This publication is cleared for public release

Cognitive Engagement in Active Cyber Defense

Robert Thomson, Ph.D.
Army Cyber Institute
United States Military Academy
robert.thomson@usma.edu

Cybersecurity is traditionally seen as an asymmetrical relationship between adversaries and defenders, however recent research has attempted to reverse this trend by operationalizing cognitive engagement for the purpose of enhancing adversary attribution. In a traditional networked environment, adversaries have low risk (low chance of getting caught) and potentially high reward (getting high-level access to privileged information), while defenders have a large attack surface to cover (all internal and external network access) and limited resources (computing and manpower) with which to defend their resources. Using deceptive techniques within networks allows defenders to better attribute attack behavior, which allows for increased data gathering and more targeted interventions. We discuss attribution techniques including the use of game theory and deception to maximize adversary interaction in a safer networked environment.

- Network defense does not need to be biased to favor adversaries' attack behavior
- Cyber deception can be used to modulate adversary engagement
- High engagement improves attribution by increasing adversary interaction in the network
- Adversaries cognitive states can be intuited through limited and indirect online interaction
- Future trends may support more interactive environments with adversaries

Background

With recent major hacks in government (e.g., OPM) and industry (e.g., Target, Home Depot), it seems like we're losing the war in cybersecurity. With a few pieces of software easily downloaded from the internet, a relatively notice hacker (referred-to as a *script-kiddie*) can severely damage the infrastructure of a large network if a vulnerability is undiscovered, or worse, *left unpatched*. Furthermore, passive defenses (e.g., firewall rules; anti-virus) can protect against known attacks, but may be vulnerable to previously unknown exploits in software (referred-to as *zero-day* exploits or *0days*). More sophisticated hackers, and groups of hackers (e.g., *Anonymous*) increase the challenges of Cyber Defense since they can penetrate sophisticated network defenses with little risk of being caught and are more likely to unleash 0days. In summary, there exists an asymmetry between the damage a single individual or small group can do and the resources required to protect our critical infrastructure against these attacks. What we need is active Cyber Defense, where we can engage adversaries in our own environments (i.e., have home-field advantage), attribute adversaries' intentions, and target our limited resources to their greatest effect.

Not all adversaries have the same goals or skills, so it is important to at least roughly-attribute adversaries to target optimal remediation. For instance, there are nation-state actors, non-state actors (e.g., groups such as Anonymous), and individuals of varying skills (e.g., the aforementioned script-kiddies). These adversaries may additionally come with differing goals. For instance, their attacks may target physical actions such as controlling a car, disabling an alarm, or shutting down a piece of technology. Alternately, these actions may be used for intelligence gathering by accessing protected files on a network. Finally, these actions may be used to gain influence by altering content and engaging in social engineering, such as hacking an individual's twitter feed to provide misinformation to their followers. Certain actions are more likely to be performed by certain adversaries; for instance, nation-

state actors may be more likely to perform interventions that lead to physical actions (e.g., in 2010 the Stuxnet worm attacked centrifuges inside Iranian nuclear facilities; or in 2015 with the Ukraine power-grid hack).

It is important to note that an attack is not necessarily a single penetration of a network in a single session. Adversaries operate according to an escalating *kill chain*, whereby they probe systems with the goal of compromise, then cycle between escalating permissions and reconnaissance until they are able to complete their attack. This attack may take time, in fact, the Ukraine power-grid hack started with a user clicking on a phishing e-mail, downloading malware onto a computer. That malware allowed hackers access to the system, then network, and over the course of six months these hackers reconnoitered the network exposing vulnerabilities until they were able to strike and simultaneously take-down numerous facilities leaving 225,000 customers without power (Sans Institute, 2016). Mandiant's (2015) annual M-Trends report claims that the average time between adversaries penetrating networks and their detection is still over 205 days (down from 243 in 2012).

A challenge in attribution is in detecting attacks as they happen. It is possible (and likely) for an adversary to wipe their trace from a compromised network. Despite knowing that an attack has occurred after the fact, there may be few or no logs with which to attribute the attacks or prevent further attacks. That said, there are mechanisms by which these adversaries can be discovered.

Honeypots and Deception for Cyber Defense

First-line deception techniques for Cyber Defense include *Honeypots*. A honeypot is a decoy that detects unauthorized use of an information system by hosting or emulating services (or entire operating systems) that appear to be real and of value to adversaries, but instead are monitored in order to *detect*, *surveil* and *distract/deflect* adversaries that attack the honeypot. Honeypots are categorized according to two general flavors: low-interaction and high-interaction.

Low-interaction honeypots (such as *Honeyd*; Provos, 2003) tend to only emulate a single service to detect attacks without exposing the entire operating system to risk. While they only offer limited interactivity for potential adversaries, the risks of the honeypot being compromised are extremely low. Using virtualization, it is possible to deploy and recover numerous honeypots using a single piece of hardware. A difficulty with low interaction honeypots is that they are relatively static, that is, because they only emulate a particular service, experienced adversaries may be aware of the honeypot's fingerprint and detect it. Still, they are a good first-line defense to detect and log simple attacks (e.g., automated attacks from script-kiddies) and to collect downloaded malware samples.

High-interaction honeypots use actual services, software, and entire operating systems to obtain a more detailed picture of adversaries' kill-chains. When using actual (as opposed to emulated) services, it is possible for high-interaction honeypots to discover new vulnerabilities (e.g., new 0days). This possibility is also a risk: by using actual services and software, it is possible for high-interaction honeypots to be themselves compromised (for which the operator may be legally liable should the compromised honeypot then attack another computer). While high-interaction honeypots may be virtualized or real machines, experienced adversaries may be able to detect virtualization. Despite this, an advantage of virtualization is that infected honeypots may be rapidly recovered.

Deceptive by nature, traditional honeypots do not detect the engagement of adversaries, but instead are tools for logging and post-hoc attributing of attacks. Using active deception techniques can shift the

apparent advantage that adversaries exhibit back to network defenders by exploiting their own desire to succeed against themselves.

Operationalizing Active Cyber Defense

The notion of *active* cyber defense means going beyond passive data collection to actively engage with adversaries. This does not mean to hack-back against adversaries (generally illegal in the United States), but instead use adversaries' own scripts/techniques against themselves and to engage adversaries' cognitive resources using techniques such as *game theory* (Vohs, Baumeister, Schmeichel, et al., 2014). By measuring and controlling cognitive engagement it is possible to maximize the effectiveness and opacity of deception techniques by modulating cognitive load such that attackers are more prone to ignore cues that may give away that they are in a honeypot environment, but also to accept the legitimacy of *deceptive credentials* and *fingerprinted documents* and make errors. In essence, the role of the active Cyber Deception is to *control adversaries' narrative*.

Passive techniques can help read adversary engagement and determine adversary goals, generally when an adversary is already engaged inside a Honeypot. These include psychometric analyses (e.g., keystrokes and temporal interaction), linguistic analyses (e.g., natural language processing, topic modeling, sentiment analysis), reverse-malware analyses (e.g., analyzing malware payloads), and attack pattern analyses (e.g., the use of which exploits and in what order). These passive techniques can help attribute attackers, for instance, by determining regional specificity within linguistic patterns in usernames and passwords. Attributing adversaries allows us to understand their motivations and behavior, making future defense easier (Rid & Ruchanan, 2014).

Active cyber defense techniques require an additional level of involvement, and come in two non-exclusive failures: (1) temptation techniques, and (2) engagement techniques. Temptation techniques attempt to entice particular adversaries into a deceptive environment by making them seem compromised and valuable. An example of a temptation technique is *reverse-phishing*, whereby a phishing e-mail containing a malicious website link or software is intentionally clicked-on within a deceptive environment with the goal of examining a particular adversary. Temptation techniques generally involve going outside a company's internal network to the Internet in order to learn more about particular adversaries/attacks. On the other hand, engagement techniques involve depleting adversaries' cognitive resources with the goal of increasing data collection, increasing the opacity of the deceptive environment, and inducing human error. To be effective, active engagement techniques involve online interaction with adversaries inside the deceptive environment.

An example of operationalizing active cyber defense is by using temporal keystroke patterns to estimate adversary engagement. A changing rate of keystrokes in a given temporal window may be used to predict stress, which also serves as a proxy for cognitive load (Vizer, Zhou, & Sears, 2009). Combining this with game-theoretic principles (Píbil et al., 2012; Carroll & Grosu, 2011), it is possible to increase adversary engagement within the system by determine which vulnerability they may access and in when they may access it. It is also possible to intermittently allow vulnerable credentials, further frustrating adversaries (Nicholson, 2015). This frustration can lead to real-world failures, such as failure to maintain/change connections through proxy servers (potentially exposing *real* IP addresses).

Wagener (2011) presents a real-life example of an active cyber defense architecture using a self-adaptive Honeypot in practice. This honeypot utilizes game-theoretic principles to produce a Nash-equilibrium (a state where neither attacker nor defender may gain by changing their current strategy). The honeypot has four potential behaviors: (1) allow execution of code, (2) block execution of code, (3)

substitute code, or (4) insult the attacker. This adaptive honeypot produced over three times the interactions against a baseline comparison honeypot (Heliza), providing substantially more information for attribution. Adversaries used more and more varied commands, spend longer online, and most interestingly, did not tend to disconnect after being insulted.

Most interesting is the option to actively insult an attacker, which generally violates the notion of being unobtrusive or opaque to adversaries. Even more interesting is that attackers only disconnected (via the *exit* command) 15.77% of the time when insulted. In one interesting anecdote, an attacker using a German IP address began to swear in Romanian after being insulting, implying that the attacker was Romanian and hijacking a German computer from which to launch an attack. This notion of interacting with hackers adds a new perspective to active defense. Minimally it may be possible to assist in determining bot-based attacks from human-based attacks via response to insult (or a lack thereof). Actively frustrating adversaries is a way of depleting cognitive resources, making it more likely to induce errors in judgment. Furthermore, as many attackers interacted with honeypot by replying to insults, this may provide additional linguistic clues beyond simply the language spoken.

In summary, active cyber defense via interacting with adversaries is a relatively new technique for measuring and modulating adversary engagement in the cyber domain. Active cyber defense provides more information for attribution, promotes human error in adversaries, and thus begins the process of reversing the asymmetry between attackers and the defensive resources required to counter them.

References

Carroll, T. E., & Grosu, D. (2011). A game theoretic investigation of deception in network security. *Security and Communication Networks, 4 (10),* 1162–1172.

Mandiant. (2015). M-Trends 2015: A View from the Front Lines. Technical Report. Retrieved July 24, 2016 https://www2.fireeye.com/WEB-2015-MNDT-RPT-M-Trends-2015_LP.html

Nicholson, A. (2015). Wide Spectrum Attribution: Using Deception for Attribution Intelligence in Cyber Attacks. *De Montfort University*. PhD Dissertation.

Píbil, R., Lisý, V., Kiekintveld, C., Bošanský, B., & Pěchouček, M. (2012). Game theoretic model of strategic honeypot selection in computer networks. In *Proceedings of the International Conference on Decision and Game Theory for Security*. Springer; Berlin Heidelberg, 201-220.

Provos, N. 2003. Honeyd-a virtual honeypot daemon. In 10th DFN-CERT Workshop, Hamburg, Germany, volume 2, 4.

Rid, T., & Buchanan, B. (2015). Attributing Cyber Attacks. *Journal of Strategic Studies, 38 (1-2),* 4-37.

Sans Institute. (2016). Analysis of the Cyber Attack on the Ukrainian Power Grid [White Paper]. E-ISAC. Retrieved July 24, 2016 http://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf

Valli, C., Rabadia, P., & Woodard, A. (2015). A Profile of Prolonged, Persistent SSH Attack on a Kippo Based Honeynet. *In Proceedings of the Conference on Digital Forensics, Security and Law*. Association of Digital Forensics, Security and Law.

Vizer, L. M., Zhou, L., & Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies, 67 (10),* 870-886.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2014). Making choices impairs subsequent self-control: a limited-resource account of decision making, self-regulation, and active initiative. *Motivation Science*, *1,* 19-42.

Wagener (2011). Self-Adaptive Honeypots Coercing and Assessing Attacker Behavior. *Institut National Polytechnqiue de Lorraine*. PhD Dissertation.