CHAPTER 5

Experimentation

Technology-specific solutions traditionally lend themselves to straightforward assessments of success via measurable results. The ability to determine whether or not a new technology provides a better metric as a solution to a problem is a foundational portion of any argument for its acceptance. The following analysis of established security paradigms and their respective evaluation via experimental methods will highlight the need for a differing process to provide defensible measurement of success or failure of human reliant cybersecurity implementation evaluations, which, given that attackers are humans, is all of them.

Unlike technologies, cybersecurity implementation assessment does not easily provide statistical metrics indicative of effectiveness. The art and tradecraft involved in such assessments mean that the same individuals could assess the same type of network and implementation multiple times and have different paths, discoveries, and recommendations. Additionally, the statistics that could be measured do not necessarily reflect the quality of work. If one type of assessment found 100 vulnerabilities and another type found 10, it might be deduced that the one which found 100 was the better assessment method. Part of what makes cybersecurity assessment methodologies difficult to compare is that it could be that the 10 vulnerabilities found in one assessment were of higher impact and importance than the 100 found in the other.

Not only is the cybersecurity implementation assessment process heavily reliant on human involvement from an attacker standpoint but the validation of its results requires implementations by yet another group of

J. G. Oakley et al., *Theoretical Cybersecurity*, https://doi.org/10.1007/978-1-4842-8300-4_5

humans performing systems administration, configuration, and operation. Then the organizational security must be reevaluated by a third group of humans to establish if there was change in the security posture. Here there is an issue where typical analysis of quantitative data is not only insufficient but likely unavailable in the way other security technologies might measure performance. Success of any given concept can be shown with defensible evaluation of the human tradecraft-driven assessment. To accomplish this, a framework for evaluating one cybersecurity implementation against another is necessary to allow for measuring their individual success and comparable novelty.

Identifying Requirements for Defensible Evaluation

Before designing an experiment to verify the novelty and quality of a cybersecurity implementation, experiment defensibility requirements need to be established. The following requirements toward defensibility should be met to standardize the actions of the human actors in the evaluation of cybersecurity paradigms:

- Controlled and realistic environment
- Defensible configuration
- Defensible operation
- Defensible Emulation
- Measurable results and metrics

Controlled and Realistic Environment

Since the goal of an experiment regarding a cybersecurity implementation is to identify how well it provides mitigation for threats and risks, it must be

conducted in an environment that represents exemplar real-world targets where such an implementation would be deployed. If assessments were done against unrealistic target networks, there would be no translation to success or failure of the paradigm in actual implementations. Control is important with regard to both users and administrators of a given network as well as outside actors attempting to compromise it. For example, if offensive cybersecurity assessors conducting one type of assessment, for instance, were able to leverage a communications path opened by the user running a Virtual Private Network (VPN), the assessment might have findings from a separate part of the organization. If assessors running another type of assessment against the same organization encountered no users running the VPN software during the time window for the assessment, they would never have a chance to generate the same findings and recommendations. This type of unfairness in an uncontrolled environment can be shown by any number of other examples such as outages in one location or another. For instance, a certain machine could be powered off during one assessment and during another, the machines might all be powered on. It is therefore clear that any evaluation of different offensive security assessments must be conducted in realistic, controlled, and identical environments.

Defensible Configuration

To determine the impact on the security posture of the test systems, configuration and administration must be performed in a repeatable and defensible way. This must also be carried out as realistically as possible. There could be a scenario where the administrator took over 100 hours to implement the changes for one cybersecurity implementation. If the implementation being compared took the administrator ten hours to complete, the comparison between the successes of either version of changes on the network might not be equal.

There is also a possibility that the configurations for one cybersecurity implementation to be evaluated are outside the realm of realistic expectations for systems administration in a real network. If the systems administration were performed improperly or unrealistically, it could provide no added security or potentially make a network more vulnerable, and therefore prevent comparison of the network's security posture. Any experiment aiming to determine the success of different cybersecurity implementations must ensure that systems administration and configuration is performed in an equal, appropriate, and realistic manner between compared paradigms.

Changes implemented by systems administration must also be accurate representations of the intent of the cybersecurity implementations. If the systems administrator misinterpreted what proper configurations were, it would also skew any ability to defensibly compare the success of one type of offensive security assessment over the other.

Defensible Operation

When comparing the effectiveness of two different cybersecurity implementations, the performance of those operating the implementations must be as defensible as possible. Imagine a scenario where one type of offensive cybersecurity assessment is conducted by someone with almost no experience in vulnerability assessment and computer exploitation and the other assessor has over ten years of such experience. The less experienced assessor is not likely to have as many or as impactful findings and is less likely to provide quality recommendations to mitigate those findings. That would be a poor basis to judge the quality of an assessment method against. Any experiment intent on evaluation of cybersecurity implementations must therefore ensure that the operators of that implementation are performed by equally qualified individuals if applicable. This is potentially not the case in a substantial portion of

cybersecurity implementations where human operation is not required post configuration. However, in the instances where human operators are involved, they need to be leveraged in a fair and defensible manner.

This is the case for both defensive and offensively oriented cybersecurity implementations. In offensive cybersecurity, the recommendations of the security assessors must be within the bounds of reason for an actual offensive security assessment. An assessor or defender could posit the recommendation of unplugging the organization network from the Internet or blocking all ports on device firewalls, which would certainly mitigate risk of remote exploitation. However, such recommendations are not likely to be applicable to any real-world scenario as they would hinder the operations of the host organization, and therefore would not be part of a real security solution.

Defensible Emulation of a Motivated and Sophisticated Attacker

With regard to evaluating the mitigating factors introduced to systems by cybersecurity implementations, the need for an appropriately emulated, motivated, and sophisticated actor is extremely important. Implementing security changes and then waiting to see if non-emulated attackers are able to compromise different portions of an organization is not defensible. It would be nearly impossible to guarantee a situation where a real cyber-attack was conducted with motivation against host organizations secured by the assessor recommendations. It would also be nearly impossible to determine the true motivation of real actors. The actor going after one network may be only a curious hacker or even an automated attack script and the attack against a second network could be an APT intent on some data or user within the network. Use of non-emulated actors creates an untenable situation for an experiment to present reliable or realistically defensible results.

Defensible emulation of the malicious actor allows the experiment to provide an equally motivated attack campaign against networks secured by cybersecurity implementations and then, as equally and defensibly as possible, determine the ability of those changes to thwart the attacker. There is a necessity to evaluate cybersecured networks to face equal levels of sophistication during the malicious attack campaigns waged against them. Equal motivation and sophistication of threats faced during experimentation is only available via emulated threat actors. This emulated actor should also represent a realistic threat commensurate with what real-world organizations may face. Regardless of actor motivation, if the capabilities for computer exploitation do not extend beyond the use of automated exploit frameworks, the experiment may result in a false sense of security where the network actually possesses little to no defense against real world threats.

Measurable Results and Metrics

If all other requirements for defensible experimental evaluation of cybersecurity implementations can be accomplished, there is still the need to provide a measurable metric. Such a metric must determine the level of success or failure that assessor-recommended changes had in enhancing the security posture and threat mitigation of an organization. Without such a metric, there is no way to determine a quantitative difference between offensive security concepts.

Without measuring the comparative effectiveness of offensive security assessments there is no way to validate a new paradigm as being an improvement upon existing methods in a given situation. As discussed earlier, such a metric must go beyond the number of findings by assessors. For the same reasons, success or failure cannot be measured by the amount of machines compromised by the emulated actor. If the emulated actor compromised ten unimportant user machines in one network, yet in the other compromised two servers, the email server and the file store

server, the two would seem to be more dangerous to the organization than the ten. To determine validity of a cybersecurity concept in comparison to others, measurable metrics representing realistic impact to the organization must be identified.

Evaluation Mediums

Potential underlying test beds for cybersecurity experimentation have four possible categorical mediums. The basic traits of these potential experiment mediums are based on the real or simulated nature of the environment and the real or simulated nature of the malicious actors. A real environment is considered for the purpose of this categorization to also have real systems administrators and operators (if necessary) and a simulated environment is considered to have its own simulated systems administration.

Real Network and Operators with Real Attackers

If this scenario were used for an evaluation medium, it would suffer from many drawbacks with regard to satisfying the defensibility requirements this dissertation has levied. With a real network and real attackers, the environment will be realistic and translate to real-world situations. However, there would be no experimental control over the organization or its network. Security assessment would not be defensible as too many environmental variables could differ across the different engagements. Using real systems administrators means that different administrators could perform different changes for the different actors and they may not want to comply with assessor recommendations if they do not agree with them. This would not allow for evaluation of the recommended changes. Relying on real attackers to engage the organization during experimental windows means there is no guarantee on similar attacks, as the sheer

breadth of variance in entities targeting organizations can be in the tens of thousands. It can be difficult to determine if a motivated attacker is trying to compromise the host organization during the evaluation period. Further, it would prove almost impossible to determine the level of sophistication of attackers between different evaluation windows, if attackers were present at all. Any metrics gathered during an experiment on such a medium would be unreliable at best and unsatisfactory as experimental results toward the validation of offensive security assessment methods.

Real Network and Operators with Simulated Attackers

If this scenario were used for an evaluation medium, it would also suffer from drawbacks with regard to satisfying the defensibility requirements this dissertation has levied against experimental validation. It is worth noting, however, that the supplement of simulated attackers for real ones does increase the potential for this option.

With a real network and simulated attackers, the environment will be realistic and translate to real-world situations. Like before, there would be no experimental control over the organization or its network. Security assessment would not be defensible as too many environmental variables still exist that may differ across the engagements of the different offensive security assessment methods being evaluated. Using real systems administrators still provides the possibility different administrators could perform different changes for the different assessors and they may not want to comply with assessor recommendations if they do not agree with them. Using simulated attackers allows for an equal level of motivation and sophistication with regard to attacks against the secured networks; however, the presence of real users and real security measures used by the organization still presents pitfalls for successful attack simulation and

evaluation. Any metrics gathered during an experiment on such a medium would still be unreliable as too many variables are left uncontrolled and potentially unequal between engagements.

Lab Network with Real Attackers

If this scenario were used for an evaluation medium, it would suffer from limited drawbacks with regard to satisfying the defensibility requirements in the attempt at validation of offensive security assessment paradigms. Use of real attackers on a controlled lab network does increase the defensibility of experimentation; however, it still has issues. A lab network in lieu of a real organization network, using real attackers, would in the immediate seem to present satisfaction for a controlled and realistic environment; this is not fully the case. Multiple real attackers could be acting against the organization at the same time and create the potential for hampering each other's progress as well as possibly creating situations that would allow for unnaturally expedited compromise of systems. There are also liability concerns in such experiments where attackers could leverage the lab network for exploitation of other targets. The lab network can be created in the image of a real organization and therefore translate to real-world situations. Yet, the inability to guarantee behavior of the actor means there is no ability to guarantee control of the lab network throughout the experiment. As long as security assessment of the lab network was conducted prior to being connected to the Internet to face real attackers, the assessment of the network will at least be defensible as environmental variables can be guaranteed to be equal during the assessment periods. As was the case previously with use of real attackers, motivation and sophistication cannot be guaranteed to be defensibly equal across the different engagements of the experiment. In such a setting, it can be difficult to distinguish between what was malicious activity or simply user mistakes. Since there is no guarantee on the effort of the attacker across given engagements, the metrics do not defensibly represent the effect of different assessor recommended changes on the security of networks.

Lab Network with Simulated Attacker

In a scenario conducted on this medium, an experiment is capable of achieving all of the defensibility requirements levied by this dissertation. Utilization of a lab network allows for a controlled environment. So long as it is created in the image of a real organization, it will be realistic, and findings of experiments conducted on it will translate to real-world scenarios. Security assessments conducted against controlled environments are defensible as the environmental variables can be maintained across assessment engagements. Systems administration conducted by experiment actors on the environment allows for defensible and equal representation of security change implementation. The motivation and sophistication of the simulated attacker can be guaranteed to be equal across the different campaigns and therefore defensible. Given the control over the realistic network and simulation of realistic actors during the experiment, this medium can provide measurable metrics that provide useable results for the validation of offensive security assessment paradigms.

Evaluation Mediums Summary

Clearly, there are pros and cons to picking a various-evaluation medium for the cybersecurity implementation evaluation to be conducted across. The most important thing is to understand the issues each of them face and to pick the most appropriate medium in a defensible manner. Doing this ensures that the evaluation medium has as little impact on the successful evaluation experiment as possible. Further, knowing the drawbacks and advantages of the chosen medium allows for experiment design to reflect further attempts at defensibility.

Experimentation Example

As an example, the following is a walkthrough of the experiment design and defensibility considerations I implemented for my doctoral dissertation, where I was evaluating the novel offensive cybersecurity assessment paradigm of *Counter-APT Red Teaming*. For more information on the concept itself, my *Professional Red Teaming* book, also by Apress, or my dissertation published by ProQuest contains exhaustive details. Here I am simply using it to illustrate what a best effort at defensibility in cybersecurity implementation evaluation looks like.

Experiment Design

With the goals of this experiment being to compare a new process for offensive cybersecurity assessment against more traditional red teaming, I determined that it requires a realistic lab network with cybersecurity implementations operated by real people, if necessary, and emulated threats and experiment actors. This is the medium I feel is best used to contrast two processes in a specific scenario.

With an evaluation medium determined for the experiment to be built upon, it is important to pick a target for the offensive security assessment that allows the experiment to provide results that would translate to a real scenario. For this purpose, there is a further requirement for identifying a simulated target that would provide an opportunity to represent the type of environment that would provide identifiable priority items for the CAPTR team model.

Target Determination to Support Realistic Network

The example of a law firm was chosen to be the basis for the lab network. A law firm contains data such as attorney–client privileged information as well as information being used in on-going legal cases. If compromised, such objects would likely be so damaging to the organization it would cease to operate. This example also allows for separate segments of a network containing operational personnel in one area and legal personnel in another. Unlike other probable targets of motivated advanced malicious actors, the legal firm example allows for a relatively small network of 40 to 50 machines to be used. This is in comparison to those of a large corporation or government institutions that would also likely be the target of such attacks. In a simulated law firm, there is no need to emulate specialized equipment such as medical or SCADA devices, which could prove difficult for experiment designers. The presence of such technology would also levy a need for specialized skills in security assessment, systems administration, and simulated attacker, which would make finding experiment actors a challenge.

Experiment Summary

CAPTR team methodology experimentation must defensibly answer two questions. Does CAPTR teaming identify findings that are unique to those found using offensive security assessors following more traditional processes? Do the recommendations from such assessments stand up in the face of advanced adversaries? Answering these questions allows for a measured representation of the uniqueness of findings generated via the CAPTR team paradigm and the ability of such findings to mitigate risk in the face of advanced motivated actors such as APTs.

With the goal of answering both questions, three identical copies of a network were created. The networks were built with only functionality in mind and were created to represent a small law firm of forty-two machines.

In this network, there were three functional LANs. There is a DMZ, a corporate LAN for devices supporting the operations of the organization such as a CEO and IT staff, as well as a LAN segmented off for the lawyers, legal aids, and customer information. Using the example of a law office allows for there to exist data and devices that, if compromised, could cripple or bring ruin to the organization. In this example, it would be confidential attorney–client privileged information from cases that would be treated as lethal compromises. The three different networks had different IP addresses, host names, user names, and domain names to appear unique to assessors and attackers, but the networks were set up identically.

One network was left unchanged as a control. The second network was assessed by an experienced penetration tester and former red team member from a machine in the DMZ using typical offensive security assessment tools and processes. This test was conducted with a scope of assessing the entire organization if possible. The third network was assessed in the CAPTR team methodology, the assessor was made to understand the intent of such an assessment and was given an initial scope of those items that would be lethal to the organization if compromised. This consisted of the case files and the servers they were stored on. These assessors then provided recommendations based on their findings. These recommendations allow for a comparison between what was identified and recommended from traditional security assessment and what was recommended by the CAPTR team resulting in a measure of uniqueness.

Lab Design

With the type of organization decided, the lab network needs to be structured such that it provides for control and realism. The types of technologies involved in the lab network must be as close to representing a real-world organization as possible and the lab must be controlled in a way as to avoid any possible external contamination to the experiment.

Lab Network Operating Systems

The most common operating system in use today is Microsoft Windows (Statistica, 2017) (Net Marketshare, 2017) and the version that is most common is Windows 7 (W3Counter, 2017) (Computer Hope, 2017) (Merriman, 2016). Therefore, the bulk of the lab network will consist of Windows 7 user devices in a domain with Windows 2008 domain controllers, as that is the closest kernel version to Windows 7 for a Windows server operating system. As a note of accountability, at the time of experiment design as well as during the offensive security assessments and simulated attacks, the remote code exploit for these kernel versions, MS17-010 (Microsoft, 2017), also referred to as ETERNALBLUE (Ullrich, 2017), had not been disclosed to the public or weaponized yet and did not impact the carrying out of this experiment.

The network required several Linux-based operating systems as well. As Ubuntu was the most popular and common Linux operating system (Hoffman, 2014), it was chosen to represent Linux platforms in the network. Another Linux distribution, Vyos (Vyos, 2018), was chosen as a routing and firewall platform for the experiment, given its proven history, administration support community, and reliability.

Lab Network Layout

As discussed earlier, the network was intended to be set up representing a law firm network. This required having multiple functional areas for the network as well as allowing communication between them and to the simulated Internet. The network would not connect to the actual Internet to avoid experiment contamination. In Figure 5-1, the three routing devices were using the Vyos operating system, the Internet, and intranet FTP servers, and Case Files Backup were using the Ubuntu operating system and the rest of the machines shown were using Microsoft Windows 7 or Server 2008 for desktop and servers, respectively.

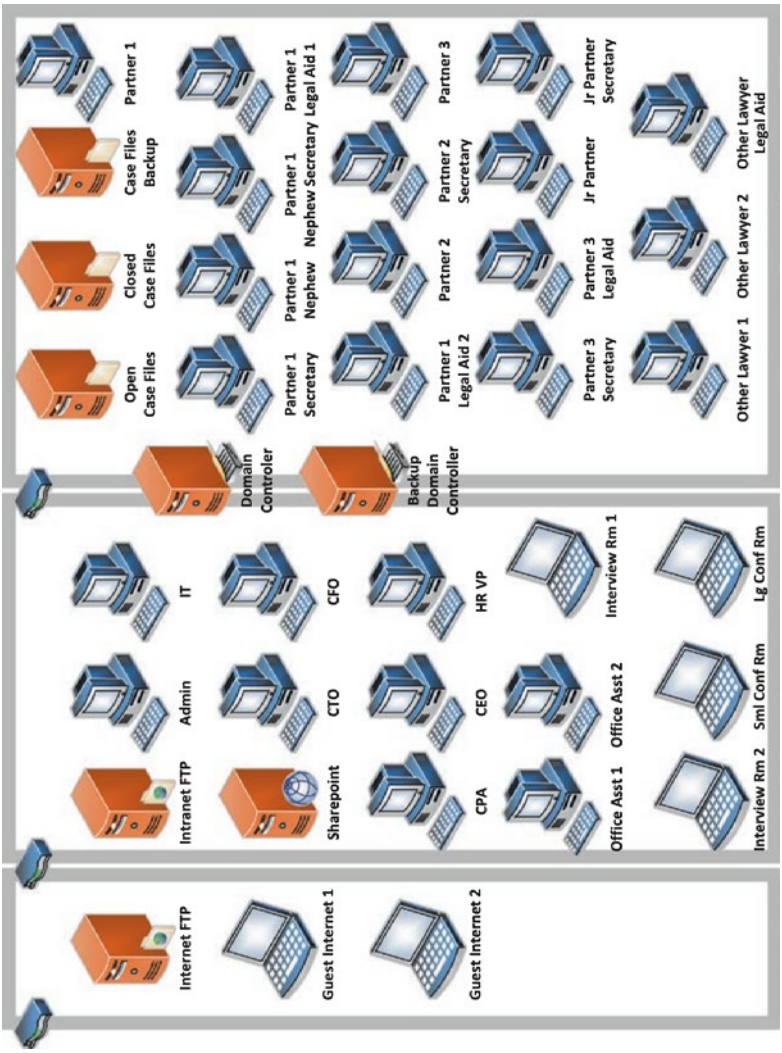

Figure 5-1. *Network Diagram*

Experiment Metrics

The purpose of this dissertation and experiment are to determine if the offensive security assessment paradigm of CAPTR teaming is a novel augment to traditional red teaming. Determining the novel nature of CAPTR teaming in comparison to traditional red teaming is shown via the categorical analysis of the assessment processes contained earlier in this dissertation. To lend a quantitative metric for novelty, this experiment will also allow for the two methods to provide findings which can be measured in their variance from one another to give a statistical idea of assessment uniqueness.

The experiment must also be able to determine the impact of recommendations to the security posture of the organization and its ability to mitigate advanced threats. To do this, the National Institute of Standards and Technology's Common Vulnerability Scoring System Calculator (NIST, 2018) was used to generate a numerical representation of the associated risk a given compromised machine would have to the organization as a whole. Typically, this calculator is used to determine a numerical score of the impact a given vulnerability has to a single system. For use in the experiment, the different machines are treated themselves as vulnerabilities and the organization is viewed as the system at risk. Therefore, the attributes that are input to create the overall score entered with this perspective. For example, if compromised by an attacker, a router within the organization would present the threat of traffic manipulation between two areas of the organization. The impact and difficulty of which are used in the CVSS calculator to give that device a score of 5.8. This value represents the device as a numerically measured vulnerability to the organization. Comparatively, a device such as a machine set up for clients to use to browse the Internet from within the DMZ is less of a vulnerability to the organization if compromised and represents a lower risk value of 3.4. This is based on the impact and difficulty of turning a compromise of this machine against the organization. The lethal compromise devices

within the organization are rated within the CVSS calculator to indicate the difficulty of turning the vulnerability of their compromise against the organization. This was done to include them within the overall risk value for the organization, even though, as lethal compromise items, their compromise would be exponentially critical in comparison to other devices.

Personnel Requirements

To provide as defensible an experiment, the performance of actions in the experiment needs to reflect expected behavior of such actors in the real world. To accomplish this, qualified personnel must be identified to perform the duties of the different actors within the experiment. Additionally, similarly qualified personnel will be identified to audit the actions of the individuals within the experiment to insure nothing is being done outside the bounds of normal activity. The following list indicates the personnel required to facilitate the experimental evaluation of the CAPTR team concept in comparison to that of traditional red teaming:

Systems Administrator

Systems Administration Auditor

Red Teamer

Red Team Auditor

CAPTR Teamer

CAPTR Team Auditor

Qualified and Sophisticated Attacker

Experiment Schedule and Walkthrough

The following is a list indicating the chronological series of events that are required for successful completion of this experiment. Following this list is an in-depth walk-through featuring the details of each phase of the experiment.

1. Control Network and related documentation created by Systems Administrator
2. Control Network audited for realism and functionality by Systems Administration Auditor
3. Control Network cloned twice by Systems Administrator and clone documentation created
4. Red Teamer assesses Network Clone 1
5. Red Team Auditor verifies the Red Teamer recommendations
6. Systems Administration Auditor verifies Red Teamer recommendations
7. Systems Administrator implements changes to Network Clone 1 based on Red Teamer recommendations
8. Red Teamer verifies changes were done in accordance with intent of Red Teamer recommendations
9. CAPTR Teamer assesses Network Clone 2
10. CAPTR Team Auditor verifies the CAPTR Teamer recommendations
11. Systems Administration Auditor verifies CAPTR Teamer recommendations

12. Systems Administrator implements changes to Network Clone 2 based on CAPTR Teamer recommendations
13. CAPTR Teamer verifies changes were done in accordance with intent of CAPTR Teamer recommendations
14. Red Teamer recommendations and CAPTR Teamer recommendations analyzed to indicate novelty metric of CAPTR team process
15. Simulated Attacker wages campaigns against Control Network, Network Clone 1, Network Clone 2
16. Metrics compiled to indicate mitigation of risk to organization in each campaign

Control Network and Related Documentation Created

The systems administrator creates a virtualized lab network in the image of one that could be utilized by a law firm. Devices within the network are configured and domains set up as well as user and administrative accounts. Documentation of the passwords, accounts, and device addresses is compiled. This lab network and its documentation will act as the control network for the experiment as it will simply have a functional level of configuration and no further security measures or alterations of configuration besides that which allow for intended communication and activity.

Network Audited for Realism and Functionality

The Systems Administration Auditor will go over the network documentation as well as network diagrams of the control network to determine if it is realistic and indicative of a functional network configuration. The network will also be audited with regard to its potential to skew the results of the experiment.

Control Network Cloned

The systems administrator will clone the now verified control network twice. This is to provide two separate swim lanes for the offensive security assessment paradigms to work within. The topology, types, and number of devices will remain identical to the control network. The hostnames, users, accounts, passwords, and IP addresses of the devices contained within the clones will be unique for each clone and separate as will the IP schemes themselves. This is to make them appear as unique as possible come the attack simulation portion of the experiment.

Red Team Assessment

One of the clone networks will be assessed in the traditional red team method by the Red Teamer. The assessment of this network will be done in a time window of ten hours to insure both assessments are concluded in equal time frames. The Red teamer will then provide recommendations based on the assessment findings.

Audit of Red Team Recommendations by Red Team Auditor

The recommendations of the Red Teamer are subjected to audit by a Red Team Auditor who is a separate qualified red team practitioner. This is to insure the recommendations from the Red Teamer fall within the scope of expected traditional red team assessment.

Audit of Red Team Recommendations by Systems Administration Auditor

The recommendations of the Red Teamer are further subject to audit by the Systems Administration Auditor. This is done to ensure that the changes suggested by the Red Teamer fall within the scope of activity a typical systems administrator would conduct and not outside the realm of realism.

Implementation of Red Team Recommendations

The Systems Administrator takes the verified recommendations of the Red Teamer and begins implementing them into the Clone 1 network using up to twenty hours of administration time. The Red Teamer is instructed to provide recommendations in an order of importance for implementation and are informed that the Systems Administrator will only have 20 hours to complete the changes to the network. This is done to keep the offensive security assessors from recommending varying amounts of changes for the security of the network which could skew results.

Verification of Red Teamer Recommended Changes

The Red Teamer is also responsible for auditing the implementation of changes conducted by the Systems Administrator based on recommendations of the offensive security assessment. The Red Teamer is to ensure that the changes were performed satisfactorily with regard to the intention of the Red Teamer. This prevents the Systems Administrator from poorly representing the assessment capabilities of the Red Teamer.

CAPTR Team Assessment

The CAPTR Teamer assesses Clone 2 of the control network. This is done in the same allotted time as the ten hours given to the Red Teamer. The CAPTR Teamer is sent network documentation and a letter indicating the spirit of the CAPTR team to the CAPTR teamer as well as scope and rules for the engagement. Recommendation guidelines are sent to the CAPTR teamer as well. The CAPTR Teamer will provide recommendations based on findings of the offensive security assessment.

Audit of CAPTR Team Recommendations by CAPTR Team Auditor

Similar to the recommendations of the red team, those of the CAPTR team are also audited by a separate party who is also qualified in offensive security and given the same intent of CAPTR team's information as the CAPTR Teamer. This will allow for third party verification that the changes suggested by this assessment method are in keeping within the spirit of CAPTR teaming.

Audit of CAPTR Team Recommendations by Systems Administration Auditor

Also, like the Red Team recommendations, those of the CAPTR team are subject to the same audit by the Systems Administration Auditor to determine that they fall within the scope of activity a typical systems administrator can be expected to perform.

Implementation of CAPTR Team Changes

The Systems Administrator takes the verified recommendations of the CAPTR Teamer and begins implementing them into the Clone 2 network also using up to twenty hours of administration time. The CAPTR Teamer is also instructed to provide recommendations in an order of importance for implementation and are informed that the Systems Administrator will only have 20 hours to complete the changes to the network. The Systems Administrator will provide a log of changes implemented into the Clone 2 network to the CAPTR Teamer.

Verification of CAPTR Teamer Recommended Changes

The CAPTR Teamer is also responsible for auditing the implementation of changes conducted by the Systems Administrator based on recommendations of the offensive security assessment. The CAPTR Teamer is to ensure that the changes were performed satisfactorily with regard to the intention of the CAPTR Teamer. This prevents the Systems Administrator from poorly representing the assessment capabilities of the CAPTR Teamer.

Recommended Changes Analyzed

The changes suggested by the two teams are compared to indicate whether or not the two offensive security assessment paradigms provided the same or different results. This is part of the basis for making the case that the CAPTR team paradigm is a worthwhile augment to established techniques. If the changes recommended by either team were nearly identical it would make a weak statement for the novelty of CAPTR teaming. If the changes were largely different then there is a stronger case for the paradigm.

Simulated Attacks

Cyber-attack campaigns are conducted against the control and clone networks. The Attacker is instructed to replicate motivated and sophisticated attacks against the organization in each of the three campaigns. The Attacker is informed that the organization for all three campaigns are legal firms and that the goal is to compromise as much of the network as possible with the specific goal of finding case files as they are the item of lethal compromise for these organizations. The attacker is given a maximum of 40 hours to conduct each of the cyber-attacks from the access provided, which is as earlier discussed, a user context implant running as if by successful spear phishing. The order of the campaigns is unknown to the attacker; however, the Control was attacked first, the Red

Team secured network second, and the CAPTR team secured network third. This was to ensure that if the Attacker gained any proficiency as the attack campaigns were completed that the attacks would be most proficient against the CAPTR team secured network, and any bias this created would make attacks against the CAPTR Team network most likely to be successful and, if anything, skew results against the CAPTR team model.

Metrics Compiled

Once the campaigns are completed, the compromised devices are tallied and a percentage of the overall risk present in the network secured is identified for each. This is done to provide a quantitative measure of the amount of risk mitigated by the changes recommended by the offensive security assessments.

Addressing Defensibility Requirements

Briefly, this section summarizes ways in which the aforementioned experiment is able to address the requisite characteristics for defensibility. The virtualized lab simulation of a network serving as a replica of potential real network servicing a law firm means that it is both controlled and a realistic situation to conduct both offensive security assessment and attack simulation. Further, the great lengths taken to guarantee remote communication of actors while maintaining a contaminant-free experiment mean that no outside actor or incident will affect the lab network.

Addressing Defensible Security Assessments

Using a lab network not connected to the Internet means that security assessment is conducted in a vacuum, free of user- and administrator-created events that may unfairly help or hinder one assessment methodology over the other. The use of industry-qualified offensive security experts in the carrying out of the assessments provides both

defensibility to their assessment as well as furthering realism. Additionally, having the assessments audited by similarly qualified separate third-party offensive security experts means there is an extra level of validation for the legitimacy of the assessments and the generated recommendations provided from them. The equal limit of time and like recommendation guidelines means that both assessment paradigms have fair assessment engagement windows and know the time restrictions on the administrator ahead of time.

Addressing Defensible Systems Administration

Ensuring the networks were created and administered across the separate assessment platforms by the same administrator insured that one network did not receive more or less qualified systems administration than the other. The audit of the networks themselves by a separate third-party qualified systems administrator prevented the lab network from failing to represent a realistic operating environment. The audit of the assessment recommendations from both teams by a third-party systems administrator insured that the implementations needed were within the scope of typical systems administration and would not skew the outcome of the test in favor of one assessment paradigm over the other. The equal limit of time for change implementation across both assessed networks kept the implementation of security fair between both assessed networks. Lastly, the presentation of change logs regarding the assessor recommendations back to the assessor insured that the changes done to the networks were in keeping with the intention of the assessors.

The use of an extremely qualified cyber operations expert and senior red team member with experience performing APT emulation allowed for an equal level of sophistication to be applied to all three attack campaigns. The level of skill maintained by the attacker meant that the networks were more likely to see deeper assessment penetration and therefore changes recommended by the assessors were more likely to face attacker scrutiny. Having a simulated attacker means that no outside attackers could influence the emulation campaigns and, therefore, it would be similarly

capable of targeting each of the three networks. The brief to the attacker on specific motivation for the legal firm's case files, in addition to wanting the whole network compromised, meant that the actor had a distinct purpose that was the same for all three networks, which achieved a fair level of motivation in all three campaigns.

Addressing Measurable Results

The comparison of number of recommendations and their uniqueness between the two evaluated assessment paradigms allowed for a measure of novelty between the suggested CAPTR Team paradigm and established red team practices. Utilization of the NIST-provided CVSS calculator to calculate the risk each compromised machine allowed for a comparable quantitative evaluation metric. This allowed the experiment to grade the success of the paradigms in protecting overall risk as well as the ability to directly compare the paradigms to each other.

Summary

The information technology industry is really good at benchmarking and evaluating newer and better security hardware or software, but not so much "wetware" (humans). That fact is problematic for innovation in industry and, I suspect, is probably the largest reason academic innovation mostly avoids research into human-driven cybersecurity implementation assessment processes. I can easily prove my encryption technique is better if it has less overhead or makes data more secure. I can readily show how my software alerts on more data than existing products. It is really hard to show my cybersecurity implementation stands up to the human-involved attack tradecraft and human-involved operations. This chapter presented defensibility requirements for experimentally comparing

cybersecurity implementations against each other. It also touched on the high level of difficulty and the dire need for continued improvement if we are going to push the envelope on theoretical cybersecurity.