Big Data for Combating Cyber Attacks

Terry Janssen, PhD, SAIC
Chief Scientist & Cyber Strategist
Cyber Operations
Washington, D.C. USA
terry.l.janssen@saic.com

Nancy Grady, PhD, SAIC
Technical Fellow, Data Science
Emerging Technologies
Oak Ridge, TN USA
nancy.w.grady@saic.com

***Abstract*—This position paper explores a means of improving cybersecurity using Big Data technologies augmented by ontology for preventing or reducing losses from cyber attacks. Because of the priority of this threat to national security, it is necessary to attain results far superior to those found in modern-day security operations centers (SOCs). Focus is on the potential application of ontology engineering to this end. Issues and potential next steps are discussed.**

Keywords—big data; ontology; cybersecurity; modeling, search; discovery; analytics; variety; metadata

I. INTRODUCTION

The last few years have seen tremendous increases in the amount of data being generated and used to provide capabilities never before possible. "Big Data" refers to the new engineering paradigm that scales data systems horizontally to use a collection of distributed resources, rather than only the earlier vertical scaling that brought faster processors and more data storage into a single monolithic data platform. Big Data technologies have the potential to revolutionize our capabilities to handle the large datasets generated in any cyber data analytics. The challenge, however, is not just in handling the large volumes and high data generation rate, but in leveraging all available data sources to provide better and faster analytics for attack detection and response. In this paper, we will discuss Big Data analytics, metadata, and semantics for data integration, and applications to cybersecurity and cyber data management.

II. BIG DATA

Big Data has several defining characteristics, including ***volume***, ***variety*** (of data types and domains-of-origin), and the data flow characteristics of ***velocity*** (rate) and ***variability*** (change in rate) in which the data is generated and collected.

Traditional data systems collect data and curate it into information stored in a data warehouse, with a schema tuned for the specific analytics for which the data warehouse was built. *Velocity* refers to a characteristic that has been previously referred to as streaming data. The log data from cell phones, for example, flows rapidly into systems, and alerting and analytics are done on the fly before the curation and routing of data or aggregated information into persistent storage. In a Big Data architecture, this implies the addition of application servers to handle the load. *Variability* refers to changes in the data flow's velocity, which for cost-effectiveness leads to the automated spawning of additional processors in cloud systems to handle the load as it increases, and release the resources as the load diminishes. *Volume* is the dataset characteristic most identified with Big Data. The engineering revolution began due to the massive datasets from web and system logs. The implication has been the storage of the data in its raw format, onto distributed resources, with the curation and imposition of a schema only when the data is read.

Big Data Analytics. Much of the development of Big Data engineering is a result of the need to analyze massive web log data. Massive web logs were first filtered by page for aggregate page counts, to determine the popularity of pages. Then the pages were analyzed for sessions (spawning the now massive "cookie" industry to make this simpler). "Sessions" are the sequence of activities that describe a customer's interaction with the site at a "single-setting," with the analyst describing what time-window is considered a session. The next step in analytics capability came from the realization that these sessions could be abstracted into patterns rather than being treated as just the literal collection of pages. With this step, traversal patterns helped site designers see the efficiencies in their link structure. Furthermore, these usage patterns could in some cases be attached to a customer account record. With this step, the site could be tuned to benefit the most valuable customers, with separate paths being designed for the casual visitor to browse, leaving the easy efficient handling for loyal customers. This pattern-oriented analysis applies to the cyber domain, in analyzing logs from a server.

The last 15 years have seen the extension of a number of analytics techniques to leverage the horizontal Big Data scaling paradigm to address both log and linked-node data found in social sites. The cyber community can leverage web log and Social Network Analysis to use the massive amounts of data to determine session patterns and the appropriateness of activity between resources. The challenge is that cyber must also deal with a richer set of attributes for the resources and their expected/allowed interconnections, which adds in a variety of other contextual datasets into the analysis.

Variety. Traditional systems handled the variety of data through a laborious integration process to standardize terminology, normalize into relational tables, choose indexes, and store into a data warehouse that is tuned for the specific analytics that are needed. This is an inflexible process that does not easily accommodate new data sources, changes into underlying data feeds, or new analytical requirements.

For web log analysis, this extension to customer session analytics only required the assignment of a customer or visitor

ID to the session, allowing integration with a purchasing history. In the cyber analytics case, the integration point is not so simple. The integration of packet data, with server log data, with port-to-port connectivity data, with server type data, with network router settings, and so forth, provides a more complex use case, needing a more sophisticated way to integrate such a variety of data, some of which carries a number of additional attributes that are needed.

Recently, *variety* datasets have been addressed through mashups that dynamically integrated a couple of datasets from multiple domains to provide new business capabilities. Early mashups demonstrated this value, for example, in the integration of crime data with real estate listings; a valuable analysis that was not possible before the availability of open datasets. There is a limitation to such mashups because of the integration of a limited number of datasets, with the integration variables being manually selected. This type of manual integration is insufficient for analytics across different large volume datasets with complex inter-relationships.

Variety is the Big Data attribute that will enable more sophisticated cyber analytics. The requirement is for an automated mechanism to integrate multiple highly diverse datasets in an automated and scalable way. This is best achieved through a controlled metadata.

III. METADATA

The executive branch has been pushing an open data initiative to move the federal government into being a data steward. The goal in releasing the data is to better serve the public and promote economic growth through the reuse of this data. The difficulty in using this data arises from the lack of the metadata descriptions. Data reuse requires as much information as possible on the *provenance* of data; the full history of the methods used for collection, curation, and analysis. Proper metadata increases the chances that datasets are re-purposed correctly—leading to analytical conclusions that are less likely to be flawed.

Two mechanisms are used for dataset integration in a relational model. In the relational model, lookup tables are established to translate to a common vocabulary for views, and a one-to-one correspondence is used to create keys between tables. In a NoSQL environment, joins are not possible so table lookups and or keys cannot be used for data integration. The connection of data across datasets must reside in the query logic and must rely on information external to the datasets. This metadata logic must be used to select the relevant data for later integration and analysis, implying the need for both standard representation and additional attributes to achieve the automated data retrieval.

A second approach is used to speed the data integration process for manual mashups of diverse datasets. Often XML wrappers are used to encapsulate the data elements, with the nomenclature for each dataset provided in the wrapper, based on user interpretation of the data elements. This approach allows rapid integration of data through the wrappers (as opposed to a lengthy data warehouse integration), but it is not an approach that can be automated, nor can it be used for large volume datasets that cannot be copied due to their volume. Even in a mashup, wrapper terms used in the metadata are themselves subject to interpretation, making reuse of data elements difficult.

Without metadata referenced to well-understood standard terminology applicable across domains, the diverse datasets cannot be integrated automatically. In addition, the integrating elements must be applied outside the big data storage, implying that the integration logic must reside in the metadata layer.

IV. SEMANTIC TECHNOLOGY

Semantic technologies are crucial for the future handling of big datasets across multiple domains. While we have methods for unique concept identification arising through the Semantic Web, these technologies have not made inroads into traditional data management systems. Traditionally, the ETL process has been used to enforce standard terminology across datasets, with foreign keys to external tables for the related information. This is not a scalable solution, since the introduction of a new data source requires the careful construction of foreign keys to each other dataset in the database. This lack of extensibility to add in additional sources highlights the limitations of horizontal scalability in current approaches. In addition, there are limitations on the continued expansion in large data warehouses, highlighting their inability to continue to scale vertically.

Semantic technologies have not yet made inroads into Big Data systems. Big datasets that consist of *volume* tend to be monolithic with no integration across datasets. The data is typically stored in its raw state (as generated), and no joins were allowed in the initial Big Data engineering. Given this, most Big Data analytics approaches apply to single datasets.

For solutions addressing the integration of *variety* datasets, the ability to integrate the datasets with uniquely defining semantic technology is a fundamental requirement. Two overarching requirements need to be addressed to use ontology for the integration of Big Data: constructing the ontology and using the ontology to integrate big datasets.

Ontology scaling. The standard method for data access through an ontology is to ingest the data into an ontological database, where the data elements are encoded along with their extant relationships. This does not work in a Big Data scenario, since ontological databases do not have the horizontal scalability needed to handle data at high volume, velocity, or diversity. Further exacerbating the problem is that some of the data needing to be integrated are not owned by the analytical organization and cannot be ingested, but only accessed through query subsets.

Separate ontology for metadata. The implementation of an integrating ontology would consequently need to reside in the metadata for browsing and querying. While this metadata could be browsed manually, the real value comes if it can be actionable; such that selections over the metadata ontology would automatically construct queries to the Big Data repository. A number of ontologies relative to the cyber domain already exist, encompassing resources, attack event ontologies, and so forth. The key is to incorporate the appropriate elements and their relationships needed to describe the elements in the desired datasets. Our intent is not to recreate a cyber ontology from scratch, but to leverage those that exist to develop a first order ontology specific to the integration of the relevant cyber datasets. Focusing on first order logic will enable the ontology to be actionable to dynamic data integration.

In order to serve as the facilitator for data integration for automated integration, this first order ontology would need to contain elements such as: data element definitions, dataset location, data producing resource characteristics, and resource connectivity.

For analytics, additional mid-level ontologies would be needed to provide reasoning over the data, such as time and location. Domain-specific ontology elements would include, for example, resource attributes by resource type, translations such as Internet protocol (IP) to location, and derived attack pattern components.

The key to the use of a semantic representation for the metadata is separating the semantic metadata from the data storage. In order to leverage the scalability and speed of high-volume NoSQL solutions, the ontology will need to reside in its own scalable environment. Data exploration would require a mechanism to browse the metadata within the ontology, with a seamless transfer mechanism to flow down into the data.

Probabilistic Challenges. One significant challenge in the use of ontology for automated data analytics across datasets resides in the need for probabilistic reasoning. Typically in ontology representations, triplets are considered *"facts,"* implying full confidence in the data elements being described. In the real world, such a luxury is typically non-existent. Resources will continually be updated, and there will be latency before the new configurations are updated in the ontology. Attack chains will have multiple possible paths with probabilistic representations of each link type. Activity counts must be evaluated with a statistical significance test to determine if an activity is truly of concern. Such counts will have variations relative to time of day and day of week. Using an ontology for such probabilistic analytics will require the ability to analyze activity under some uncertainty. Much work has been done on probabilistic ontology, like MEBN, which inserts Bayes' theorem in ontology nodes [1].

V. APPLICATION TO CYBERSECURITY

Practical application to countering cyber attack is achievable in the near-term. The following questions can be answered with properly implemented Big Data technologies that span the variety of datasets: What data is available on malware X attacks globally? How many machines did an event land on? What ports were leveraged? What users were affected? What machines were compromised? What was leaked? Was sensitive information lost? Who did it? Was it an insider or outsider? More difficult questions for the future would be: What is the composite activity globally of this attacker that penetration tested (pentested) my perimeter? What are all the locations globally of <malware name> attacks? What should I expect from this attacker within the next hour? Next week? Next month? (Based-on the historical data on this attacker.) What unsafe actions are my users doing, rank ordered by risk significance? What suspicious activity occurred today? Where is the greatest risk within the enterprise? It would also be useful to tabulate statistics on vulnerabilities versus attacks, and visualize the results.

The latter "future set" of questions requires more research and development in topics like machine learning and reasoning, and is well beyond this paper's scope. For example, can ontology as proposed in this paper help us reason about risk based on the topology of devices and controls? Theoretically, this is deterministic and machines should be able to do better than man. Our intent is to model perimeter security of a large, enterprise network and collect real-time data, reason about risk in real-time based on the topology of devices and controls, and respond to threats in attempt to prevent loss. Given the appropriate set of data and generation of a set of reasonable hypotheses, can we use Big Data to do evidence collection to support or refute those security risk and threat hypotheses, in time to prevent loss?

Progress-to-Date. As a first step in preparing to instantiate an ontology, we have been mindful of what hundreds of organizations do in the current cybersecurity management process in a global networked enterprise. Description of this workflow is beyond this paper's scope. System awareness currently resides in the minds of hundreds of professionals who track threats and malware, maintain the security devices like firewalls and the configurations and patches of thousands of network devices, monitor events and log files, create tickets when an anomaly is observed, and perform remedial actions such as Incident Response; Configuration Management; Vulnerability and Patch Management; Firewall, Intrusion Detection and Prevention; Deep Packet Inspection and Cyber Threat Assessment; Security Architecture and Design; and so forth.

We propose to elicit all knowledge necessary assessment, decision, planning, and response into this ontology. At first glance, this may appear daunting, but based on the successes with ontology engineering in recent years, and the high stakes,

we believe this not only practical, but necessary, to better understand how to solve this national priority problem.

Cyber-security management has the characteristics of a successful knowledge elicitation and ontology engineering endeavor. The information is in digital form, and cyber-security processes are repetitive—meaning that the same indications of an attack are well documented and observed in typical network operations routinely and the remedial steps are documented and used routinely. This is not to say the cybersecurity experts are not highly knowledgeable and skilled—just the opposite. This knowledge can be coded and reused in the parts the machine does best; man should continue to do the parts that it does better than machines. With this expectation, we will meet the goal stated up front of flipping the current situation to one where a network's defense is optimized and efficient, lowering cost of defense, and making it very hard and expensive for the attacker.

Cyber Ontology for Countering Attacks. The top levels are illustrated in Figures 1 and 2.

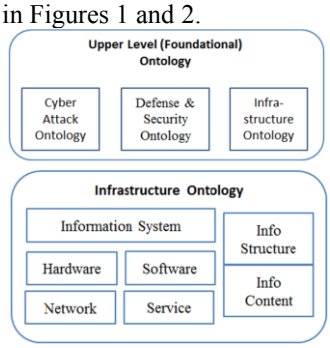

Figure 1. Upper Level and Lower Level Infrastructure Ontology.

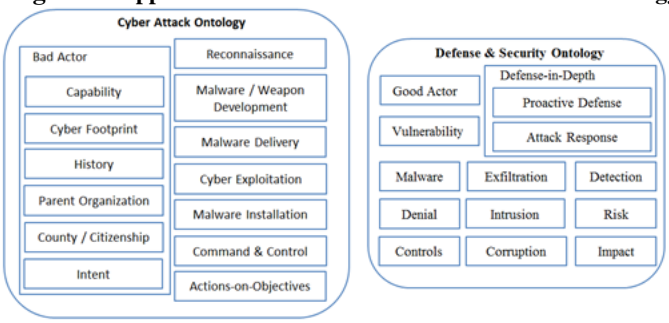

Figure 2. Lower Level Ontology for Attack and Defense.

Our goal is a proof-of-concept prototype of the entire process, but only for a few appropriate types of attacks and respective plans as defined by a fairly rigorous test set. Big Data elements for proof-of-concept have been partially selected.

Ontology engineering tools are being evaluated for "most suitable" for implementing this ontology for use in the system as previously described. A trade study will need to be conducted, for tools that can be selected for implantation of a production system capable of meeting the aforementioned objectives in a large, global enterprise network. For the purpose of demonstrating the concept we selected an ontology engineering tool from highfleet.com that reportedly provides an implementation of first order logic that is decidable and tractable (by simple programming constraint). It is a tool that one of the authors has used in the past. Results here are positive from the little done to-date; we cannot do an assessment until the ontology is populated. There are other ontology engineering tools, for example the description logic Protégé ontology editor. We have not made a decision; eventually we will need to identify appropriate metrics and conduct assessments to determine what would be needed for production grade deployment to address this problem space

Due to page limit constraints, it is impossible to discuss all aspects of the cyber ontology development, but a few aspects need to be mentioned. For example, there are many good resources for specifying and instantiations these ontologies to a level useful in cyber, most notable are efforts by MITRE [2]. Research issues remain unanswered and they can be categorized into big data and analytics, ontology and probabilistic reasoning, decision making and design and architecture. Cybersecurity is a hard problem and it is doubtful that the approach taken in this paper, or any other, will be a complete solution. Furthermore, the cyber attack sophistication is advancing rapidly which compounds this problem significantly [3].

VI. Future Steps

We are in the planning phase for continued research and development, beginning with the Big Data analytics necessary to more fully identify, understand, and respond to cyber attacks. In parallel, we would like to develop a proof-of-concept prototype to test how well this ontology and Big Data integration would work in practice in a large enterprised network with high traffic and large number of cyber attacks. The key to the success of this prototype will be to focus on one narrow aspect of cyber attack defense; if one is implemented and demonstrated, it can be used to extrapolate the resources needed for development and implementation in large production environments.

References

[1] Laskey, K.B, *MEBN: A Language for First-Order Bayesian Knowledge Bases*, Department of Systems Engineering and Operations Research, George Mason University, Fairfax, VA, 2007

[2] Obrst, L., Chase, P., Markeloff, R., Developing an Ontology of the Cyber-security Domain, *Semantic Technology for Intelligence, Defense and Security* (STIDS) 2012, GMU, Fairfax, VA, 2012.

[3] http://www.cnas.org/technology-and-national-security