Hybrid Modeling of Cyber Adversary Behavior

Amy Sliva, Sean Guarino, Peter Weyhrauch, Peter Galvin, Daniel Mitchell,
Joseph Campolongo, Jason Taylor

Charles River Analytics
Cambridge, MA, USA
asliva, sguarino, pweyhrauch, pgalvin, dmitchell, jcampolongo, jtaylor@cra.com

Abstract. Cyber adversaries continue to become more proficient and sophisticated, increasing the vulnerability of the network systems that pervade all aspects of our lives. While there are many approaches to modeling network behavior and identifying anomalous and potentially malicious traffic, most of these approaches detect attacks once they have already occurred, enabling reaction only after the damage has been done. In traditional security studies, mitigating attacks has been a focus of many research and planning efforts, leading to a rich field of adversarial modeling to represent and predict what an adversary might do. In this paper, we present an analogous approach to modeling cyber adversaries to gain a deeper understanding of the behavioral dynamics underlying cyber attacks and enable predictive analytics and proactive defensive planning. We present a hybrid modeling approach that combines aspects of cognitive modeling, decision-theory, and reactive planning to capture different facets of adversary decision making and behavior.

Keywords: Cyber Defense, Adversary Modeling, Cognitive Models, Decision Theory, Predictive Analytics, Cyber Simulation

1 Introduction

Over the last decade, the rapid increase in the number of networked devices, from desktop workstations to large-scale servers to ad hoc mobile devices, has vastly expanded the possible cyber attack surface. As modern cyber adversaries become more proficient and sophisticated, these network systems are increasingly vulnerable to cyber attacks. Despite significant investment in addressing cyber security, cyber attacks still remain a major threat to personal information, the global economy, and national security. In 2013, 7% of US organizations lost $1 million or more due to cybercrime, and 19% of entities had claimed losses between $50,000 and $1 million. Domestically, it is estimated that cyber-attacks cost $300 billion per year and cost $445 billion worldwide [1].

Many existing defensive mechanisms are based on analysis of network traffic or host-based observations, identifying anomalous behavior, looking for known patterns of alerts from monitoring appliances (e.g., intrusion detection systems), or finding malware application signatures. The biggest challenges currently facing cyber

defenders are that these defenses tend to be reactive and static, addressing attacks after the damage has already been done and failing to adapt to evolving strategies of advanced adversaries. These reactionary postures mean that defenders are often one step behind the adversaries. However, what if it were possible to get ahead of the adversaries and proactively deploy defenses that can deter or derail their attack strategies? To gain this advantage, we must first develop analytic methods that provide insight into the cyber adversaries themselves. Rather than viewing cyber attacks simply as a sequence of network traffic or a malicious application, we can look at the human component of an attack, recognizing that at the other end of this attack is a human adversary with specific goals and characteristics that influence their decisions. In this paper, we propose a novel hybrid modeling approach to understanding the cognitive biases, decision-making processes, motivations, and behaviors of cyber adversaries that can be used to simulate cyber attacks to predict and proactively address vulnerabilities.

2 Hybrid Cyber Adversary Models

To understand and analyze behaviors of cyber adversaries, we need a multi-faceted modeling approach that enables us to capture the multiple dimensions and characteristics of human behavior. We have developed a hybrid modeling methodology that combines decision theoretic models, cognitive models, and a reactive agent framework informed by sociocultural context to enable representation of a realistic decision process that manages multiple competing goals and the tradeoffs and biases associated with risks and rewards when planning an attack. This flexible hybrid modeling approach is based on the AgentWorks™ computational modeling framework [2][3], which provides a mechanism for merging disparate modeling formalisms into coherent, executable agents. Under this paradigm, different types of decision making and behaviors are captured by different mechanisms.

2.1 Decision Theoretic Models

When executing an attack, a cyber adversary inherently assumes some risk, such as the risk of getting caught or the risk of failure. Risk-based models derived from decision theory have been effective at modeling cyber adversary behavior [4], and capturing their decision-making process where they must balance the inherent risk of an attack against the potential reward if successful.

Decision theory posits that an agent will attempt to maximize their utility in any given situation, calculating the potential payoff given the likelihood of success. Using this insight, the authors developed a risk-based formalism for modeling cyber adversaries, defining a mathematical representation of the risk assumed by adversaries according to their characteristics [4]. In our approach, attacker risk is represented as a function of two attributes: (1) attack complexity; and (2) security features. The amount of risk an adversary assumes increases with the complexity of an attack and the security features. However, the rate of this increase in risk depends on two characteristics of an attacker—(1) skill level; and (2) access to resources—represented as parameters in the

risk-based models that determine the risk an attacker assumes for an attack. For example, a low-skilled attacker may find a complex attack too risky since there is a high chance of failure; however, a highly skilled adversary may find the same attack less risky and be more likely to choose this option.

We augmented these basic risk models with concepts from utility curves in microeconomics, assuming an increase in risk is not uniform, but has a marginal growth in the impact of each additional attack step or security feature. We quantify these factors in Equation (1), where A and S are the observable attack complexity and security features, $1 < \alpha, \beta$ are parameters representing adversary skill level and resources, and d and e are risk constants based on the particular network under study [4].

$$R(A,S) = \frac{d * \frac{1}{\alpha}^{A} \beta^{S}}{\frac{1}{\alpha_{min}}^{A_{max}} \beta_{max}^{S_{max}}} - e \quad (1)$$

The risk model above are based only on characteristics of skill level and access to resources, which may be inferable based on the types of attack patterns seen. However, we identified an additional characteristic—adversary motivation or will—that has a large impact on risk-reward. To incorporate will into the risk function, we considered will as an additive endogenous factor that shifts the risk curve, capturing changes in how the attacker perceives or weighs the risk of a particular situation (e.g., as will increases, the perception of risk decreases—or, said another way, an attacker becomes less risk averse and is willing to assume a greater amount of risk to achieve their goal).

2.2 Cognitive Decision-Making Models

The way in which an adversary makes judgments about the risk-reward tradeoff not only depends on external characteristics, such as skill, but is also influenced by their cognitive and cultural biases. For example, Chinese culture tends to be skeptical of secret information, assuming these sources to be potentially deceptive, making adversaries less likely to value certain types of information in cyber espionage operations. Similarly, most humans will be susceptible to classic biases, such as confirmation bias, which may impact their willingness to take certain risks.

These biases are indicative of the vagueness and imprecision inherent in human decision making. To model this aspect of cyber adversaries, we can use the fuzzy logic component of AgentWorks to assess the tradeoffs of different attacks based on bias. Our fuzzy logic component constructs a fuzzy set of possible states based on the attack options. Once these sets have been generated, they are input into a computational rule-base that uses mathematically-based Boolean functions (e.g., minimum for <and>; maximum for <or>) to combine members from the sets. The rules are weighted based on the importance of the rule to a particular adversary. This weighting also includes aspects of cultural and cognitive biases, enabling us to model the impact on an adversary's perception of their options. Combining these rules, the fuzzy logic component derives a value for each potential goal and action, which can then be used to weight the decision-theoretic risk functions described in the previous section.

2.3 Grammatical Representation of Cyber Attack Vectors

In addition to modeling the decision making of cyber adversaries, to understand how they might exploit vulnerabilities on a network we also need a representation of the actual actions that an adversary can take in a given circumstance. To represent adversary actions, we use a formalism adapted from the sociolinguistic theory of Systemic Functional Grammars (SFGs) [11]. SFGs can model the goals and actions of cyber adversaries, capturing the decomposition of an attack into the actions and the conditions necessary to successfully carry out each goal. SFGs are powerful due to their rich knowledge representation and reasoning capabilities. They are designed to account for contextual information—such as the state of the network, the goals of an attacker, and external factors, such as economic conditions—making them particularly well-suited to modeling complex attack structures for cyber adversaries.

As a sociolinguistic theory, SFGs are widely used in seminal natural language processing (NLP) systems, such as Winograd's language understanding system [12] and the Penman language generation system [13]. We adapted this approach to the cyber domain, representing the sequence of functional choices that can be made by cyber adversaries to achieve their goals. There is a large body of work on using structured approaches to represent cyber attacks. Attack graphs [14][15] can be used to describe vulnerabilities in a network and support detection of attacks along known vectors. However, these methods tend to be developed from a network rather than a behavioral perspective, requiring redesign each time there is a change. Conversely, SFGs are attack-centric. They focus on the goals and techniques of the attack itself, representing general attack patterns and their constraints. Therefore, SFGs do not need to be redesigned for different scenarios or networks.

The SFG structure has two layers, or strata: the grammatical stratum and the contextual stratum. In language, the grammatical stratum consists of an ontology of grammatical functions where each node may be associated with structural constraints. When applied to a cyber grammar, a node can represent *information gathering* and have the structural constraint that it must appear before a node that *exfiltrates data*. Just as there are millions of possible sentences in the English language, there are millions of possible cyber attacks that can occur in different scenarios. The contextual stratum consists of an ontology of intents and contexts where each node may map to one or more nodes in the grammatical ontology. This mapping from the contextual to the grammatical stratum allows us to generate elements of the grammatical stratum best suited to the current context.

2.4 Reactive Agent Framework for Realistic Goal Prioritization

Finally, our hybrid modeling approach draws from reactive planning [16] to represent the way an adversary might prioritize the goals and actions described in the SFGs in accordance with their biases and risk-reward decision making. Reactive planning models dynamic, adaptive decision making in reaction to beliefs about the state of the world. We have adapted Hap [17], a believable agent architecture developed to drive reactive, realistic agents in simulation environments.

The Hap framework is designed to support highly parallelized behavior and manage the exchange between potentially competing goals. Hap uses information from the probabilistic assessment of the belief state to factor in possible effects of its actions on the current state of the world and reprioritize its goals. Agent behavior is represented by decomposing each goal into subgoals and specific actions designed to accomplish that goal. Using the SFG representation, the Hap model will identify the grammatical stratum behavioral structure consistent with the current context, including the goals of the adversary.

At the top level of the behavior hierarchy are broad sets of activities, such as intelligence gathering and attack methods. At the lowest level, goals are specific actions that manipulate an environment to execute a particular activity. Goals can be either sequential, meaning one item must be completed before another can begin, or parallel, meaning the items can be executed at the same time. Each goal has prerequisites that must be completed before the goal or action can be executed

3 Discussion and Conclusions

The hybrid modeling approach presented here enables cyber defenders to create rich, realistic models of cyber adversaries. Using this approach, we have constructed models for several behavioral templates, representing a hostile nation state, a hacktivist group, and a "script kiddie" hacker. The modular nature of this methodology allows cyber defenders to design new adversary profiles by recomposing elements of existing models. Adversary models will provide cyber defenders with a deeper understanding of possible vulnerabilities and what types of defensive postures may be most successful by clearly illustrating which vulnerabilities are likely to be targeted and the responses of adversaries to a variety of defensive postures.

As we continue to refine and mature this modeling approach, we are cognizant of a number of significant challenges that lay ahead. For example, the profile above was developed solely through human research and expertise. We plan to explore various automated machine learning approaches that can augment part of this process, particularly in developing the mathematical decision-theoretic models of risk-reward tradeoffs. In addition, validation and verification of models of human behavior are always challenging tasks, and perhaps more challenging for cyber adversaries due to the anonymity of cyberspace. We plan to verify this modeling approach through high-fidelity simulations of cyber attack scenarios, as well as use these methods to make forecasts of future attacks that can be validated against open source attack reporting. Finally, we plan to refine the components of our hybrid modeling approach, enabling more complex representations of biases for risk-reward and situation-assessment, such as more advanced probabilistic models or soft-logic representations. Further, we are exploring integration with simulation engines, such as the OneSAF framework, for using these models in for predictive analytics to develop proactive defenses.

4 Acknowledgements

This material is based upon work supported by the Communications-Electronics, Research, Development and Engineering Center (CERDEC) under Contract No. W56KGU-15-C-0053 and the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0108. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of CERDEC, ODNI, IARPA, AFRL, or the US Government."

5 References

[1] I. Bremmer, “These 5 Facts Explain the Threat of Cyber Warfare,” TIME, June 19, 2015.

[2] B. Rosenberg, M. Furtak, S. Guarino, K. Harper, M. Metzger, S. Neal Reilly, J. Niehaus, and P. Weyhrauch, “Easing Behavior Authoring of Intelligent Entities for Training, ” In Conference on Behavior Representation in Modeling and Simulation (BRIMS), 2011.

[3] M. Furtak, “Introducing AgentWorks,” In 14th Intelligent Agents Sub-IPT, 2009.

[4] S. Li, R. Rickert, and A. Sliva, “Risk-Based Models of Attacker Behavior in Cybersecurity,” In the International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP), 2013.

[5] A. Pfeffer, “Probabilistic Relational Models for Situational Awareness,” In AIAA Infotech@Aerospace, 2010.

[6] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, “Learning Probabilistic Relational Models,” In Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), 1999.

[7] K. Murphy, “Dynamic Bayesian Networks: Representation, Inference, and Learning,” U.C. Berkeley, 2002.

[8] A. Pfeffer and T. Tai, “Asynchronous Dynamic Bayesian Networks,” In Uncertainty in Artificial Intelligence, 2005.

[9] S. Hongeng and R. Nevatia, R, “Large-Scale Event Detection Using Semi-Hidden Markov Models,” In International Conference on Computer Vision, 2, 1455-1462, 2003.

[10] P. A. Schrodt, “Forecasting Conflict in the Balkans using Hidden Markov Models,” In *Programming for Peace*, Springer, 2006.

[11] M. A. Halliday, *On Language and Linguistics (Volume 3)*, New York: Continuum, 2003.

[12] T. Winograd, “Understanding Natural Language,” Cognitive Psychology, 3, 1-191, 1972.

[13] W. Mann and C. Matthiessen, C. “Nigel: a Systemic Grammar for Text Generation,” USC/Information Sciences Institute, 1983.

[14] C. Phillips and L. P. Swiler, “A Graph-Based System for Network-Vulnerability Analysis,” In Proceedings of the 1998 Workshop on New Security Paradigms, 71-79, 1998.

[15] P. Ammann, D. Wijesekera and S. Kaushik, “Scalable, Graph-Based Network Vulnerability Analysis,” In Proceedings of the 9th ACM Conference on Computer and Communications Security, 217-224, 2002.

[16] J. R. Firby, “Adaptive Execution in Complex Dynamic Worlds,” Department of Computer Science: Yale University, 1989.

[17] A. B. Loyall, “Believable Agents: Building Interactive Personalities,” Pittsburgh, PA: Carnegie Mellon University, 1997.