SAND2017-12585R

Built on LDRD Laboratory Directed Research and Development

SAND20XX-XXXXR
LDRD PROJECT NUMBER: 201113
LDRD PROJECT TITLE: Temporal Cyber Attack Detection
PROJECT TEAM MEMBERS: Joe B. Ingram, Timothy J. Draelos, Meghan Galiardi, Justin E. Doak (JD)

ABSTRACT:

Rigorous characterization of the performance and generalization ability of cyber defense systems is extremely difficult, making it hard to gauge uncertainty, and thus, confidence. This difficulty largely stems from a lack of labeled attack data that fully explores the potential adversarial space. Currently, performance of cyber defense systems is typically evaluated in a qualitative manner by manually inspecting the results of the system on live data and adjusting as needed.

Additionally, machine learning has shown promise in deriving models that automatically learn indicators of compromise that are more robust than analyst-derived detectors. However, to generate these models, most algorithms require large amounts of labeled data (i.e., examples of attacks). Algorithms that do not require annotated data to derive models are similarly at a disadvantage, because labeled data is still necessary when evaluating performance.

In this work, we explore the use of temporal generative models to learn cyber attack graph representations and automatically generate data for experimentation and evaluation. Training and evaluating cyber systems and machine learning models requires significant, annotated data, which is typically collected and labeled by hand for one-off experiments. Automatically generating such data helps derive/evaluate detection models and ensures reproducibility of results.

Experimentally, we demonstrate the efficacy of generative sequence analysis techniques on learning the structure of attack graphs, based on a realistic example. These derived models can then be used to generate more data. Additionally, we provide a roadmap for future research efforts in this area.

INTRODUCTION:

In order to quantify the performance of machine learning algorithms on a cybersecurity problem or assess the efficacy of a cyber defense system, realistic attack data is necessary. Open-source data is limited due to sensitivity issues and concerns, and is generally limited in its representation. In addition, most cyber data is unlabeled and cyber attacks are very rare, so good labeled training sets are scarce. Therefore, data is typically collected and annotated by hand for experimentation, which is time-consuming and difficult to replicate. This lack of data for experimentation and evaluation makes it difficult to perform foundational research and advance state-of-the-art detection methods.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Built on LDRD Laboratory Directed Research and Development

Automatically generating cyber data is a difficult problem. Realistic cyber attacks contain a time component and potentially complex dependencies. So, how can realistic, meaningful labeled cyber data be generated?

A naïve approach would be to randomly alter the payload, or to insert random noise between stages of a known, observed attack. This approach is similar to fuzzing, a technique to test software by randomly generating inputs. Although a good baseline for comparison, more intelligent techniques to actively adapt the data may prove more realistic and successful.

In this work, we explore the use of generative sequence learning algorithms for automatically learning the structure and transition probabilities of an attack graph, based on observed attack state transitions. Generative models learn the joint distribution between the inputs and outputs of a system. By directly modeling the joint distribution, these models can then generate likely input examples for a desired output. These models can then be used to generate more attack state transition data as needed.

To measure the efficacy of this approach, we specify a ground-truth probabilistic model of an attack graph, generate attack states from this model, and train two different generative models on the resulting observed sequences. We explore Markov models and deep learning (long short-term memory; LSTMs). Measuring the difference between the estimated transition probability matrices of both models against that of the ground-truth provides a metric of performance for each algorithm.

Our results indicate that both methods do well at learning the transitions and probabilities of an attack graph. LSTMs might be of interest for learning and generating more complex attack graphs with long-term dependencies. We leave this investigation to future work. This work is a step toward automatically inducing generative models from observed cyber data, which can then be used for consistent, reproducible experimentation.

DETAILED DESCRIPTION OF EXPERIMENT/METHOD:

In order to reason and evaluate attack data generation methods, a rigorous, well-defined representation of an attack graph is necessary. Additionally, models of a given attack graph must be probabilistic in some sense in order to be able to generate interesting data that fully explores the adversarial space. Fortunately, finite–state machines and other more complex automata provide a useful formalism for defining attack graphs succinctly. Also, these representations have strong ties to generative modeling, which should make them effective for automatic data generation.

A finite-state automata (FSA) can be thought of as an abstract computer that contains a finite amount of memory. It consists of states of activity that describe a problem space (e.g., elements of a cyber attack) and conditions defining transitions to and from each state. Additionally, automata can be represented graphically by a series of states enclosed in circles and transitions

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

between these states represented as arrows between states as shown in Figure 1. This graphical representation helps aid understanding and reasoning. FSAs are a good start for representing attack graphs, but are only accepting devices and not generating devices. That is, they can only determine if an observed sequence is "accepted" by the FSA via a binary flag ("yes" or "no"). In order to formulate a generating attack graph, a probabilistic finite–state automata (PFA) is necessary, which defines state transitions probabilistically [1]. By randomly traversing the PFA based on its possible states and transition probabilities, it is able to generate more interesting, slightly random data.

The goal of this methodology is to use an existing cyber attack to generate similar (but different) attack data to create a richer set of training data for learning algorithms. The process can be described as follows:

1) Choose initial cyber attack.
2) Analyze the attack and determine states and transitions.
3) Define or estimate the transition probabilities.
4) Use the PFA to generate data by randomly traversing the graph, based on the defined probabilities.

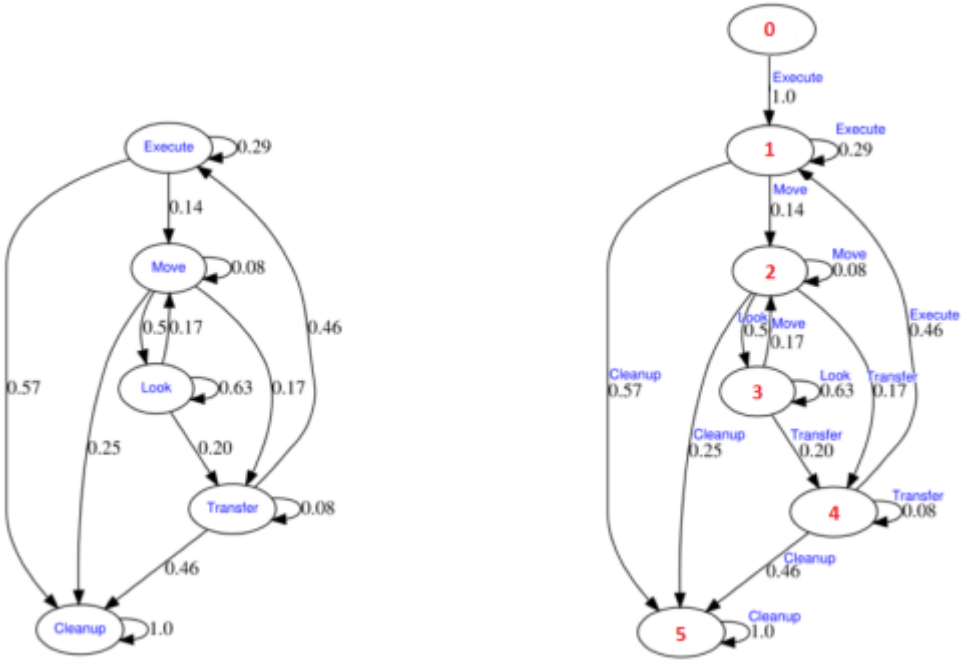

Figure 1: Markov chain of the attack graph (left) and the associated Probabilistic Finite Automata (right)

Suppose the Markov chain representation of the attack graph is defined by the left side of Figure 1, which is a realistic example borrowed from [2]. It can easily be converted into a PFA, by having the output along any transition simply be the current state, as shown on the right side of the figure. Then this model can be run probabilistically to generate attack data. For example, one generated attack sequence may be:

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Execute → Move → Look → Move → Transfer → Cleanup

RESULTS:

Experimentally, we demonstrated the efficacy of generative sequence analysis techniques on learning the structure of attack graphs, based on a realistic example. That is, given an observed sequence of attack state transitions, we investigated how many samples were necessary to learn the state transitions and probabilities of the attack graph using maximum likelihood estimation (MLE) of Markov models and deep learning (long short-term memory (LSTM)).

Markov models are a standard, well-known technique and baseline for sequence analysis. These models are based on the Markov assumption that the next state is only dependent on the current state. They are typically trained using maximum likelihood estimation, which estimates the transition probabilities from the observed transitions as:

$$\frac{w_{ij}}{\sum w_j}$$

where w_{ij} is the observed count associated with the transition from $i \rightarrow j$. The estimated transition probabilities of a Markov process converge to the actual transition probabilities at rate $1/\sqrt{n}$, where n is the sample size.

Long short-term memory (LSTM) [3] is a recurrent neural network that can be used for sequence prediction and classification, in addition to data generation. Its recurrent structure assumes that sequences are dependent, which allows this model to learn long-term dependences. This recurrence allows it to represent a more complex class of models than Markov models. They have shown significant success on various sequence learning problems, such as speech recognition [4].

Using the probabilistic finite-state automata (PFA) from Figure 1 as our ground truth attack graph, we measured the difference between the estimated transition probability matrices of the models with that of the PFA's probability matrix. The difference between the true transition probability matrix P and the estimated probability matrix P' is measured by the Frobenius norm between the two matrices, defined as: $||P - P'||$.

Figure 2 shows the norm between the models' estimated probability matrices and the ground-truth as a function of the number of training samples provided to the model. Both methods asymptotically approach the theoretical error rate with negligible difference, although MLE approaches faster. This result is not surprising given the consistency and efficiency of MLE. However, LSTMs are able to represent more complex models with long-term dependence, which may make them useful for generating more sophisticated attack data.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Additionally, both of these models can be used to generate more example data based on what has been "learned," which we also demonstrated. For example, some generated sequences from the LSTM are:

Execute → Move → Look → Look → Look → Move → Transfer → Cleanup

Execute → Move → Look → Transfer → Cleanup

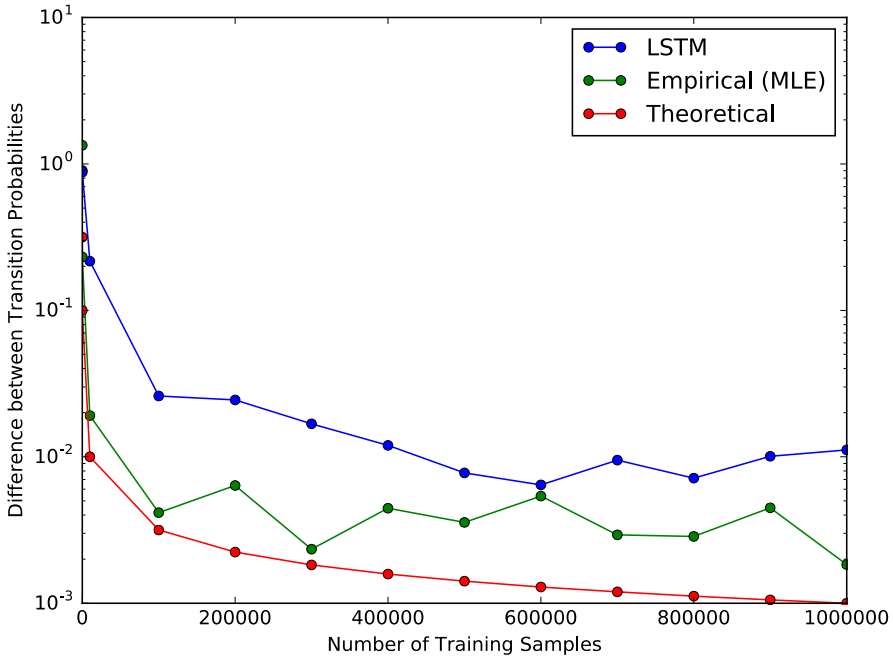

Figure 2: Difference between estimated transition probabilities and ground-truth as a function of the number of samples that the model is allowed to observe

DISCUSSION:

These results are a first step toward systematically generating full attack data. In future work, it would be useful to explore the automatic identification of attack states, as well as the internal dynamics of those states. That is, in addition to learning and generating state transitions, it might also be possible to learn and automate the generation of data that is representative of each state. Additionally, if the states can be detected automatically, an analyst would not need to research and derive each state by hand.

Learning generative models of full attack graphs would allow game-theoretic learning mechanisms to adapt and evolve attack data in order to preemptively generate data that avoids detection. Such proactive mechanisms could be used to enhance and augment machine learning models and cyber systems in order to make detection capabilities more robust.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Providing other primitives (e.g., packet crafting) for data generation would allow cybersecurity analysts to define their own attack use cases. These mechanisms would allow analysts to define attack scenarios based on their own expertise or open-source threat reporting. In order to demonstrate such a use case, we explored a real-world attack described in open-source threat reporting. We chose the malicious advertising attacks that were downloading ransomware on visitors to major news sites, such as New York Times and BBC [1] [2] [3]. After researching the attack from multiple sources, we derived the following states:

1) Delivery of Malicious Advertising (Malvertising) to Popular Websites
    1. The adversary does a little reconnaissance to determine which advertising networks to load their malvertising to.
    2. Adversary crafts malicious ad, perhaps specifically to evade detection by the ad networks.
    3. Adversary uploads ad onto ad network.
    4. Ad networks push ad to websites (e.g., BBC and New York Times).
2) User Visits Website Containing Malicious Ad
    1. User enters the URL of the site (e.g., nytimes.com) or gets referred to the site by clicking on a link.
    2. User is redirected to two malicious servers owned by the adversary.
3) Malicious Server Attempts to Compromise Browser
    1. The second malicious server uses Angler Exploit Kit (AEK) to look for exploitable vulnerabilities in the browser itself or in plugins.
    2. AEK is used to try to exploit identified vulnerabilities.
    3. If successful, the adversary downloads the BEDEP backdoor onto the user's system, which in turn drops malware known as TROJ_AVRECON.
4) Adversary Elevates Privileges and Executes Ransomware
    1. Adversary may need to exploit a different vulnerability in order to execute the malware.
    2. Adversary executes the malware.
    3. Dropper unpacks itself.
    4. Malware may need to elevate privileges by exploiting yet another vulnerability.
    5. Malware encrypts the user's hard drive.
    6. Malware requests ransom from user to unlock hard drive.
    7. Malware collects payment via bitcoin.
    8. Malware provides key for decrypting hard drive.
5) Malware Cleans Up
    1. Malware removes all artifacts (e.g., Windows Registry keys).
    2. Malware removes itself.

If there were provided tools for generating (or replaying) data from each state (e.g., network traffic containing the exploit), then it would be possible to generate data without first needing to observe it. Also, the graph and/or states could be tweaked (e.g., altering probabilities, adding noise, etc.) to perform sensitivity analysis on the resulting detection methods.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

ANTICIPATED OUTCOMES AND IMPACTS:

Generating consistent, annotated attack data will allow for more principled evaluation of cyber defense systems and statistical/machine learning models. The ability to quantitatively assess the detection capabilities of current cyber defense systems is crucial to trusting the nation's cyber assets and improving resiliency.

Since generative models can create data based on a model's learned representation, they can actively synthesize new data points, pass them through current analytics/models, observe the output, and adaptively "learn" what triggers detection /subversion. Successive iterations of this process would allow tracking false positives/negatives, and thus allow quantitative summarization of expected performance.

Additionally, such data generation mechanisms will be useful for machine learning and statistical models by: 1) providing data for training and deriving such models, 2) alleviating the need for data collection and associated sensitivity issues, and 3) ensuring the reproducibility of experimental results.

Emulytic environments provide an opportunity to instantiate virtualized environments to test vulnerabilities in systems and collect corresponding artifacts. However, knowledge of attacks is still required and creating diverse data is laborious. Also, automatically generating realistic data for simulation is an open problem in such environments.

This work has strong ties to the Linkography work at Sandia as the linkograph formulation and representation of attack graphs are directly related to Markov models [2]. This work directly leveraged the LinkShop library as well [6].

Additionally, given a generative model that can manufacture data, game-theoretic approaches can be used to train a model to detect the newly generated data, thereby making detection capabilities more robust. This method can be viewed as a more advanced way of correcting learning bias in highly-skewed samples, which is typically accomplished by oversampling, undersampling or randomly perturbing existing data to create new inputs. Game-theoretic techniques have had demonstrative success on images using generative adversarial networks (GANs).

To be successful in future research in this area, correctly representing the structure and constraints of cyber data may be crucial; future research will need to include investigating the necessary realism and validity of the generated data. Cybersecurity will continue to be challenged by attacks hiding amid large amounts of data, thus negatively impacting national security. The ability to quantitatively assess the detection capabilities of current cyber defense systems is crucial to trusting the nation's cyber assets and improving resiliency.

This work ties to the Cyberspace Mission Area and various future research goals in the Data Science Research Challenge. Attack graphs are activity-based models of adversarial intent.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Intelligently evolving such models allows quantifying uncertainty and identifying system subversions and vulnerabilities, which will allow for proactive remediation to increase resilience/confidence.

ACKNOWLEDGEMENT:

This work was funded by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories under Project Number 201113 and Title "Temporal Cyber Attack Detection."

The idea for this research stemmed from collaboration between the Algorithms and Cyber teams during the Hardware Acceleration of Adaptive Neural Algorithms (HAANA) Grand Challenge LDRD (PI: Conrad James, Algorithms PI: Brad Aimone).

The authors would like to thank Kevin Dixon for serving as Project Manager and providing valuable insights regarding both the programmatic and technical aspects of this project. The authors would also like to thank Michael Reed Smith, John Jarocki, and Rob Mitchell for useful suggestions and discussions.

CONCLUSION:

Although machine learning has made significant advances in applications such as speech recognition, handwriting recognition, image labeling, language translation, and sequence (e.g., character) prediction, its success has been less prominent in cybersecurity applications. One factor contributing to this lack of success is the need for realistic, consistent data for experimentation and evaluation.

In this work, we explored the use of generative sequence learning to learn attack graph representations from observed sequences in order to automatically induce data generation methods for cyber data. We performed an experiment using a few generative learning algorithms on a realistic example and demonstrated the efficacy and potential of this idea. We also provided some suggestions for future research directions.

BIBLIOGRAPHY

[1] E. Vidal, F. Thollard, C. D. L. Higuera, F. Casacuberta and R. Carrasco, "Probabilistic Finite-State Machines-Part 1," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] A. Fisher, C. Kent, D. Zage and J. Jarocki, "Using Linkography to Understand Cyberattacks," in *IEEE Conference on Communications and Network Security (CNS)*, 2015.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.,* vol. 9, no. 8, pp. 1735-1780, 1997.

[4] A. Graves, A. Mohamed and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *CoRR,* 2013.

[5] A. Hern, "Major sites including New York Times and BBC hit by 'ransomware' malvertising," The Guardian, 16 March 2016. [Online]. Available: https://www.theguardian.com/technology/2016/mar/16/major-sites-new-york-times-bbc-ransomware-malvertising. [Accessed 13 March 2017].

[6] J. Segura, "Large Angler Malvertising Campaign Hits Top Publishers," Malwarebytes, 4 November 2016. [Online]. Available: https://blog.malwarebytes.com/threat-analysis/2016/03/large-angler-malvertising-campaign-hits-top-publishers/. [Accessed 13 March 2017].

[7] C. Mihalick, "New York Times, BBC and others inadvertently serve up dangerous ads," CNET, 16 March 2016. [Online]. Available: https://www.cnet.com/news/new-york-times-bbc-dangerous-ads-ransomware-malvertising/. [Accessed 13 March 2017].

[8] "LinkShop," May 2017. [Online]. Available: https://github.com/sandialabs/linkshop. [Accessed 25 October 2017].

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.