, but there does not exist a commonly accepted standard scheme. Automated malware detection and classification systems currently classify malware inconsistently across products, and their results tend to be incomplete (Bailey, et al., 2007). Some of the available analysis systems are Cuckoo Sandbox, Malwr, VirusTotal, and Yarae. One of the major problems with these classification systems is that they are inconsistent, incomplete and fail to be concise in their semantics (Bailey, et al., 2007).

The first recommendation in the JASON cyber report (Science of Cybersecurity, 2010) is that the cybersecurity community should develop vocabularies and ontologies such that a common language and a set of basic concepts can be developed for a shared understanding. This report was commissioned by the United States Department of Defense. Mundie and McIntire (2013) state that: "Nowhere in the cybersecurity community is the lack of a common vocabulary, and the problems it causes, more apparent than in malware analysis." Mundie also stated on a podcast (Mundie & Allen, n.d.): "And in my view, all the other aspects of a science – the statistics, hypothesis testing, etc. – all of that can only be built on top of that shared understanding that the report highlighted."

Wang and Zhong (2017) and Botes, Leenen and De La Harpe (2017) used machine learning techniques for malware classification. A growing number of researchers are investigating the use of semantic technologies to develop more efficient malware

classification systems (Mundie & McIntire, 2013) (Huang, Chuang, Tsai, & Lee, 2010) (Chiang & Tsuar, 2010).

Ontology-Based Approaches

The modern military environment is faced with an overwhelming amount of information from heterogeneous sources that has to be processed, integrated, interpreted, and exploited in order to gain situational awareness. The development and application of military related ontologies have grown tremendously the past 15 years. Curts and Campbell (2005) stated that "…the sorts of semantic interoperability provided by ontology technology are indispensable" in attempting to improve our understanding of Command and Control. A number of efforts have been devoted to developing ontologies for military applications and we mention a few of these below. Lombard, Gerber and van der Merwe (2012) developed an ontology for countermeasures against military aircraft. Belk & Noyes (2012) used an ontology to categorise all operations in cyber space. There are a number of ontology-based approaches to counter-terrorism (Turner & Weinberg, 2011) (Chmielewski, Galka, Jarema, Krasowski, & Kosinki, 2009).

Orbst, Chase and Markeloff (2012) have done work in support of the development of an ontology of the cybersecurity domain that will enable data integration across disparate data sources. They propose a number of resources for the envisioned ontology that range from domain specific resources, languages, vocabularies, ontologies and schemas. Their ontology is currently focussed on malware but they propose the inclusion of actors, victims, infrastructure, and capabilities.

Ontologies have been developed to support cybersecurity policy implementation: Jansen van Vuuren, Leenen and Zaaiman (2014) developed an ontology to support the implementation of the South African National Cybersecurity Policy. Due to the many role players, functions and relations that are involved in such an implementation, the authors present an ontology to represent the environment in which the policy implementation is to be done. Cuppens-Boulahia et al. (Cuppens -Boulahia, Cuppens, de Vergara, & Vazquez, 2008) proposed an ontology-based approach to instantiate new security policies to counteract network attacks.

Oltramari & Lebiere (2013) represent requirements for building a cognitive system for decision support with the capability of simulating defensive and offensive cyber operations by employing a semantic approach that includes the use of ontologies.

EMERGING RESEARCH AREAS

One of the necessary steps to obtain interoperability is to encourage research disciplines such as the big data and linked data communities to collaborate with the semantics research community. Linked data refers to a way of representing structured data so that it can be interlinked and become enhanced. There is a need for more research to be done in this area: Grobelnik, Mladenic and Fortuna (2012) performed a quick test in 2012 by looking at the number of hits for key words such as "big data" (20 million), "semantic web" (9 million) and "big data & semantic web" (0.3 million). They also searched for the number of appearances of "semantic" in the four leading books published in 2011 on "big data" and found very few incidences.

Janssen & Grady (2013) explored the use of big data technologies augmented by ontologies to improve cybersecurity. They note that these technologies have the potential to revolutionise the handling of large volumes of cyber data. One way in which big data analytics will be effective in the cyber domain is to identify patterns rather than processing collections of pages. Janssen and Grady also maintain that semantic technologies are crucial for the handling of big data sets across multiple domains. Little inroads have yet been made to integrate big datasets. These researchers argue that integration ontologies will have to be developed to provide metadata for browsing and querying: the integrating ontology should automatically construct queries to the big data repository. A significant challenge in using ontologies for automated data analytics across data sets that requires attention is probabilistic reasoning. This is due to the fact that analysis will have to be done under some uncertainty.

Although there are a number of emerging trends in both the semantic research and big data communities, we focus on four main trends relevant for cyber defence: the creation of interoperability and platforms for the sharing of information, the importance of intelligence-led approaches, blockchain and disruptive technologies. Scalable reasoning methods and stream reasoning are two emerging areas in the semantics community that should be noted by the cyber defence community. These methods can support the building of more efficient cyber defence systems.

Cybersecurity Information Sharing, Knowledge Representation and Interoperability

[The] Semantic Web in its most general aim is about interoperability being needed in almost all areas of research and business (Grobelnik, Mladenic, & Fortuna, 2012).

Formal models of cybersecurity information, vocabularies, standardised representations, data formats and exchange protocols are required to share

cybersecurity information effectively in the cybersecurity community. Significant effort has been made to categorise cybersecurity information and standardise data formats and protocols (Hernandez-Ardieta & Tapiador, 2013). According to Dandurand and Serrano (2013) current practices and supporting technologies limit the ability of organisations to share information securely with trusted partners. These authors give an overview of a number of cybersecurity standards and initiatives that have been developed such as the European Information Sharing and Alert System (EISAS (enhanced) report on implementation., 2011) and languages and structures developed by the MITRE corporation: Common Vulnerabilities and Exposures (CVE), Common Platform Enumeration (CPE) and others (Martin, 2008). Adoption of standards is improving but recently, subject-matter experts from the RSA organisation stated that: "Data standards for describing and transmitting threat information have advanced significantly, but much progress is needed to extend existing standards and drive wider adoption in vendor solutions. Threat information-sharing and collaboration programs help organizations augment their expertise and capabilities in detecting and remediating advanced threats, but most sharing programs are hindered by a heavy reliance on manually intensive, non-scalable processes and workflows." (Hartman, 2012).

Janowics & Hitzler (2012) cite the usefulness in publication of own data as one of the examples of the added value of semantics: the creation of intelligent metadata enables researchers to support the discovery and reuse of their data. They also stress the shift from developing increasingly complex software to the creation of metadata, and that smart data will make all future applications more usable, flexible and robust. Ontologies should be used to restrict the interpretation of domain vocabularies towards their intended meaning and reduce the risk of combining unsuitable data and models, something which purely syntactic approaches or natural language representation often fail to do (Kuhn, 2005).

The availability of cybersecurity datasets for machine learning and other AI-based systems is an issue that is important for the creation of benchmarks. Yavanoglu and Aydos (2017) raise the problem that although there are several studies on particular datasets, there are limited studies on the comprehensive state of security related datasets. They give a comprehensive review of publicly available datasets as well as an assessment of AI and machine learning applications using these sets.

Numerous organisations across the globe detect and gather information regarding cyber-attacks, network intrusions and malware. Standard, shared systems should be developed to collate and encourage information sharing to enable improved protection against cyber events. However, due to the vast amount of information and the speed at which cyber-attacks take place, timely decision making and automated responses are required and the use of ontologies to accomplish this goal is important (Dandurand & Serrano, 2013). A 2008 review of existing security ontologies stated

that the security community requires a complete security ontology that addresses insufficiencies in existing ontologies and provides reusability, communication and knowledge sharing (Blanco, et al., 2008). Similarly, there should be a standard malware classification system and vocabulary.

Orbst, Chase and Markeloff (2012) has made an attempt at creating an ontology for the cyber domain. They are using an initial ontology that is mainly focussed on malware but present a discussion of the development of an ontology for the whole domain. They give a description of the potential ontologies and standards that can be used in the global ontology. These resources include cyber and malware standards, schemas and technologies, foundational or upper ontologies, utility ontologies. An overview of the possible architecture is also given. Janssen & Grady (2013) also proposed the development of a cyber domain ontology that will contain all knowledge necessary for assessment, decision, planning and response in this domain. They base their proposal on the fact that system awareness currently resides in the minds of large numbers of cyber professionals. This information should be gathered in a single repository. Although it is a daunting task, the researchers argue that the recent successes of ontology engineering and the high stakes in the cybersecurity domain makes it necessary to solve on a national level. This argument can also be applied on an international level in the view of the authors of this paper.

There are issues such as trust and willingness to share which will also have to be addressed.

Intelligence-Led Approaches

Intelligence-led security is depicted by the *Information Age* as one of the 11 trends that will dominate cyber security in 2016 (Rossi, 2015). Intelligence-led security approaches in cybersecurity will be able to produce better results in terms of tracking security incidents and analysing huge amounts of information. Traditional technologies cannot cope with the rate at which information is generated and are unable to tie together unlinked pieces of information in order to create situational awareness. Real-time monitoring, advance warning and speedy analysis time will support quick reaction to security incidents.

The research director of Gartner, Lawrence Pingree, said the lack of automated intelligence sharing prevents human and business processes from responding to breaches. Pingree also said 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" (Marko, 2014).

According to the Ponemon Institute's 2015 and 2018 cybersecurity studies (Study on Global Megatrends in Cybersecurity, 2018) (Global Megatrends in Cybersecurity,

2015), technology innovation will shift towards big data analytics, forensics, intelligence-based cyber solutions, AI and threat intelligence feeds. They predict that the following technologies will gain the most in importance over the next three years: encryption for data at rest, big data analytics, SIEM and cybersecurity intelligence, automated forensics tools, encryption for data in motion, next generation firewalls, web application firewalls, threat intelligence feeds and sandboxing or isolation tools.

Blockchain Technology for Cybersecurity

Blockchain technology can protect data from tampering. Every participant in a blockchain has a copy of the list of transactions. This distributed basis of the chain makes tampering and modification to data very difficult because every action on the chain is fully transparent. (Wolfson, 2018)

Applications of the technology in cybersecurity include (Wolfson, 2018) (Horbenko, n.d.):

- One example of an application of this technology is the protection of datalogs.
- The prevention of fraud and data theft. A hacker would have to take down a whole chain which is almost impossible; each node will continue to verify data in that chain.
- Preventing Distributed Denial of Service (DDOS) attacks by fully decentralising the Domain Name System (DNS).

Disruptive Technologies

Disruptive technologies that can pose cyber risks are for examples the Internet of Things (IoT), the acceptance of virtual currencies, the use of AI, big data analytics, the use of drones and the use of cloud services. According to the latest Ponemon study on Cybersecurity (Study on Global Megatrends in Cybersecurity, 2018), enabling and disruptive technologies are currently having the second and fifth greatest impact cybersecurity, respectively. In the same study, disruptive technologies are rated to pose the biggest threat to cybersecurity over the next three years. The impact and reduction of these risks should receive attention from the cybersecurity domain.

Scalable Reasoning Methods

Scalability is a feature of a system that enables it to accommodate growth. The primary purpose of providing meaning to data is to facilitate reasoning about the data, so as to be able to perform sophisticated tasks such as intelligent search and data integration. Reasoning is an expensive computational endeavour, however.

One of the major challenges in this regard is the development of scalable reasoning methods. In recent years there have been a number of breakthroughs in the design of scalable ontology languages. The most important of these are the three profiles of the Web Ontology Language OWL 2: OWL 2 EL, OWL 2 DL, and OWL 2 RL (Motik, et al., n.d.). All three profiles are sub-languages of OWL 2, each designed expressly for representing a particular class of ontologies. The focus on specific classes of ontologies makes it possible to design reasoning methods with very attractive computational properties. To get a sense of the difference between the three profiles, it is important to understand that there is a distinction to be drawn between data and an ontology, the latter being used to provide meaning to the data.

OWL 2 EL is designed for scenarios in which the ontology is large and complicated, but with fairly small amounts of data underlying it. A representative example of a large OWL 2 EL ontology is the medical ontology SNOMED CT[2], with more than 300 000 active concepts and more than 1 000 000 relationships between the concepts. With SNOMED being represented as an ontology in OWL 2 EL, modern reasoning methods are able to classify all the concepts in SNOMED CT within a matter of milliseconds – a feat that was considered impossible about 15 years ago.

OWL 2 DL, on the other hand, is designed for cases in which an ontology is relatively small but spans large amounts of data. It is frequently used by employing the ontology as a semantic layer into which large database systems are being plugged. This enables users to query a database through the semantic layer, thereby obtaining truly intelligent responses from the system. The power of OWL 2 DL querying lies in the development of techniques where queries posed through the ontology are rewritten as standard database queries. This makes it possible to exploit existing efficient database querying methods, and has the potential for very fast and efficient querying.

Finally, OWL RL exploits the fact that many domains of interest can be represented using rule-like statements, and adopts existing techniques for reasoning efficiently with rule-based systems. OWL 2 RL is aimed at applications that require scalable reasoning without sacrificing too much expressive power. It is designed to accommodate OWL 2 applications that can trade the full expressivity of the language for efficiency, as well as RDF(S) applications that need some added expressivity. OWL 2 RL reasoning systems can be implemented using rule-based reasoning engines. The ontology consistency, class expression satisfiability, class expression subsumption, instance checking, and conjunctive query answering problems can be solved in time that is polynomial with respect to the size of the ontology. The RL acronym reflects the fact that reasoning in this profile can be implemented using a standard Rule Language.

One of the recent advances regarding scalability is the recognition that the so-called knowledge graphs used in many practical applications, and increasingly being used by big players such as Google and Facebook can be considered as a semantic technology. Knowledge graphs can be considered as shallow versions of ontologies (Krötzsch, 2017). The inexpressivity inherent in these representations provides for some limitations, but allow for enormous scalability. In doing so, the technology dovetails well with the availability of large amounts of data. In short, the advantage of the use of knowledge graphs is perfectly summarised by the phrase "Ontologies meets Big Data".

Stream Reasoning

Most of the currently available semantic technologies are based on the assumption that information is static. This is, of course, not a realistic assumption, and one of the important trends in this area is the development of tools able to deal with dynamic information that changes over time. A particularly useful scenario to consider is one where an incremental flow of data is available. Examples of this include data obtained from sensor network monitoring, traffic engineering, RFID tags applications, telecom call recording, medical record management, financial applications, and clickstreams, and are frequently referred to as streams of data. Clearly, information needed for ensuring cybersecurity falls in this category as well. Reasoning over such streams of data is referred to as stream reasoning (Della Valle, Ceri, van Harmelen, & Fensel, 2009). The goal of stream reasoning is to draw relevant conclusions and react to new situations with minimal delays. It is needed to support a variety of important functionalities in autonomous systems such as situation awareness, execution monitoring, and decision-making.

What is needed for efficient, intelligent stream reasoning is the provision of the abstractions, foundations, methods, and tools required to integrate data streams and existing reasoning systems, and there is broad consensus that the ability to reason about streaming data to cope with the increasing amount of dynamic data on the web is the next big step in semantic technologies. The research agenda for this challenge has been picked up by a number of research groups internationally ((Stuckenschmidt, Ceri, Della Valle, & van Harmelen, 2010). At its core is the goal to combine existing semantic technologies with data streams in order to perform stream reasoning. Work has been done on the foundations of real-time reasoning on data streams as they become available (Beck, Dao-Tran, Eiter, & Fink, 2014). It has also led to alternative abstractions for representing and querying semantic streams of data. Various forms of deductive and inductive stream reasoning have

been investigated (Barbieri, et al., 2010). In terms of improving the efficiency of stream reasoning methods, the exploitation of the temporal order of data streams has been recognised as a key optimisation method for stream reasoning. In a similar vein, parallelisation and distribution techniques for stream reasoning have been investigated (Albeladi, 2012).

CONCLUSION

This paper considers the application of big data analytics and semantic technologies for cyber defence by giving an overview of the current state of affairs, and identifying emerging trends in the combination of these fields. Big data analytics provides the ability to deal with large sets of diverse data, structured and unstructured, in almost real-time while semantic technologies provide the ability to make sense of the resulting information. Semantic technologies allow one to tie together seemingly unrelated pieces of information. The emerging trends also serve as the authors' recommendations for future research areas in the domain.

REFERENCES

Addressing Big Data Security Challenges: The Right Tools for Smart Protection. (2012). *Trend Micro.* Retrieved from http://www.trendmicro.com/cloud-content/us/pdfs/business/white-papers/wp_addressing-big-data-security-challenges.pdf

Albeladi, R. (2012). Distributed Reasoning on Semantic Data Streams. In *11th International Semantic Web Conference (ISWC) in Lecture Notes in Computer Science. 7650* (pp. 433-436). Berlin: Springer.

Alrajeh, N., & Lloret, J. (2013). Intrusion Detection Systems Based on Artificial Intelligence Techniques in Wireless Sensor Networks. *International Journal of Distributed Sensor Networks*.

Bailey, M., Oberheide, J., Andersen, J., Mao, Z. M., Jahanian, F., & Nazario, J. (2007). Automated Classification and Analysis of Internet Malware. In *International Symposium on Recent Advances in Intrusion Detection (RAID'07)* (pp. 187-197). Springer. 10.1007/978-3-540-74320-0_10

Balepin, I., Maltsev, S., Rowe, J., & Levitt, K. (2003). Using specification-based intrusion detection for automated response. In *Recent Advances in Intrusion Detection* (pp. 136–154). Berlin: Springer. doi:10.1007/978-3-540-45248-5_8

Barbieri, D., Braga, D., Ceri, S., Valle, E., Huang, Y., Tresp, V., ... Wermser, H. (2010). Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics. *IEEE Intelligent Systems*, *25*(6), 32–41. doi:10.1109/MIS.2010.142

Beck, H., Dao-Tran, M., Eiter, T., & Fink, M. (2014). Towards a Logic-Based Framework for Analyzing Stream Reasoning. In *3rd International Workshop on Ordering and Reasoning* (*Vol. 1303)*. CEUR-WS.org.

Belk, R., & Noyes, M. (2012). *On the Use of Offensive Cyber Capabilities* (Master's thesis). Harvard Kennedy School. Retrieved from http://www.dtic.mil/docs/citations/ADA561817

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 28–37. doi:10.1038cientificamerican0501-34 PMID:11341160

Bhandari, P., & Guiral, M. (2014). *Ontology Based Approach for Perception of Network Security State. In Recent Advances in Engineering and Computational Sciences* (pp. 1–6). RAECS.

Big Data Analytics for Security Intelligence. (2013). *The Cloud Security Alliance (CSA)*. Retrieved from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf

Big Data Analytics in Cyber Defense. (2013). *Ponemon Institute*. Retrieved from http://www.ponemon.org/local/upload/file/Big_Data_Analytics_in_Cyber_Defense_V12.pdf

Bio, L. (n.d.). *Semantic Web vs. Semantic Technologies*. Retrieved from Cambridge Semantics: http://www.cambridgesemantics.com/semantic-university/semantic-web-vs-semantic-technologies

Blanco, C., Lasheras, J., Valencia-García, R., Fernández-Medina, E., Toval, A., & Piattini, M. (2008). A Systematic Review and Comparison of Security Ontologies. *Third International Conference on Availability, Reliability and Security (ARES 08)*, 813-820. 10.1109/ARES.2008.33

Botes, F., Leenen, L., & de La Harpe, R. (2017). Ant Tree Miner Amyntas: Automatic, Cost-Based Feature Selection for Intrusion Detection. *Journal of Information Warfare*, *16*(4), 73–92.

Chiang, H.-S., & Tsuar, W.-J. (2010). Ontology-based Mobile Malware Behavioural Analysis. *IEEE Second International Conference on Social Computing (SocialCOM 2010)*. 10.1109/SocialCom.2010.160

Chmielewski, M., Galka, A., Jarema, P., Krasowski, K., & Kosinki, A. (2009). *Semantic Knowledge Representation in Terrorist Threat Analysis for Crises Management Systems*. Military University of Technology.

Cuppens-Boulahia, N., Cuppens, F., de Vergara, J., & Vazquez, E. (2008). An Ontology-based Approach to react to Network Attacks. *3rd International Conference on Risks and Security of Internet and Systems (CRiSIS '08)*, 280-305.

Curts, R., & Campbell, D. (2005). Building an Ontology for Command & Control. *10th International Command and Control Research and Technology Symposium*, McLean, VA.

Dandurand, L., & Serrano, O. (2013). Towards Improved Cyber Security Information Sharing. In *5th International Conference on Cyber Conflict.* Talinn: NATO CCD COE Publications.

Della Valle, E., Ceri, S., van Harmelen, F., & Fensel, D. (2009). It's a Streaming World! Reasoning Upon Rapidly Changing Information. *Intelligent Systems*, *9*(6), 83–89. doi:10.1109/MIS.2009.125

Dilek, S., Cakir, K., & Aydin, M. (2015). Applications of Artificial Intelligence Techniques to Combating Cyber Crimes. *International Journal of Artificial Intelligence & Applications*, *6*(1), 29–39. doi:10.5121/ijaia.2015.6102

EISAS (enhanced) report on implementation. (2011). *ENISA*. Retrieved from https://www.enisa.europa.eu/activities/cert/other-work/eisas_folder/eisas-report-on-implementation-enhanced

Global Megatrends in Cybersecurity. (2015). Ponemon Institute.

Graham, A. (2018, March 22). *Artificial Intelligence in Cyber Security*. Retrieved from IT Governance: https://www.itgovernance.co.uk/blog/artificial-intelligence-in-cyber-security/

Grobelnik, M., Mladenic, D., & Fortuna, B. (2012). Semantic Web in 10 years. *11th International Semantic Web Conference (SWC2012) Workshop on "What will the Semantic Web look like in 10 years from now",* Boston, MA.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. doi:10.1006/knac.1993.1008

Hartman, B. M. (2012). *Breaking Down Barriers to Collaboration in the Fight Against Advanced Threats.* RSA Security Brief February 2012. Retrieved from http://www.emc.com/collateral/industry-overview/11652-h9084-aptbdb-brf-0212-online.pdf

Hernandez-Ardieta, J., & Tapiador, J. (2013). Information Sharing Models for Cooperative Cyber Defence. *5th International Conference on Cyber Conflict.* NATO CCD COE Publications.

Horbenko, Y. (n.d.). *Using Blockchain Technology to Boost Cyber Security*. Retrieved from Steel Wiki: https://steelkiwi.com/blog/using-blockchain-technology-to-boost-cybersecurity/

Huang, H.-D., Chuang, T.-Y., Tsai, Y.-L., & Lee, C.-S. (2010). Ontology-based Intelligent System for Malware Behaviour Analysis. *The International Conference on Fuzzy Systems (FUZZ)*, 1-6.

IBM Security Intelligence with Big Data. (n.d.). Retrieved from IBM: http://www-03.ibm.com/security/solution/intelligence-big-data/

Janowicz, K., & Hitzler, P. (2012). *Key Ingredients for Your Next Semantics Elevator Talk.* Berlin: Springer.

Jansen van Vuuren, J., Leenen, L., & Zaaiman, J. (2014). Using an Ontology as a Model for the Implementation of the National Cybersecurity Policy Framework for South Africa. *9th International Conference on Cyber Warfare & Terrorism (ICCWS 2014)*.

Janssen, T., & Grady, N. (2013). Big Data for Combating Cyber Attacks. *Semantic Technology for Intelligence, Defense, and Security (STIDS 2013).*

Krötzsch, M. (2017). Ontologies for Knowledge Graphs. *30th International Workshop on Description Logics*.

Kuhn, W. (2005). Geospatial semantics: Why, of what, and how? *Journal on Data Semantics, 3534*.

Leenen, L., & Meyer, T. (2016). Semantic Technologies and Big Data: Analytics for Cyber Defence. *International Journal of Cyber Warfare & Terrorism*, *6*(3), 53–64. doi:10.4018/IJCWT.2016070105

Litan, A. (2014, January 14). *Reality Check on Big Data Analytics for Cybersecurity and Fraud*. Retrieved from Gartner: https://www.gartner.com/doc/2651118

Lohr, S. (2012, March 29). *New U.S. Research will Aim at Flood of Digital Data*. Retrieved from The New York Time: from http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html?

Lombard, N., Gerber, A., & van der Merwe, A. (2012). Using Formal Ontologies in the Development of Countermeasures for Military Aircraft. *Eighth Australasian Ontology Workshop*.

Marko, K. (2014, November 9). *Big Data: Cyber Security's Silver Bullet? Intel makes the case*. Retrieved from Social Innovation Reports: http://www.forbes.com/sites/kurtmarko/2014/11/09/big-data-cyber-security/#5d833caa294e

Martin, R. (2008). Making Security Measurable and Manageable. In *IEEE Military Communications Conference (MILCOM 2008* (pp. 1-9). IEEE.

Martindale, J. (2018, August 3). *What is a Blockhain?* Retrieved from Digital Trends: https://www.digitaltrends.com/computing/what-is-a-blockchain/

Masimbuka, M., Grobler, M., & Watson, B. (2018). Towards and Artificial Intelligence Framework to Actively Defend Cyberspace. In *European Conference on Cyber Warware and Security.* ACPI.

MecElwee, S., Heaton, J., Fraley, F., & Cannady, J. (2017). Deep Learning for Prioritizing and Responding to Intrusion Detection Alerts. *Cyber Security and Trusted Computing, 3*.

Motik, B., Cuenca Grau, B. I. H., Whu, Z., Fokoue, A., & Lutz, C. (n.d.). *OWL 2 Web Ontology Language Profiles*. Retrieved from http://www.w3.org/TR/owl2-profiles

Mundie, D., & Allen, J. (n.d.). *Using a Malware Ontology to Make Progress Towards a Science of Cybersecurity*. Retrieved from CERT: http://www.cert.org/podcast/show/20130509mundie.html

Mundie, D., & McIntire, D. M. (2013). An Ontology for Malware Analysis. In *Eighth International Conference on Availability, Reliability and Security* (pp. 556-558). IEEE.

Neditz, J., Dial, A., Jones, J., Grundman, J., & Bush, J. (2018). *Companies Begin to Protect Their Key Assets.* Retrieved from Kilpatrick Townsend and Ponemon Institute: https://www.kilpatricktownsend.com/en/Insights/Publications/2018/4/2018-Ponemon-Survey

Oltramari, A., & Lebiere, C. (2013). Towards a Cognitive System for Decision Support in Cyber Operations. Semantic Technology for Intelligence, Defense, and Security (STIDS 2013), 49-56.

Orbst, L., Chase, P., & Markeloff, R. (2012). Developing an Ontology of the Cyber Security Domain. *Semantic Technology for Intelligence, Defense, and Security (STIDS 2012).*

Ponemon, L. (2016). *Ponemon Institute and Cloudera announce a webinar on the state of cybersecurityBig Dataanalyctics on October 11 at 10am/1pm ET*. Ponemon Institute and Cloudera.

Riviera, J. (2014, February 6). *The Cloud Times. By 2016, 15 Percent of Large Global Companies will have Adopted Big Data Analytics for at least one Security or Fraud Detection Use Case*. Retrieved from Gartner: http://www.gartner.com/newsroom/id/2663015

Rossi, B. (2015, December 4). *11 Trends that will Dominate Cyber Security in 2016*. Retrieved from Information Age: http://www.information-age.com/technology/security/123460617/11-trends-will-dominate-cyber-security-2016

Saurabh, K. (2017, October 6). *7 Surprising Facts about AI and Big Data in Cybersecurity*. Retrieved from insideBIGDATA: https://insidebigdata.com/2017/10/06/7-surprising-facts-ai-big-data-cybersecurity/

Science of Cybersecurity. (2010). *MITRE Corporation*. Retrieved from http://fas.org/irp/agency/dod/jason/cyber.pdf

Stuckenschmidt, H., Ceri, S., Della Valle, E., & van Harmelen, F. (2010). Towards Expressive Stream Reasoning. *Semantic Challenges in Sensor Networks, Dagstuhl Seminar Proceedings, 10042.*

Study on Global Megatrends in Cybersecurity. (2018). Ponemon Institute Report.

Turner, M., & Weinberg, D. T. (2011). *A Simple Ontology for the Analysis of Terrorist Attacks.* UNM Digital Repository. Retrieved from http://digitalrepository.unm.edu/ece_rpts/41

Tyugu, E. (2011). Artificial Intelligence in Cyber Defence. In *3rd International Conference on Cyber Conflict (ICCC)* (pp. 1-11). IEEE.

van Heerden, R., Leenen, L., & Irwin, B. (2013). Automated classification of computer network attacks. *International Conference on Adaptive Science and Technology (ICAST 2013)*, 157-163.

Vasudevan, V. (2018, July 24). *How AI is transforming Cyber Defense*. Retrieved from Forbes: https://www.forbes.com/sites/forbestechcouncil/2018/07/24/how-ai-is-transforming-cyber-defense/#127873be3bb2

Wang, B.-S., & Zhong, Q.-Z. (2017). Automatic Malware Classification and New Malware Detection using Machine Learning. *Frontiers of Information Technology and Electronic Engineering*, *18*(9), 1336–1347. doi:10.1631/FITEE.1601325

Wolfson, R. (2018, July 3). *How a Leading Cyber Security Company Uses Blockchain Technology to Prevent data Tampering*. Retrieved from Forbes: https://www.forbes.com/sites/rachelwolfson/2018/07/03/how-a-leading-cyber-security-company-uses-blockchain-technology-to-prevent-data-tampering/#5230bb004529

Yavanoglu, O., & Aydos, M. (2017). A Review on Cyber Security Datasets for Machine Learning Algorithms. *IEEE International Conference on Big Data, Symposium on Data Analyctics for Advanced Manufacturing*. 10.1109/BigData.2017.8258167

ENDNOTES

1 This chapter is an extension of a previous paper by the authors (Leenen & Meyer, 2016).

2 http://www.ihtsdo.org/snomed-ct/