Harnessing Machine Learning, Data Analytics, and Computer-Aided Testing for Cyber Security Applications

Achieving Sustained Cyber Resilience for Typical Attack Surface Configurations and Environments

Thomas J. Klemas
Decision Engineering Analysis Laboratory
Cambridge, USA
tklemas@alum.mit.edu

Steve Chan
Decision Engineering Analysis Laboratory
San Diego, CA
stevechan@alum.mit.edu

***Abstract*—While media reports frequently highlight the exciting aspects of the cyber security field, many cyber security tasks are quite tedious and repetitive. At the same time, however, strong pattern recognition, deductive reasoning, and inference skills are required, as well as a high degree of situational awareness. As a direct consequence, the field of cyber security is replete with potential opportunities to apply data analytics, machine learning, computer aided testing, and other advanced approaches to reduce the frustration of cyber security operators by easing key challenges. In fact, given a typical range of cyber attack surfaces, leveraging these machine-enhanced analysis and decision approaches in conjunction with a robust defense-in-depth posture is a crucial step towards achieving sustained, predictable performance across typical cyber security tasks and promotes cyber resilience. This paper will both outline details for a near-term research effort and explore a variety of key opportunities to exploit these approaches with the objective of raising awareness, providing initial guidance to aid potential adopters, and developing effective strategies to incorporate them into existing cyber security constructs.**

Keywords- artificial intelligence; expert systems, machine learning; supervised learning, unsupervised learning, pattern recognition, spectral methods, k-means, modularity, Lagrange multiplier, optimization, anomaly detection, data analytics, data science, networks, cyber security operator, cyber defensive tools, cyber resilience.

I. INTRODUCTION

It is well known that cyber security defense is a very challenging task [1]. One significant contributor to this defensive complexity is the inherently dynamic nature of the cyber environment. Client workstations, servers, other computers and devices, operating systems, and software become obsolete in a timescale of years and must be replaced with newer models or versions. Frequently, organizations will undergo transformations in size, focus, or organization that result from business mergers, growth, or decline and cause dramatic changes in the enterprise network composition. Because of any of these types of changes, organizations are constantly assembling unique new networks or modifying existing networks. Members may join, depart, shift to different sub-units, or change roles within the organization. In addition, network users themselves can be sources of great variability and can frequently frustrate cyber security operations by taking short-cuts for security measures or actively resisting inconvenient policy controls. Finally, external dynamics contribute to the defensive challenge: Criminals and other cyber attackers are perpetually scanning, searching, and finding new vulnerabilities to exploit, building new tools, developing new approaches to disguise their tracks, and refining their techniques to achieve their objectives.

Beyond the internal and external dynamic factors described above, paradigm-shifting technological changes are dramatically altering the cyber landscape, introducing innovations and improvements but potentially simultaneously increasing the attack surfaces and vulnerabilities within them [5]. The rapid evolution of new technologies, both hardware and software varieties, and increasing integration levels suggest that the challenges of cyber security will continue to grow for the foreseeable future [2]. For example, increasing Internet of Things (IoT) capabilities will incorporate new types of devices into the internet-accessible realm, thereby increasing attack surfaces and possibly exposing new types of vulnerabilities due to new Application Programming Interfaces (APIs) associated with new classes of devices. Thus, the addition of these technologies can increase the internal defensive complexity. Resourceful attackers may very well be able to find ways to exploit the increasing connectivity to access these new devices for their purposes.

While there are efforts to build in security into designs and even standards [3], frequently, security lags novel capabilities, sometimes to a significant degree. The dynamics, persistent advancement of malicious actors, and revolutionary technological change combine to elevate defensive complexities facing cyber security operators. As many defensive tasks are tedious and repetitive in nature, such fertile grounds will incubate the growth of errors. To counter this state of affairs, many cyber security innovators are leveraging data analytics, computer-aided testing, and machine learning to enhance or even replace human operator activities.

The remainder of this paper is organized as follows: Section II discusses the role of machine learning and other automated decision support tools in cyber security and presents applications. Subsequently, Section III explores how data analytics encompasses a crucial part for both supporting the function of these techniques and the decisions of cyber security operators. Then, Section IV delves into the benefits of leveraging computer-aided testing for the benefit of a wide variety of cyber security activities. Finally, the paper reviews and emphasizes key points in Section V, the conclusion.

II. Role of Machine Learning

Numerous cyber security tasks can be eased and accelerated by incorporation of pattern recognition and machine learning approaches. In particular, many modern cyber monitoring tools already incorporate machine learning and enhanced visualization to provide insights that guide and accelerate decision-making [1] [2]. Pattern recognition is important capability for cyber security monitoring support tools. For example, signature-based detection of malicious activity often involves scrutinizing specific attributes of packets, email, or other data for values of identifying features that may match the corresponding values of previously known feature set entry in a pre-loaded data set captured from previous attacks of malicious actors. Thus, signature-based methods are quite effective against previously seen attacks but typically ineffective against first-time attacks for which the feature set entries would not yet be included in threat profiles. However, these technologies play a vital role in a layered, defense-in-depth system [11] [12].

To deal with first time events, other anomaly detection tools would normally be utilized that learn about the "normal," non-anomalous patterns of behavior. Once the defensive systems can recognize the patterns of activities that comprise normal behaviors, then the leap is not so great to be able to distinguish when a new event is an anomaly. As an example, time is the feature dimension that would be used to analyze user login behavior relative to employee work patterns. Thus, once login data is captured in logs, it would be straightforward to detect anomalous login times for users that work regular business hours. Anomaly detection can be challenging because there are many feature dimensions across which an event could be deemed anomalous, and there are many conditions and states of the enterprise, network, and associated devices required to properly characterize the normal activity patterns for each.

Another dimension for anomalous events might include login Internet Protocol (IP) addresses. In companies with brick and mortar work sites, login entries will be dominated by internal IP addresses, perhaps followed by laptops used at home, such that IP addresses associated with foreign origins would easily fall outside the normal login entry patterns enough that they would be detected as anomalies. Certain dimensions (or categories of log or traffic data) should be flagged high priority due to high risk associated with malicious activities across the dimension; for example, events that include downloads and uploads should be scrutinized particularly carefully.

The reason that machine learning, which can fall a bit short amidst many human dominated chores, can fit the bill within a layered defense in that each component of a defense-in-depth approach is only responsible to achieve subset of goals. It is the integrated combination of layers and components, each supplying their specific contributions to the sum, that comprises the ultimate level of defensive strength of the system. This includes machine learning decision support subsystems and machine-aided tools, as well as human operators. Hybrid tools that incorporate pattern recognition, signature-based, and machine learning, can dramatically enhance the performance of humans in many of the tedious defensive tasks.

TABLE I. Machine-aided Aproaches to enhance Cybersecurity Operator and system performance

Approach	Alternative uses and cybersecurity considerations
Machine-learning	Signature matching and anomalous event detection. Classification of log entries and packet traffic data. Acceleration of asset management tasks. Semi-intelligent adversary attack agents for automated-testing activities.
Data analytics	Pre-processing and ingestion of event data records. Meta-data tagging of event entries for rapid filter searches
Automated-testing	Randomized agents capturing insider and external threat actor behaviors, pen-testing aids, and fuzz-type testing tools for software and systems

Machine learning has a strong role to play in anomaly detection. Pattern recognition techniques utilized for cyber security applications may involve identification of patterns that exist within a data set and then classifying both old and new data items into those categories or classes of patterns. Characterizing classes of normal traffic and classifying new traffic into existing patterns are well within the capabilities of machine learning algorithms, such that recognizing anomalies that do not fit into any of the existing patterns is possible.

For this classification and anomalous data detection task, a combination of signature-based systems, expert systems, supervised, unsupervised, and semi-supervised learning approaches may be advantageous to address the various challenges posed by different components of the data [7]. One of the objectives of this research was to examine the efficacy of applying a combination of classification approaches to categorize packet traffic data and log data. Previous work with clustering [13] [15] demonstrated some success in community detection. For these methods, feature set selection and determination of the number of classes are key steps to develop a successful classification tool. If feature sets are well chosen, clustering methods may be able to learn the numbers of basic types and the feature-based characteristics of the basic types of data with minimal human assistance, and then collect traffic or log entry statistics based upon those groupings.

Then, once group statistics are accumulated, it becomes possible to consider detection of anomalous data points that do not fit into any of the previously learned groupings. In this manner, anomaly detection systems can be constructed that classify data or messages into normal categories if they fit, and items that fall outside of the regular categories may be deemed anomalous or suspicious, potentially triggering some sort of alert, perhaps even associating a concern level for the degree of anomaly. Thus, these sorts of machine learning tools can support identification of traffic types, detection of anomalies and alerting for monitoring.

III. Data Analytics

Data analytics are crucial to the success of most machine learning approaches as well as human-led cyber security defensive operations. Data analytics includes pre-processing to clean and prepare the data, capturing essential auxiliary data that maximize the usefulness of data, ingesting the data into an extensible infrastructure, and analysis to squeeze the most insight possible from the data.

One of the often-overlooked aspects is the post-data-collection labeling or metadata tagging, which may involve detailed parsing of the data element. Using a network packet as an example, some typical fields that might require parsing include the time field, the protocol type, the source and destination IP addresses (if any), packet length, status fields, among other fields. This labeling step is vitally important to maximize the usefulness and efficacy of subsequent analysis stages. For packet traffic data, this step may require some special guidance to provide network subnet structure to group packet traffic by subnet since the detailed subnet structure would not necessarily be clear from the traffic itself, either implicitly or explicitly. Creating and populating a database with the dataset and the metadata tags of interest to facilitate further analysis, ensures the correct fields and values are available for rapidly searching data sets for samples of interest and associating data points that match in specified metadata attributes.

Another key objective of data analytics is computation of critical statistics that aid in preliminary decision-making and assist in selection of optimal approaches. A wide variety of statistics might be computed for network traffic that could include the relative composition of network traffic by protocol, subnet distribution of traffic across the enterprise network, IP addresses statistics, activity timing statistics, and many more. Once these statistics are computed, they can be utilized as additional dimensions for machine learning activities or for human decision-making.

IV. Computer Aided testing

The value of computer- assisted testing cannot be overstated in cyber security. While poor testing approaches can lead to outcomes that are worse than no testing, many testing efforts do not require a tremendous amount of intelligence to be successful but require thoroughness and are necessarily quite tedious, stretching the boundaries of human patience. By its very nature, computer-aided testing is comprehensive, methodical, and yet, can also incorporate randomness, and so this one of the key reasons why computers can fill this testing niche remarkably well.

Computer-assisted testing that includes randomized parameters is crucial because engineers frequently make assumptions during the development process, and, as these assumptions accumulate, the aggregative effect can be very hard to track and lead to inconsistencies. Thus, randomized testing approaches will occasionally violate these assumptions, causing tests to fail, by selecting test vectors within the engineers' "blind spot".

An obvious application for computer-aided testing that is used in software engineering and also applicable to cyber security defense is fuzzing or fuzz testing of software applications [8]. This sort of testing can help find website errors, database issues, and other application bugs. Similarly, a hybrid of fuzz and Monte Carlo testing can be used to aid validation of tools [9]. Recently, we developed an algorithm for calculating network complexity of virtual cyber ranges [10], but one key remaining task was validation of the algorithm. Using guided random-parameter point testing and by comparing the network complexity scores to subject matter expert expectations, we were able to make rapid improvements because the random value-generated models frequently represented attribute value permutations that fell outside our design assumptions.

One critical application area that could benefit from computer assistance is penetration testing. Some areas require human leadership, but computer-aided capabilities can assist other areas, such as tools to scan the surrounding network to enumerate devices and discern network structure and services, as well as tools to help find and infiltrate user accounts with weak passwords.

Like penetration testing, but with more requirements for stealth, automated attack tools can be used in war-gaming to challenge defensive teams or test tools. By measuring and observing the characteristics of the normal usage patterns, these tools could automatically enforce limits to ensure that communication and control traffic remains below the standard thresholding to avoid triggering cyber defensive tools. These tools could learn by passively observing and/or actively scanning its environment for vulnerabilities, potential pathways, defensive activities of concern or interest, and other exploitable opportunities that could yield the desired access or information.

V. Approach Details

The goal of this research effort is to gain insight into the efficacy of utilizing elements from each of the sections above to enhance and simplify the process of potentially determining anomalous events, traffic composition, groupings of interest, structure, and other attributes from passive analysis of collected packet traffic data. It is our hope that these results would enable operators to gain an understanding of both the full scope of the possibilities and limitations of this approach to accelerate detection of anomalies, identification, asset management, and other important cybersecurity functions. This multi-layer decision support system will incorporate machine aided learning to derive insights from higher level data produced by a data analytics platform that includes a variety of pattern recognition capabilities and other automation support. The remainder of this section will describe the experiment design and the technical approach details underlying the machine learning decision support methods.

TABLE II. PROPOSED ALTERNATIVES FOR EXPERIMENT DESIGNS

Candidate Independent Variables	**Candidate Dependent Variables**	**Candidate Control Variables**
Number of clusters	Cluster sets	Network size
Packet traffic data set	Packet time clusters	Network structure
Source IP addresses	IP clusters	Feature weightings
Target IP addresses	Traffic composition statistics	Feature vector
Protocols	Host associations	Data set size
Packet lengths		
Packet times		
Log entry data sets		
Initial cluster centroids		

Multiple alternatives for the higher-level experiment design, required to achieve desired end goals to include traffic characterization, anomaly detection, identification/asset management, and related cybersecurity objectives are outlined in table 2. Clearly, once the exact details are specified, using this approach, a similar (subset) table would be created for each desired experiment to enable determination of the number of repetitions required for statistics that satisfy desired hypotheses acceptance/rejection thresholds and support determination of confidence levels.

Although there are some crucial pre-processing steps to clean up and label data appropriately, in the interest of focusing on the technical challenges, we will omit details here and directly skip ahead to posit that clustering may serve well as initial approach to achieve rudimentary classification of the preprocessed data. First, we will share the fundamentals of various clustering approaches. We will represent packet traffic data or log entries by a graph, H, consisting of vertices or nodes, X, that represent items of interest and edges, F, that represent the connections between the items of interest.

$$H = (X, F)$$

The edges that connect node pairs capture specific associations of interest between the items, discerned in the data. The graph could potentially be multi-partite because the packet traffic or log data might identify the source and target IP addresses. Devices are typically distributed throughout the various subnets of the network, so there could be an additional layer of mapping required between the IP addresses and subnet nodes. A grouping C_m is comprised of a cluster of nodes, orthogonal to every other grouping, because no vertex exists in more than one grouping.

$$X = \text{U}C_m,\ C_m \cap C_n = \{\}$$

Each item, x, can be assigned a feature vector, g_x. Figure 1 depicts an example of a multi-dimensional feature vector. Our objective is to use the feature vectors with a metric to facilitate grouping of vertices into k clusters, although, for some applications, the feature vector could be as basic a notion as connectivity. Each element of the adjacency matrix, **B**, represents a measure of the events that relate a pair of IP addresses, forming connections between the corresponding nodes in the graph formed by the interconnections (or perhaps distance in the feature space) between the devices in the network [14]. If the IP pairing vectors that arise from the columns of the adjacency matrix are compared with a proximity measure (for example: a similarity measure) then connectivity patterns can be compared between nodes with straightforward operations, such as inner products.

Also, the similarity matrix, **S**, formed by computing inner products of the adjacency matrix column vectors is another useful concept:

$$\boldsymbol{S} = \boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B}$$

Thus, the higher valued elements of the similarity matrix will reveal node pairs, represented by the adjacency matrix column vectors, that have common patterns of connectivity.

For binary classification decisions, graph partitioning approaches may be employed that leverage spectral methods. To achieve larger numbers of classes, k-means, modularity-informed spectral methods, or hybrid k-means approaches [4][13]-[15] can be employed to compute clustering. Lagrange multipliers may be used in conjunction with these approaches to capture constraints for cluster memberships as part of standard optimization procedures.

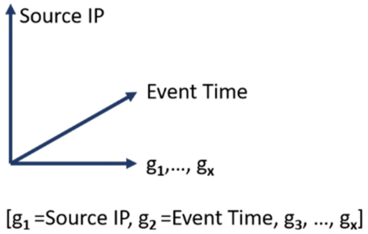

Figure 1. Example of a feature vector, g_x, which could potentially include other elements like target IP, protocol, and many more features.

We have obtained positive results with these methods [13] [15] to improve performance of the clustering algorithms for social community structure in cell phone data, so this study will explore its utility revealing insights that arise from potential groupings of traffic data or log entry items. As in our previous research, we may adopt the silhouette metric [13] [15] [16] to assess the degree of grouping structure in a proposed clustering, in which the silhouette value of one item or vertex, m, is specified

$$\text{silhouette}(m)=(p(m)-r(m))/\max\{r(m), p(m)\}$$

and r(m) represents an average dissimilarity between m and the remaining items or nodes within that cluster and p(m) is the minimum of the dissimilarities computed between m and all other clusters. Node or item dissimilarity is computed as the "distance" (e.g. Euclidean distance) between their respective feature vectors.

Accumulating the insights from multiple classification engines based on the outputs of the clustering processes, as well as outputs (e.g. alerts) of the other defensive tools, a multidimensional vector can be directed as an input to a multi-layered neural network that will form the basis for operator decision support. This research can explore the efficacy of alternative neural network methods [7], such as artificial neural networks, deep neural network, convolutional neural networks, and others to provide decision support in conjunction with the prior clustering/anomaly detection subsystem. One potential benefit of such an arrangement is that by working simultaneously, along-side the operators, essentially under continuous supervision, the neural net subsystems can improve performance with each detection decision, even as it offers suggestions to aid the operators in making their final determination. In large enterprise systems with multiple operators, this feedback loop may prove to accelerate performance improvement.

VI. Conclusion and Future Work

In this paper, we shared some cyber applications that can benefit from computer aided enhancements, such as machine learning, data analytics, and computer-aided testing. Numerous tools are entering the market, which incorporate these techniques, but it is challenging to leverage these novel tool capabilities effectively without a firm understanding of the underlying methods and the assumptions upon which they are based. Furthermore, many tools are ascribed far better performance in marketing literature than is achievable in a typical environment. As a result, there is rationale to develop internal tools and conduct thorough testing and tuning optimization for both internal and external tools of this kind. The testing and tuning iterations should measure success and accumulate statistics against common threat scenarios to ascertain overall performance.

We hope to employ the approach described in Section V to conduct a series of experiments to characterize the statistics associated with test networks as a baseline and then to study performance of enhanced systems that employ selected tools in conjunction with machine learning approaches outlined. The results should help shape new approaches to provide decision-aids and other support to cyber security operators that will help in providing insights and countering the rising challenges associated with enlarging attack surfaces that accompany the rapid evolving cyber environment and dynamics of typical enterprise networks.

Acknowledgment

The authors would like to thank the Design Engineering Analysis Laboratory (DEAL) and its affiliates for their encouragement throughout the process of completing this research. Without the assistance of the DEAL, as well as the ideas, guidance, and resources made available through that support, much of this research would have been delayed with unpredictable consequences. The authors would also like to thank the International Academy, Research, and Industry Association (IARIA) for the constant motivation to excel as well as the opportunity to serve as a contributing member within the cyber and data analytics domains.

References

[1] D. Schatz, R. Bashroush, and J. Wall. "Towards a More Representative Definition of Cyber Security". Journal of Digital Forensics, Security and Law, vol. 12, iss. 12, art. 8, 2018.

[2] https://www.telegraph.co.uk/connect/better-business/cyber-security/cyber-security-challenges-threats-in-2018/

[3] https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp18_CyberSecurity_Ed1_FINAL.pdf

[4] R. Lleti, M. Ortiz, L. Sarabia, and M. Sanchez, "Selecting variable for k-means Cluster Analysis by using a Genetic Algorithm that optimizes Silhouettes", Analytica Chimica Acta, vol. 515, iss. 1, pp. 87-100, 2004.

[5] T. Klemas, R. Lively, and N. Choucri, "Cyber Acquisition Policy Changes to Drive Innovation in Response to Accelerating Threats in Cyberspace", Proceedings CYCON 2018, press.

[6] https://www.trendmicro.com/vinfo/us/security/news/security-technology/is-big-data-big-enough-for-machine-learning-in-cybersecurity

[7] S. Theodoridis and K. Koutroumbas, "Pattern Recognition", Elsevier Inc, 2009.

[8] https://searchsecurity.techtarget.com/definition/fuzz-testing

[9] D.P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev,"Why the Monte Carlo method is so important today". WIREs Comput Stat., vol. 6, no. 6, pp. 386–392, 2014.

[10] T. Klemas and L. Rossey, "Network Complexity Models for Automated Cyber Range Security Capability Evaluations", ThinkMind, The First International Conference on Cyber-Technologies and Cyber-Systems, pp. 1-6, 2016.

[11] https://www.techrepublic.com/blog/it-security/understanding-layered-security-and-defense-in-depth/

[12] https://searchnetworking.techtarget.com/answer/What-is-layered-defense-approach-to-network-security

[13] T. Klemas and D. Rajchwald, "Evolutionary clustering analysis of multiple edge set networks used for modeling Ivory Coast mobile phone data and sensemaking", ThinkMind, The Third International Conference on Data Analytics, pp. 100-104, 2014

[14] M. Newman, Networks, An Introduction. Oxford : Oxford University Press, 2010.

[15] T. Klemas and S. Chan, "Automating Clustering Analysis of Ivory Coast Mobile Phone Data, Deriving Decision Support Models for Community Detection and Sensemaking", ThinkMind, The Fourth International Conference on Data Analytics, pp. 25-30, 2015.

[16] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Computational and Applied Mathematics , vol. 20, pp. 53-65, 1987.