AFRL-RI-RS-TR-2014-189

DATA MINING IN CYBER OPERATIONS

JULY 2014

INTERIM TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2014-189 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /
MISTY BLOWERS
Work Unit Manager

/ S /
BRENT HOLMES
Chief, Cyber Operations Branch
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2014	INTERIM TECHNICAL REPORT	MAR 2012 – MAR 2014

4. TITLE AND SUBTITLE

DATA MINING IN CYBER OPERATIONS

5a. CONTRACT NUMBER
IN-HOUSE

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)

Misty Blowers, Stefan Fernandez, Brandon Froberg, and Jonathan Williams (AFRL/RI)

George Corbin and Kevin Nelson (BAE Systems)

5d. PROJECT NUMBER
ACRE

5e. TASK NUMBER
IH

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGB
525 Brooks Road
Rome NY 13441-4505

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGB
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2014-189

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2014-0954
Date Cleared: 7 Mar 2014

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The dynamic nature of the cyberspace environment presents opportunities for both attackers and defenders to conduct complex cyber operations in serial or parallel across multiple networks and systems. Defensive operators must be vigilant to identify new attack vectors, real-time attacks as they happen, and signs of attacks that have gotten through the security perimeter. This means that defenders must continuously sift through vast amounts of sensor data that could be made more efficient with advances in data mining techniques to accurately map the attack surface, collect and integrate data, synchronize time, select features, develop models, extract knowledge and produce useful visualization. Effective techniques would enable models that describe dynamic behavior of complicated attacks and failures and allow defenders to detect and differentiate simultaneous sophisticated attacks on a target network.

15. SUBJECT TERMS
Cyber Operations, data mining, learning models

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**MISTY BLOWERS**
U	U	U	UU	16	19b. TELEPHONE NUMBER (*Include area code*) **315-330-3438**

Book: Network Science and Cybersecurity, Springer, July 2014
Chapter Title: Data Mining in Cyber Operations
Authors: Dr. Misty Blowers, Lt. Stefan Fernandez, Lt Brandon Froberg, Capt. Jonathan Williams, AFRL/ RI, Rome, NY
George Corbin and Kevin Nelson, BAE Systems, Rome, NY

Introduction

Cyber operations has been roughly defined as the employment of cyber capabilities to achieve military objectives or effects in or through cyberspace.[1] Defending cyberspace is a complex and largely scoped challenge which considers emerging threats to security in space, land, and sea.

Joint Publication 1-02, Department of Defense (DoD) Dictionary of Military and Associated Terms defines cyberspace as a global domain within the information environment consisting of the interdependent network of information technology infrastructures, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers.[1] Cyberspace operations is defined as the employment of cyber capabilities where the primary purpose is to achieve military objectives or effects in or through cyberspace. Such operations include computer network operations and activities to operate and defend the Global Information Grid. The global cyber infrastructure presents many challenges because of the complexity and massive amounts of information transferred across the global network daily. The cyber infrastructure is a made up of the data resources, network protocols, computing platforms, and computational services that bring people, information, and computational tools together.

Data Mining

According to Han and Kamber, [2] data mining is a process of discovering interesting patterns in large amounts of data which as previously noted is often a challenge in cyber operations. In order to gain a tactical edge, a warfighter must be able to apply data mining techniques to be maneuverable in cyber space. Maneuverability in cyberspace allows attackers and defenders to simultaneously conduct actions across multiple systems at multiple levels of warfare. For defenders, this can mean hardening multiple systems simultaneously when new threats are discovered, killing multiple access points during attacks, collecting and correlating data from multiple sensors in parallel or other defensive actions.[3] The complexity and dynamics of cyber operations is only weakly understood, especially when a nation is engaged in cyber-warfare.

The dynamic nature of the cyberspace environment presents opportunities for both attackers and defenders to conduct complex cyber operations in serial or parallel across multiple networks and systems. [4] Defensive operators must be vigilant to identify new attack vectors, real-time attacks as they happen, and signs of attacks that have gotten through the security perimeter. This means that defenders must continuously sift through vast amounts of sensor data that could be made more efficient with advances in data mining techniques to accurately map the attack surface, collect and integrate data, synchronize time, select features, develop models, extract knowledge and produce useful visualization. Effective techniques would enable models that describe dynamic behavior of complicated attacks and failures and allow defenders to detect and differentiate simultaneous sophisticated attacks on a target network. [4] Defensive operators that

manage an enterprise-level network, distributed networks or multiple, interoperating networks face a significant challenge of strategic coordination to defend against complex cyber-attacks. These operators clearly face a "big data" problem. [5]

"Big Data" is about the growing challenge in how we deal with the large and fast-growing sources of data or information. It presents a complex range of analysis and use problems. [6] There are many considerations when dealing with massive amounts of data. One challenge is in having a computing infrastructure that can ingest, validate, and analyze high volumes (size and/or rate) of data. Another is in assessing mixed data (structured and unstructured) from multiple sources. It is often very challenging to deal with unpredictable content with no apparent schema or structure, and often a challenge enabling real-time or near-real-time collection, analysis, and answers. [6]

Before one attempts to extract useful knowledge from data, it is important to understand the steps in the data mining process. Simply knowing many algorithms used for data analysis is not sufficient for successful data mining (DM). The figure below outlines the process of mining data that leads to knowledge discovery.

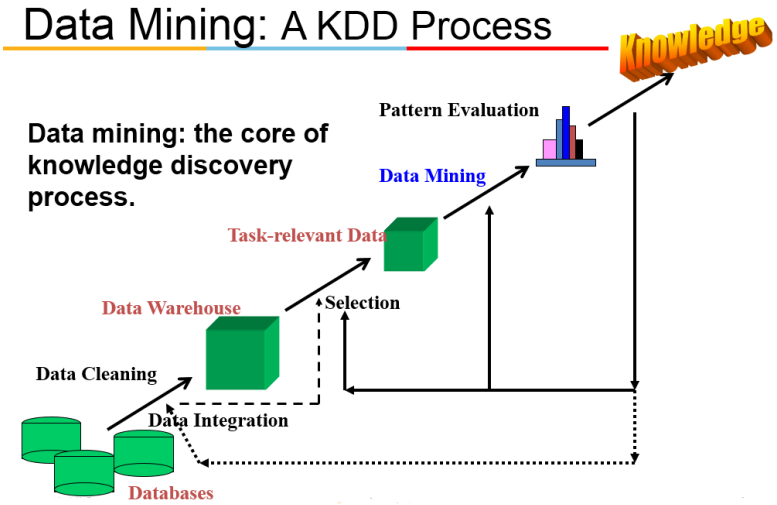

Figure 1: The Knowledge Discovery from Data process allows for the "mining" of valuable knowledge from vast amounts of data just as a miner mines for gold [2]

Fayyad et al. [38] describe the knowledge discovery from data model as a series of nine steps.

1. Develop and understand the application domain. This step includes learning the relevant prior knowledge and considers the goals of the end user.
2. Create a target data set. Here the data miner selects a subset of variables (attributes or features) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.

3. Data cleaning and preprocessing. This step consists of considering outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes. Outliers may be irrelevant or be significantly relevant depending on the task at hand.
4. Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.
5. Choosing the data mining task. Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.
6. Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
7. Data mining. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc. More advanced machine learning methods also may apply here.
8. Interpreting mined patterns. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.
9. Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge. In the cyber domain, metrics to measure the effectiveness of detection or battle damage assessment is considered.

The traditional approach to understanding and protecting the cyber domain is a highly manual and human intensive process. It is growing increasingly difficult for these manual processes to keep up with both the massive amount of data and the quickly changing landscape of the cyber domain. It has become necessary to utilize automated techniques to maintain situational awareness and effective offensive and defensive strategies in the cyber realm. Data mining within cyber operations provides some techniques to address these challenges. Through the data mining process described above, one can find hidden patterns, interesting data, or relevant correlations within large datasets. It provides techniques to automate the discovery of structure or patterns which would otherwise be out of reach from human analysts. This analysis is typically performed in an automated process with a variable amount of human interaction, depending on the application.

The scope of data mining for cyber operations is large enough to be its own book, so for purposes of this chapter the scope will be limited to intrusion and malware detection, social networking for cyber situational awareness, and emerging topics for data mining in cyber operations.

Data Mining for Intrusion Detection

Intrusion Detection and Prevention Systems (IDPS) are automated software designed to monitor traffic or mine through select data sources in search of evidence of an intruder attempting to compromise the network. An IDPS is created to monitor characteristics of a host, the network, and combination of both host/network. [9] IDPSs use three basic types of detection to discover intrusions: signature-based detection, anomaly-based detection, and stateful protocol analysis [10].

Signature-based IDPS use signatures, patterns known to indicate a threat, to compare to observable event patterns in order to identify a current threat [10]. A signature-based IDPS is used in firewalls as a first line of defense as it can efficiently identify threats and act before damage is done for very precisely defined and common threats. A disadvantage to this approach is that it relies entirely on a database of known attack signatures to compare against the current network activity. Data mining may be applied to a signature-based IDPS by observing and analyzing known and suspected attacks to discover new signatures and patterns indicative of an intrusion [11].

Applying data mining techniques allows not only for these previously undiscovered signatures to be found, but also for generalized patterns of attacks to be seen. New and novel attacks, which may not exactly match a previously observed signature, may still match the general patterns of an attack that were learned through data mining techniques.

Anomaly-based detection depends on understanding normal patterns of network activity and looking for activity which appears abnormal relative to normal activity [10].The vast majority of new threats will come in as anomalous traffic and yet will likely be undetectable by Signature-based Detectors until new signature rules can be created once they are detected, countered and accounted for in the signature database. An anomaly-based IDPS can be successful in detecting attacks which are novel or vary too far from a signature to be detectable by the signature-based IDPS. They are slow to train and heavily dependent upon having very good "normal" data to upon which to base the training. Data mining is very applicable to this approach, as anomaly detection relies entirely on defining a baseline of normalcy. Various data mining techniques may be effectively used to learn a meaningful definition of normalcy based on known benign network connections. [12]

Stateful Protocol Analysis also looks at behavior outside of known signature patterns to precisely how protocols are designed to be used and what the protocol creators expect to see when those protocols are used [10]. The key is not only in finding anomalous behavior, but also in finding an anomalous behavior beyond what is typical for a specific network activity. Part of understanding a stateful interaction between a user and a network resource is the series of communications between them and not just individual packets as signature-based and most anomaly-based detectors are usually looking at. Looking at the state of the transactions, the intent of the user is revealed. Monitoring state in a network is complex and requires a lot of processing power in high volume networks. As new normal uses for protocols are developed, these systems need to be modified to understand them to ensure that they are not producing false positives. Again, data mining proves useful for defining what constitutes normal use based on previous network activity.

Data Pre-processing

Feature selection is a fundamental part of the Data Mining Process. The main goal is to identify features that are important to the mining effort. The effort of feature selection is to reduce the dimensionality of the data to make processing the data more efficient. Within the study of data mining there is a phenomenon called "the curse of dimensionality" in which all the dataset

members appear isolated and unique from the others. According to Dartigue, Jang and Zeng, the areas to analyze for feature selection and extraction can be in [12]:

- Intrinsic features which exist in all network traffic such as protocol, port, destination server name, and requester IP address
- Time-based features which connect traffic from "same host" or "same service" which is valuable in identifying DoS and fast probing exploratory attacks
- Host-based traffic features include grouping connections based upon the same server destination to help to identify slower probing attacks
- Content-based features that are designed to consider long term asynchronous conversations between the target server/service and the attacker's software client. These can be characterized as being slow, methodical and thorough attacks over wide windows in time

Model Development

Various data mining techniques have been explored in existing research to create Intrusion Detection Systems. Tsai, Tsu et al. performed a survey of machine learning techniques for intrusion detection seen in research papers between 2000 and 2007 [13]. Much of the research utilized training data to create classifiers which map input data to an output (benign or an intrusion). New incoming network traffic would be put through this classifier to determine if it represents an intrusion or not. The classifiers were generally one of three types: single, hybrid, or ensemble. Single classifiers utilize one single machine learning algorithm to create a single model which is used to make classifications. The most common single classifiers used to create IDPSs in the research are as follows:

- K-nearest neighbor (KNN) [17][18]: instance based learning to classify a new vector based upon it's calculated nearest neighbor from the training set
- Support vector machines (SVM) [19]: a supervised model defining the decision boundary, gap between the most divergent training examples, based upon support vectors rather than the whole training set to classify new events
- Artificial neural networks (ANN) [20]: information processing units intended to mimic the network of neurons in the human brain for performing pattern recognition
- Self-organizing maps (SOM) [21]: an artificial neural network that uses unsupervised training to produce discretized representation of the training data in the form of a low-dimensional map
- Decision trees [18][22]: maps feature observations about an event to conclusions learned from the features of a training dataset in the form of a classification/regression tree
- Naïve Bayes network [23]: analyzing the features independently of each other along a normal distribution as established by the training dataset
- Genetic algorithms [24]: a meta-heuristic designed to mimic natural selection in finding the most effective classification of new events based upon the features trained from the training dataset
- Fuzzy logic [25]: based upon a real world concept that things are never just black and white, rather they are in the spectrum of grey between the two extremes. It treats the training data as more benign and compares new data to be processed as more or less benign in comparison to the training set.

Hybrid classifiers combine multiple machine learning techniques to improve performance. This approach represents a more customized implementation to suit specific intrusion detection objectives. Hybrid classifiers may include multiple levels of processing/filtering of the training data where later phases are fed subsets of results from earlier filtering [26].

Ensemble classifiers are another effort to improve on single classifiers. They apply a collection (ensemble) of learning algorithms to different training samples to collectively provide improved performance [27].

As data mining and machine learning tools become more popularly utilized methods for intrusion detection, they also become popular targets for adversaries to attempt to undermine. In computing, a denial-of-service (DoS) or distributed denial-of-service (DDoS) attack is an attempt to make a machine or network resource unavailable to its intended users. One common method of this attack involves saturating the target machine with external communications requests, so much so that it cannot respond to legitimate traffic or it responds so slowly it is rendered essentially unavailable. Such attacks usually lead to a server overload. For these types of attacks, the feature selection process becomes exceedingly more important. Computational resources can be optimized if critical features are detected and the noise is filtered away.

Barreno, Marco, et al provide an excellent taxonomy of other approaches adversaries may use against typical IDPS [7]. These taxonomies are shown in Figure 2.

	Integrity	*Availability*
Causative: *Targeted*	*The intrusion foretold*: mis-train a particular intrusion	*The rogue IDS*: mis-train IDS to block certain traffic
Indiscriminate	*The intrusion foretold*: mis-train any of several intrusions	*The rogue IDS*: mis-train IDS to broadly block traffic
Exploratory: *Targeted*	*The shifty intruder*: obfuscate a chosen intrusion	*The mistaken identity*: censor a particular host
Indiscriminate	*The shifty intruder*: obfuscate any intrusion	*The mistaken identity*: interfere with traffic generally

Figure 2 Taxonomy of attacks against IDPS [8]

According to the taxonomy, an attack is broken down into three different axes, influence, specificity, and security violation. The influence of an attack defines whether it is causative or

exploratory. A causative attack modifies the training set that patterns are mined from in order to influence the learning model. An exploratory attack does not alter the training process, but rather uses other techniques to take advantage of existing weaknesses or blind-spots in the model. An attack is further classified by its specificity as being either targeted or indiscriminate. A targeted attack focuses on a specific intrusion or creating a specific misclassification while an indiscriminate attack looks for any possible intrusion. The third axis, security violation, focuses on the CIA (confidentiality, integrity, availability) model of a network by describing an attack as either an integrity attack or availability attack. An integrity attack results in the IDPS incorrectly classifying an intrusion as benign (false negatives) while an availability attack causes so many misclassifications (both false negatives and false positives) that the IDPS becomes unusable.

Malicious Code Detection

Within the scope of intrusion detection is the more specific security concern of malware or malicious code detection. As the prevalence of malware infections has reached epidemic proportions, it is becoming increasingly important to choose the right defenses to prevent costly malware infections that are targeted at stealing sensitive corporate secrets and mining critical user information records. With today's Internet, malware researchers are seeing a large spike in malware activity and estimate that thousands of new malware variants are being released into the wild daily. Working with large datasets and feature sets to discover hidden patterns has proved extremely applicable to the area of malware detection. Malware can be defined as a program that performs malicious behavior, compromises the security of the system, or performs a function against the wishes of the user. The spread of malware represents an increasing threat to maintaining the security of cyber systems. According to the Symantec Global Internet Security Threat Report, there were over 5,000 reported vulnerabilities in 2012[28].

As mentioned in the previous section, traditional signature based detection is a standard approach for finding and detecting malicious behavior on a system. However, these methods are inherently less effective for detecting novel and polymorphic malware. Signature based detection cannot reliably detect new malware until after it has been identified and given a signature. Polymorphic malware attempts to continuously modify itself in order to evade detection from a previously assigned signature. These concepts pose a serious challenge to existing anti-virus solutions.

Automatic detection of malicious code is a common application of data mining techniques. One method for this detection is through the mining of auspicious binary executables. In order to perform this analysis, appropriate features must be selected to determine whether the sample is benign or malicious. These features may include a list of function calls, strings, headers, byte sequences, or other attributes of the binary [29]. These features can then be processed and fed into a classification algorithm. Some methods assign each sample a classification probability based on the Naive Bayes algorithm, a rules based classifier, or a multi-classifier system [29]. Oulette et al. proposed deep learning algorithms to classify related malware families using a more comprehensive understanding of the malware's intrinsic properties [30].Others have developed solutions which extract n-gram features from both binary and assembly code [31].

Anon-trivial challenge of these approaches is finding and extracting relevant and useful features for the data mining. Another challenge of these approaches is that it can only classify new malware samples based on previous known samples. Also, various obfuscation techniques attempt to hide the true intent of the malicious code to skirt detection. In order to overcome these challenges, some solutions look for relevant features in a dynamic environment. These systems may search for anomalies within network traffic or other previously unseen behavior patterns. Thuraisingham et al. developed models using support vector machines to detect intrusions or malicious behavior based on deviations from normal network patterns [31]. In order to detect novel classes, Masud et al. proposed techniques for the detection of concept-drifts in data streams, which may be applied to the domains of network intrusion or fault detection [32]. These approaches must continually refine their techniques to gain acceptable detection and false positive rates. Since these detection methods are typically utilized with the oversight of a human analyst, a high false positive rate will quickly cause frustration for both the analysts and end users.

Although few commercial IDPS products currently utilize data mining, this is a topic of growing importance with a large (and growing) corpus of research supporting its use. As the number and complexity of existing exploits increases and it becomes easier and easier to morph and obfuscate attacks, most common IDPSs which rely on an updated database of known attack signatures will become less effective. Data mining techniques for learning generalized patterns indicative of attacks will soon become more prevalent and effective.

Data Mining for Improved Cyber Situational Awareness

Handling cyber threats unavoidably needs to deal with uncertainty and imprecise information. What is observed as potential malicious activities can seldom give us 100% confidence on important questions about which machines have been compromised, the extent of damage that has been incurred, and who and why the systems have been targeted. It is through Social Network Analysis (SNA) that some of these questions may be answered. Again, this is a very complex problem which must take into consideration a wealth of information from multiple sources.

Efficient and reliable analysis of such large datasets is a challenge faced by both intelligence agencies and law enforcement. Data mining can yield results which would be impractical or impossible through manual efforts alone, due to the massive amount of relevant data available. These techniques are often performed semi-autonomously, delivering additional support for human analysts. Within the cyber security field, data mining processes may be applied in the defense of computer networks and cyber infrastructure to identify malicious actors or organizations that pose a threat. In addition, if some threatening entities have already been identified, then these techniques may be applied to expand the search in order to identify other related attackers.

Data mining provides the ability to correlate and condense data into a social network structure, in order to discover patterns and relationships between humans, organizations, or other entities. By representing a social network as a graph, with entities as nodes and relationships as edges, automated techniques can provide deeper insight into the social relationships present within that

system. SNA techniques help the human analyst discover interesting factors or patterns that have previously unrecognizable. SNA provides mathematical constructs to model and predict useful patterns of social interactions. This analysis can greatly bolster the efforts of human analysts by identifying areas of interest, spotting emerging leaders, and predicting behavior. Krebs utilized SNA to identify core members of a terrorist network involved in the 9/11 attacks [906]. In this example, the relationships and structure were built from surveillance data released by government authorities and publicly available information on the web. This analysis discovered strong mutual connections between the hijackers, while also revealing an emerging leader within the network structure [33].

In addition to discovering individual entities within a social network, analysis can reveal the strength and influence of a network as a whole. Shang et al. developed an indicator model that measures the degree of connectivity of a network in order to find and predict criminal networks [34]. Iqbal et al. demonstrated the feasibility of the collecting online chat logs, identifying topics of conversation, and analyzing these messages for possible criminal activity [35]. Chen et al. developed techniques to identify strong subgroups within a network, and to find central members within a subgroup of a potential criminal network [36]. These data mining processes can provide key information in developing a clear understanding of the social dynamics in play within the social network.

In addition to passively understanding the social connections, this analysis can also provide direction for actively influencing the social network. This intelligence may help determine a course of action produce a desired effect within the organization. For example, if the key members of an organization can be identified, then crucial lines of communication may be intercepted or denied to alter the effectiveness of the group. Other techniques may be applied to relevant areas of the graph to achieve a certain desired effect.

This social network analysis often relies heavily on the mining of large datasets to construct these networks. Public social media sites are a common source for this data. Lau et al. produced mining methods which discovered both implicit and explicit relationships derived solely from public social media sites, through extracted words supplied to a probabilistic model [37]. While this analysis can be extremely powerful, it depends strongly on the quality of the data collected. If the data is biased, misrepresented, or incorrect, the results will similarly be erroneous.

Emerging Challenges for Data Mining in Cyber Operations

Modern and emerging networks are rated by the amount of billions of bits they can transport in a second, which uses the metric prefix of giga- to represent a one billion multiplication factor. A common rating of network bandwidth is the term Gigabit, and this rate is abbreviated as Gbps or Gb/s. In a single minute there can up to 60 Gigabits transferred, which is equivalent to 7.5 Giga Bytes and is close to 1.5 DVDs worth of content. This number seems impressive at first, but quickly becomes shadowed when considering there are 1,440 minutes in a day, and the ratification of the IEEE standard 803.3ba defines both a 40 Gb/s and 100Gb/s network [39]. In a single day, at a maximum sustained bandwidth of 100 Gb/s, over 219,142 DVDs worth of content could be transferred. These Internet bandwidth speeds are slowly moving to replace commercial infrastructure as the status quo. Google Fiber advertises it is 100 times the speed of

broadband connections and is only at a bandwidth of 1 Gb/s [40]. However, there is a growing threat in cyberspace that will be able to block network traffic even at these high data rates

History was made in February 2014 when the largest ever Distributed Denial of Service (DDoS) attack was recorded by Cloudflare Incorporated [40]. Cloudflare is a content distribution network that hosts websites and applications for Internet users. The company recorded an attack of over 400 Gigabits per second against one of its hosted sites from a series of 4,529 vulnerable NTP servers [41]. Cloudflare also reports that Network Time Protocol (NTP) DDoS attacks see an amplification factor, of corrupted input to malicious-amplified output, of over 200 times, and they have observed the Simple Network Management Protocol (SNMP) protocol to have DDoS attacks with an amplification factor of 650 times [42]. Even with the worlds expanding Internet infrastructure of Fiber technologies these DDoS attacks will be able to saturate a company's Internet bandwidth, since they have currently shown an attack capability 4 times greater than the maximum 100 Gb/s bandwidth. Considering Google Fiber's speed claims: the NTP DDoS is 400,000 times greater in size than modern broadband cable bandwidth. Ultimately, data mining will be center stage in defense of growing DDoS and other unknown capabilities, since this focus is on massive amounts of data and bandwidth.

Analysts not leveraging data mining would become instantaneously saturated in extremely large data sets if they were to experience an NTP DDoS attack. Sifting through information that was being transmitted in a 1 Gb/s connection, or higher speeds, would need data mining to determine what activities and actions are occurring within this space. Data mining would allow the detection, determination, and prevention of cyber threats, which would enable IDPS to mitigate or even thwart such an attack. Any interesting scenario would be if an attacker was able to combine a DDoS with the execution of malicious code. The DDoS would then be used to obfuscate this malicious activity, and without data mining capabilities there could be a large delay time in discovering this code if it occurred at all. Future data mining will need to be able to be optimized in order to mitigate these near future cyber threat, and without leveraging data mining there seems to be no other solutions that would be able to allow maneuverability during an attack.

Maneuverability is key to cyber operations for both parties in a conflict. A DDoS attack is designed to effectively remove any movement of the target. Data mining could provide mitigation strategies that would allow a target partial survivability, which would allow transmission or even migration of operations to a non-attacked platform. Having extended periods of blocked transmissions in cyberspace could greatly cripple a system or asset with respect to denial, disruption, degrading, and even deception. A Key feature of future defenses must be able to survive and mitigate an attack to prevent a full stop of cyber maneuverability.

Furthermore new and different areas should start considering the application of data mining to potential big data problems. According to Kamal and Muccio (2011) mission awareness is at the heart of cyber situational awareness, which gives an understanding of mission to asset dependencies. In light of these new threats, and having this goal of situational awareness, new systems must incorporate data mining to stay relevant when analyzing big data. Having the ability to understand and interpret data through data mining will enable the ability to predict and provide potential courses of actions to defense systems. Lastly, there are applications for data

mining in current and future cyber modeling, simulation, and war gaming. From participating in these events it has been observed that many Department of Defense war games rely heavily upon analyst input and interpretation of data. The addition of data mining in war games could provide a deeper analysis of the results, or even add the potential of multiple iterations of scenarios where currently there are only a few iterations. The application of data mining to cyberspace is endless, but it provides a greatly exciting future to all of those involved.

Bibliography

1) Jabbour, Kamal, and Sarah Muccio. "The Science of Mission Assurance." Journal of Strategic Security 4.2 (2011).
2) Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.
3) Applegate, Scott D. "The principle of maneuver in cyber operations." Cyber Conflict (CYCON), 2012 4th International Conference on. IEEE, 2012.
4) Gregorio-de Souza, Ian, et al. "Detection of complex cyber attacks." Defense and Security Symposium. International Society for Optics and Photonics, 2006.
5) Grant, Tim, Ivan Burke, and Renier van Heerden. "Comparing Models of Offensive Cyber Operations." Proceedings of the 7th International Conference on Information Warfare and Security: Iciw 2012. Academic Conferences Limited, 2012.
6) Villars, Richard L., Carl W. Olofson, and Matthew Eastwood. "Big data: What it is and why you should care." White Paper, IDC (2011).
7) Barreno, Marco, et al. "Can machine learning be secure?." Proceedings of the 2006 ACM Symposium on Information, computer and communications security. ACM, 2006.
8) Barreno, Marco, et al. "The security of machine learning." Machine Learning 81.2 (2010): 121-148.
9) Sabahi, F., and A. Movaghar. "Intrusion detection: A survey." Systems and Networks Communications, 2008. ICSNC'08. 3rd International Conference on. IEEE, 2008.
10) Scarfone, Karen, and Peter Mell. "Guide to intrusion detection and prevention systems (idps)." NIST Special Publication 800.2007 (2007): 94.
11) Han, Hong, Xin-Liang Lu, and Li-Yong Ren. "Using data mining to discover signatures in network-based intrusion detection." Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on. Vol. 1. IEEE, 2002.
12) Lee, W., & Stolfo, S. J. (1998). Data Mining Approaches for Intrusion Detection. Proceedings of the 7th USENIX Security Symposium. San Antonio.
13) Tsai, Chih-Fong, et al. "Intrusion detection by machine learning: A review." Expert Systems with Applications 36.10 (2009): 11994-12000.
14) Dartigue, Christine, Hyun Ik Jang, and Wenjun Zeng. "A new data-mining based approach for network intrusion detection." Communication Networks and Services Research Conference, 2009. CNSR'09. Seventh Annual. IEEE, 2009.
15) Michalski, Ryszard S., Ivan Bratko, and Avan Bratko. Machine Learning and Data Mining; Methods and Applications. John Wiley & Sons, Inc., 1998.
16) Theodoridis, S., & Koutroumbas, K. (2006). Pattern recognition. Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo: Academic Press

17) Bishop, C. M. (1995). Neural networks for pattern recognition. England: Oxford University.
18) Mitchell, T. (1997). Machine learning. New york: McGraw Hill.
19) Vapnik, V. (1998). Statistical learning theory. New York: John Wiley.
20) Haykin, S. (1999). Neural networks: A comprehensive foundation (2nd ed.). New Jersey: Prentice Hall.
21) Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59–69.
22) Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, P. J. (1984). Classification and regressing trees. California: Wadsworth International Group.
23) Pearl, Judea. (1988). Probabilistic reasoning in intelligent systems. Morgan Kaufmann
24) Koza, J. R. (1992). Genetic programming: On the programming of computers by means of natural selection. Massachusetts: MIT.
25) Zimmermann, H. (2001). Fuzzy set theory and its applications. Kluwer Academic Publishers.
26) Jang, J.-S., Sun, C.-T., & Mizutani, E. (1996). Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. New Jersey: Prentice Hall
27) Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226–239.
28) Symantec. 2013 Internet Security Threat Report. Volume 18 Vol. , 2013. Print.
29) Schultz, M. G., et al. "Data Mining Methods for Detection of New Malicious Executables". Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on. Web.
30) Ouellette, J., A. Pfeffer, and A. Lakhotia. "Countering Malware Evolution using Cloud-Based Learning". Malicious and Unwanted Software: "The Americas" (MALWARE), 2013 8th International Conference on. Web.
31) Thuraisingham, B. "Data Mining for Malicious Code Detection and Security Applications". Intelligence and Security Informatics Conference (EISIC), 2011 European. Web.
32) Masud, M. M., et al. "Classification and Novel Class Detection in Concept-Drifting Data Streams Under Time Constraints." Knowledge and Data Engineering, IEEE Transactions on 23.6 (2011): 859-74. Web.
33) Krebs, Valdis E. "Mapping networks of terrorist cells." Connections 24.3 (2002): 43-52.
34) Xufeng Shang, and Yubo Yuan. "Social Network Analysis in Multiple Social Networks Data for Criminal Group Discovery". Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2012 International Conference on. Web.
35) Iqbal, F., B. C. M. Fung, and M. Debbabi. "Mining Criminal Networks from Chat Log". Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on. Web.
36) Chen, Hsinchun, et al. "Crime data mining: an overview and case studies." Proceedings of the 2003 annual national conference on Digital government research. Digital Government Society of North America, 2003.
37) Lau, R. Y. K., Yunqing Xia, and Yunming Ye. "A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media." Computational Intelligence Magazine, IEEE 9.1 (2014): 31-43. Web.

38) Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), 1996. Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge
39) McCabe, Karen. "IEEE-SA - IEEE Launches Next Generation of High-Rate Ethernet with New IEEE 802.3ba Standard." IEEE Standards Association. Institute of Electrical and Electronics Engineers Standards Association, 26 May 2010. Web. 21 Feb 2014. https://standards.ieee.org/news/2010/ratification8023ba.html.
40) Prince, Matthew. "Technical Details Behind a 400Gbps NTP Amplification DDoS Attack.". Cloudflare, Inc, 13 Feb 2014. Web. 21 Feb 2014. http://blog.cloudflare.com/technical-details-behind-a-400gbps-ntp-amplification-ddos-attack.
41) Graham-Cumming , John . "Understanding and mitigating NTP-based DDoS attacks." . Cloudflare, Inc, 9 Jan 2014. Web. 21 Feb 2014. http://blog.cloudflare.com/understanding-and-mitigating-ntp-based-ddos-attacks.
42) Google Fiber Inc. "Plans and Pricing.". 2014. Web. 21 Feb 2014. https://fiber.google.com/cities/kansascity/plans.