Applying the scientific method to cybersecurity research

Conference Paper · May 2016

DOI: 10.1109/THS.2016.7568886

CITATIONS
2

READS
1,555

12 authors, including:

Mark Tardiff
Pacific Northwest National Laboratory
28 PUBLICATIONS **151** CITATIONS
SEE PROFILE

George T. Bonheyo
Pacific Northwest National Laboratory
49 PUBLICATIONS **1,202** CITATIONS
SEE PROFILE

Katherine Allen Cort
Pacific Northwest National Laboratory
35 PUBLICATIONS **142** CITATIONS
SEE PROFILE

Christopher S Oehmen
Pacific Northwest National Laboratory
64 PUBLICATIONS **677** CITATIONS
SEE PROFILE

Applying the Scientific Method to Cybersecurity Research

Mark F. Tardiff, George T. Bonheyo, Katherine A. Cort, Thomas W. Edgar, Nancy J. Hess, William J. Hutton III, Erin A. Miller, Kathleen E. Nowak, Christopher S. Oehmen, Emilie A. H. Purvine, Gregory K. Schenter, Paul D. Whitney

National Security Directorate
Pacific Northwest National Laboratory
Richland, Washington

The cyber environment has rapidly evolved from a curiosity to an essential component of the contemporary world. As the cyber environment has expanded and become more complex, so have the nature of adversaries and styles of attacks. Today, cyber incidents are an expected part of life. As a result, cybersecurity research emerged to address adversarial attacks interfering with or preventing normal cyber activities.

Historical response to cybersecurity attacks is heavily skewed to tactical responses with an emphasis on rapid recovery. While threat mitigation is important and can be time critical, a knowledge gap exists with respect to developing the science of cybersecurity. Such a science will enable the development and testing of theories that lead to understanding the broad sweep of cyber threats and the ability to assess trade-offs in sustaining network missions while mitigating attacks.

The Asymmetric Resilient Cybersecurity Initiative at Pacific Northwest National Laboratory is a multi-year, multi-million dollar investment to develop approaches for shifting the advantage to the defender and sustaining the operability of systems under attack. The initiative established a Science Council to focus attention on the research process for cybersecurity. The Council shares science practices, critiques research plans, and aids in documenting and reporting reproducible research results. The Council members represent ecology, economics, statistics, physics, computational chemistry, microbiology and genetics, and geochemistry.

This paper reports the initial work of the Science Council to implement the scientific method in cybersecurity research. The second section describes the scientific method. The third section in this paper discusses scientific practices for cybersecurity research. Section four describes initial impacts of applying the science practices to cybersecurity research.

I. INTRODUCTION

The cyber environment has rapidly evolved from a curiosity to a critical element in many aspects of the contemporary world including commerce, law enforcement, national defense, medicine, and the conduct of our personal and professional lives. Cybersecurity research emerged in response to adversarial attacks interfering with or preventing normal cyber activities. As the cyber environment has expanded and become more complex, so have the nature of adversaries and styles of attacks.

Historical response to cybersecurity attacks is heavily skewed to tactical responses with an emphasis on rapid recovery. All cyber systems, including government and corporate entities with top-of-the-line security systems, are vulnerable to attacks [1, 2]. Despite continuous innovations in defensive systems, cyber incidents are an expected part of life. The demand for rapid progress and the diversity of defensive research presents new challenges: how to make the development cycle more efficient and rigorous and how to enable meaningful testing and comparisons. The careful analysis of systems, defenses, threats, attacks, consequences, and recovery is needed to provide useful observations and to develop an understanding of fundamental phenomenologies associated with cyber systems and their security.

A knowledge gap exists with respect to developing the science of cybersecurity. In other technical fields, a scientific method is typically developed and applied to address these needs. The benefit of such a science is the development and testing of theories that lead to understanding the broad sweep of cyber threats and the ability to assess trade-offs in sustaining network missions while mitigating attacks. Applying the scientific method also leads to findings and conclusions that are repeatable, and verifiable, thus providing a dependable knowledgebase to the broader cybersecurity community.

The need for a formal scientific approach to cybersecurity is widely acknowledged. The JASONs published a report on applying the scientific method to cybersecurity [3]. Cornell University has a science of cybersecurity blueprint [4]. The National Security Administration launched its 'Science of Security Lablets' program to engage universities [5], and 21CT has papers addressing the science of cybersecurity [6]. These published reports help demonstrate the substantial agreement that a more formal approach to cybersecurity is necessary and the conversation on how to do so is becoming pervasive.

The research described in this paper is part of the Asymmetric Resilient Cybersecurity initiative conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

The Asymmetric Resilient Cybersecurity (ARC) Initiative at Pacific Northwest National Laboratory is an internal multi-year, multi-million dollar investment to develop approaches for shifting the advantage from the attacker to the defender and sustaining the operability of cyber systems under attack. The initiative is predicated on the proposition that significant advances in sustaining cyber resources in the face of ongoing attacks requires more strategic approaches that sustain mission functionality and raise the costs or risks to attackers.

The ARC Initiative established a Science Council comprising active researchers from multiple disciplines to focus on the research process as it is applied to cybersecurity. The Council charter is to share common science practices for conducting research, critique research plans and experiment designs, and document and report research results that can be reproduced by others. The members of the Science Council represent ecology, economics, statistics, physics, computational chemistry, microbiology and genetics, and geochemistry. The Council also includes two members who are active cybersecurity researchers to keep the Council grounded in the realities of the cyber domain.

This paper reports the initial work of the Science Council to implement the scientific method in cybersecurity research. Section two describes the research context along a science continuum and the scientific method. Section three discusses scientific practices for cybersecurity research. Section four describes the experiences and initial impacts of applying the science practices to cybersecurity research. Section five presents a discussion and conclusions.

II. Scientific Method

Topic areas in research domains undergo development and maturation. Very often, emerging topics are poorly understood and initial investigations are invested in characterizing the phenomenology and developing the initial conceptual model (inductive reasoning – using specific observations to pose a general statement). Initial conceptual models support the development of research questions that can be formalized as falsifiable hypotheses and investigated with experiments (deductive reasoning – testing a general statement or theory using specific observations). The goal for applied research is to eventually develop a sufficient understanding such that mature models can be validated for operational use. Figure 1 represents this sequence as a science continuum of discovery.

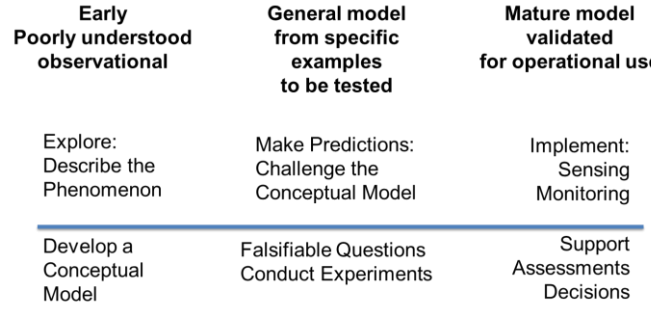

Fig. 1. Science continuum of discovery.

Research has many different starting points; it is not always in an orderly left-to-right progression along the science continuum. Conceptual models fail in the face of experimental evidence and attempts to support decisions become muddled with extenuating circumstances because the model does not generalize sufficiently to operational environments. To this end, failures in the form of unexpected experimental outcomes are often critical to the eventual success of research. Rigor in experimental design and analysis provides the means to assess results to find causes or identify new questions that reshape the model.

The "scientific method" serves two purposes in advancing our knowledge of the world around us. At its core, it is a set of techniques or steps for investigating some phenomenon. In practice, the steps are often taken out of order and very likely iterated to achieve a better understanding of the thing of interest. Implementing the steps eventually leads to defensible advances in our understanding of the phenomenon because it subordinates human beliefs or preferences to empirical evidence. The scientific method is also a philosophy of discovery that depends on the rules of logic, healthy skepticism, and the integrity of investigators to accurately represent their research. The philosophy of science is a formal discipline in its own right with a substantial literature, [7, 8, 9]. A consequence of the scientific method is that for research to have impact, results must be made available to the research community where they can be reviewed, challenged, repeated, and verified.

Researchers have made many attempts to articulate the fundamental steps of the scientific method. The difficulty lies in the differences between how science is described and how it is conducted. Science, as practiced, is often roiled with false starts, miscommunications, hidden assumptions, disciplinary dogmas, and occasional insights. Nonetheless, listing the steps is valuable for identifying where a research effort is relative to the desired outcome of defensible and repeatable results. An example of the scientific method from Crawford and Stucki [11] is as follows:

1. Define a question
2. Gather information and resources (observe)
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment and collecting data in a reproducible manner
5. Analyze the data
6. Interpret the data and draw conclusions that serve as a starting point for new hypothesis
7. Publish results
8. Re-test (frequently done by other scientists).

Implementing the scientific method relies on a conceptual model that represents the researcher's current understanding of the investigation target. The conceptual model documents the mental model of how the thing of interest works and what opportunities exist for making measurements and advancing our understanding. When data are collected, an underlying

conceptual model informs what types of measurements to make and where and how often to make them. Failing to make an explicit conceptual model often leads to ill-posed research questions, poorly designed experiments, and conclusions that are not supported by the evidence. The conceptual model also establishes the scale at which the phenomenon of interest is investigated. A useful hypothesis in the form of a falsifiable question directly challenges the conceptual model, leading to increased confidence in the model or revisions based upon the experimental outcomes.

III. Science practices for cybersecurity

The first key hypothesis for the Science Council is that science practices developed for other disciplines can be applied to research in cybersecurity to improve the design of experiments and thereby generate meaningful, repeatable research outcomes. Achieving experimental results that can be repeated and refuted or confirmed is fundamental to science.

The second key hypothesis for the Science Council is that large complex problems are intractable because it is impossible to conduct controlled experiments directly on them. They can be addressed by identifying and investigating smaller sub-problems. The results from multiple sub-problem experiments can be integrated to gain insights and develop a generalized understanding and theories about the large problem. An example of this progression is the research in atmospheric sciences to investigate climate change. While it is impossible to conduct controlled experiments on the atmosphere, numerous laboratory-scale experiments are being conducted to understand physical and chemical processes associated with atmospheric phenomena.

Science practices from other domains that address complex systems, including health care, social and organizational dynamics, and computer network systems should have recurrent themes. Crawford and Stuck [11] reviews a standard taxonomy of studies in the health sciences along with standard practices for each type of study. The experiments described in this paper are similar to randomized controlled trials and diagnostic test studies from that reference. Lohr [12] collects papers on experimental work in political sciences. There are linkages between the approach described here and best practices in that paper.

The Science Council identified the following eight practices that can be applied to cybersecurity research to improve the quality of experiments and generate repeatable outcomes.

A. Define a Tractable Problem

Many cybersecurity research problems are difficult to resolve because they are sensitive to the scale at which they are defined and have many explicit and latent variables that influence results. It can be difficult to be confident that the experimental outcome is due only to the variables investigated by the experiment and not due to other uncontrolled or uncontrollable variables. problem is common with other domains that conduct observational science such as economics and ecology.

A solution is to divide large complex problems into sub-problems or phenomena that can be constrained and investigated, much like the atmospheric sciences example given above. The goal is to investigate the phenomenology, develop mechanistic models of how things work, and use the models to examine the problem at larger scales. This is often challenging because it is not obvious how to cleave the large problem into sub-problems that are useful to understanding the whole. Bounding a tractable research problem often requires iterations. Whether the problem is sufficiently constrained to generate useful results can only be determined by conducting experiments.

In the process of identifying a tractable problem, researchers should be explicit about the coherence of the proposed research to the larger and more complicated question. To that end, it is valuable to consider and document responses to the following questions:

1. What is the general problem that is being addressed?
2. How does the project (sub-problem) relate to the general problem?
3. What specific research questions does the project pose?
4. If the project is successful, how will it impact progress on the general problem?

An example complex problem is to develop a metric for the relative security of a cyber system. The problem statement infers that the idea of “relative security” is meaningful and that it can be quantified at least at an ordinal level (i.e., this state is better than that state). The cyber system and the type of attack are both undefined and the components of the cyber system to be monitored are not stated, indicating that the whole undefined system is to be monitored with a single metric to determine the state of security.

A more tractable problem statement is to develop a metric for detecting data exfiltrations that are 50MB or larger from a data storage device accessed by multiple users with password-protected privileges. The problem is better constrained in that detection infers a binary indicator of exfiltration or no exfiltration. The subsystem is a data server with multiple users indicating that legitimate data access is ongoing and will need to be addressed as false positives for exfiltration. The size of the exfiltration sets an initial target for what constitutes a meaningful loss of information and helps to manage the false positive problem. The sub-problem statement is directly relatable to the larger problem statement and progress in addressing this research question would very likely be useful in addressing the larger problem.

B. Preliminary Data Assessment

Research builds on prior knowledge and preliminary work is often necessary to collect observations to build a rudimentary understanding of a system. Collecting the results from prior work by others helps to frame the problem by evaluating what is known about the system of interest. There is also value in understanding how past investigations were

conducted and their limitations or shortcomings. When data are available to address the current research problem, questions regarding sufficiency and provenance also need to be addressed. Often new data are required. Careful attention and planning are necessary to make sure the context of the experiment is documented along with the observations.

C. Develop Falsifiable Research Questions

Generating falsifiable hypotheses is a key component to conducting useful science that turns a naturally inductive process into a deductive one. The logical underpinning for needing falsifiable questions is that it is impossible to prove a negative. The classic example is that we may see a thousand white swans but have no basis for saying that swans cannot be black. Even if we collect more observations of swans, there is always another swan that we have yet to see that might not be white. This is the inference of a general statement from a set of individual cases. Conversely, we can pose the hypothesis that all swans are white and reject that conjecture when we find one black swan. In this case, we state a general swan model and reject it with a single instance that proves it to be false.

In the cyber environment, we might have the following research question: Can our event log metric detect differences in unauthorized file access by different attacks? The falsifiable form of this question is: The event log metric shows no significant differences in response across attacks A through D on network configuration alpha. An experiment can be designed to test this question by collecting data on instances of each of the four attacks on the specified network configuration. A statistical comparison of the metric response to the four attacks can be used to reject or fail the hypothesis.

Hypothesis testing in the context of conducting an experiment results in one of three outcomes tied to one of two implications. One outcome is to revise the hypothesis or experiment because the results indicate that the wrong features were measured to address the original question. A second outcome is to reject the null hypothesis of no difference leading to the statement that, based on the experimental data, there is a significant difference. The alternative outcome is that based on the experimental data, there is insufficient evidence to reject the null hypothesis. This is a very different from "accepting the null hypothesis." This distinction is important because the significance level for rejecting the null hypothesis is not the same as the significance level associated with correctly deciding that there is no difference. Additional information on hypothesis testing and rejecting the null can be found in many sources, [13, 14].

The implications of the hypothesis test can be that the detected difference is practically meaningful or that the difference, while statistically significant, is not practically meaningful. The distinction highlights the point that statistical tests do not make decisions. Humans make decisions using the outcome of statistical tests as evidence to support their choices. Large amounts of data with small variances can lead to statistically significant differences with little or no practical value either because small changes in the environment being measured generate false positives or the attributes being measured actually do not vary sufficiently to matter.

D. Identify Ground Truth

Ground truth data are observations of the phenomenon where the state for particular variables is known with certainty. If we are interested in measures of network resilience, then we need instances of networks that have no attacks (benign) and networks with known attacks (compromised) to determine whether we can sense a change in resilience. We might be interested in a binary resilience indicator (compromised, yes/no) or an incremental indicator that shows both change and magnitude. Initial investigations may be purely phenomenological in order to get initial ideas of "how things work." Eventually, the research needs to progress to investigations where hypotheses are posed and the experimental environment has ground truth for the outcomes. Developing a ground truth is often challenging, but without it, the experimental results are only anecdotal.

The need for ground truth also motivates having an experimental framework or testbed where the system behavior is sufficiently characterized such that outcomes to perturbations can be anticipated. The testbed is represented by a conceptual model and the falsifiable questions challenge that conceptual model to advance our understanding of the testbed and ultimately the system that the testbed represents. In the example above, developing a measure of resilience by investigating the World Wide Web is an intractable challenge. Developing a testbed network with a known size and configuration that can be compromised with specified attacks at known times makes it possible to develop measures of resilience. Eventually the measures need to be evaluated for their usefulness on unconstrained networks, but those unconstrained networks are not the place to develop them.

E. Document Assumptions

It is very unlikely that experimental investigations can be conducted without making assumptions. The assumptions run the gamut of how the experimental environment is defined, which variables are measurable, how variables and parameters are related, which variables are more important, which variables can be controlled, what measurement and data analysis methods will be used and why, and what defines a useful outcome. These assumptions need to be stated and documented at the beginning of the experiment. As the work progresses, the assumptions need to be revisited and updated. Very often there are latent or implicit assumptions that are second nature to the investigators and turn out to be significant to the outcomes. Some basic assumptions for computer networks are that the domain name system will use port 53, hypertext transfer protocol uses port 80, and all security patches on the system are up to date.

There are also significant assumptions regarding how a tractable sub-problem relates to the larger complex problem, how results from experiments on the sub-problem will inform insights to the larger problem, and the relevance of a testbed or simplified experimental environment to operational conditions. Inadequate attention to assumptions often diminishes the impact of research because the context for the experiment and results is not properly represented. Attempts to replicate research results without sufficient knowledge of the assumptions are very likely to fail.

F. Test Tools and Assumptions

Science depends on collecting data, conducting analyses, and making inferences. Conducting an experiment necessarily involves ways to sense the experimental environment, including instruments, measurements, and/or algorithms. A critical first step is to test the tools against simple problems with known outcomes to confirm that they are working correctly, much like calibrating an instrument before and after taking measurements to confirm that the instrument is responding as expected (in control).

This practice is especially important for cybersecurity research where the interesting questions are typically not subject to constraints such as gravity, momentum, or conservation of mass, which tend to bound measurement principles and expected outcomes. The advent of "large data" also increases the value and impact of testing measurement methods. Measurement techniques need to be validated to reduce the possibility of generating data that are artifacts of the collection method instead of measurements of the phenomenon of interest. Digital sensor systems often increase in their response to increasing signal to a point and then fall off because the rate of events exceeds the signal processing time. Under these conditions a very large signal can look like the absence of any signal because the measurement system is paralyzed.

Assumptions are often beliefs or widely accepted concepts within a domain. It is important to watch for evidence that assumptions are not valid. Depending on how fundamentally important they are to the experiments and research designs, it is often highly beneficial to conduct simple proof-of-principle investigations to validate assumptions before interrogating the problem directly. Failing to test assumptions can lead to substantial errors in interpreting results because the design and execution of the experiment was flawed with misunderstandings.

G. Start with Simple Experiments

Investigating simple versions of the problem first and adding complexity as the results, form well-founded confidence that helps ensure the phenomenon is correctly represented by the model or models. Simple experiments entail a limited number of variables and serve to determine the range of effects and sensitivity of measurements. However, as illustrated in Figure 2, simple experiments rarely emulate real-world conditions and it is possible that synergistic effects between variables may be absent if these variables are only tested individually. It is also important to review experimental results with healthy skepticism and to look for flaws in the approach that could have generated great results for the wrong reasons. If we cannot perform well on a simplified problem, it is unlikely that performance will improve with increased complexity. The progression from simple to complex often reveals unexpected elements and complications that challenge our understanding of the problem and often cause us to rethink what we knew to be true.

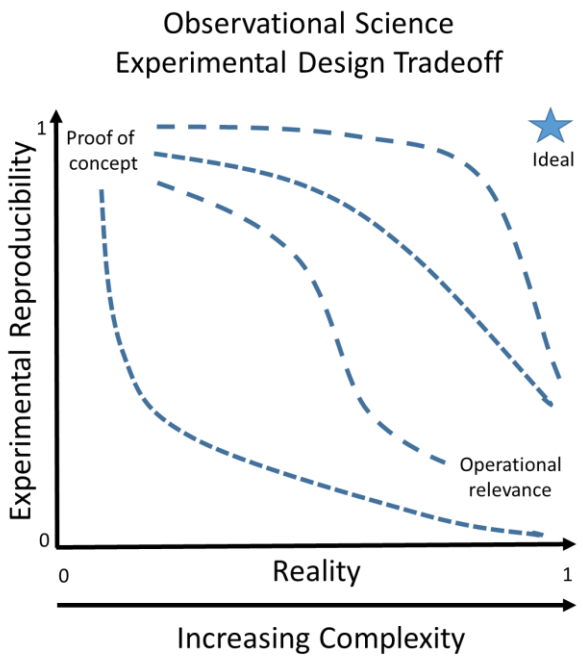

Fig. 2. Schematic view of experiment reproducibility related to reality. Scale bars are normalized to a value of one, representing absolute reproducibility or a true, real-world setting.

H. Assess Progress Toward the Larger Problem

The goal of the simple experiments is to make progress on understanding the large complex problem. This requires periodically evaluating progress toward useful models of the sub-problems and how these inform the larger problem. If the components of the problem are well-characterized and represented as mechanistic models, they should respond to perturbations in predictable ways. As the models mature, the models and experiments may become intertwined in the sense that a mature model can be used to predict responses in experiments. If these responses differ from expectations, the results are used to update the model. As understanding and models progress, both need to be validated against earlier results to avoid over-fitting the model to a particular experimental result and losing generalizability to the larger problem.

IV. APPLYING SCIENCE PRACTICES TO CYBERSECURITY

The ARC Initiative is investigating approaches to enhance the resilience of cyber networks given pervasive attacks. While the goal is to develop approaches to maintain mission capabilities during attacks, the ability to detect attacks is a necessary part of assessing the efficacy of resilience strategies. Toward that end, the initiative invested in developing a variety of sensing capabilities. This section will describe one of the attack sensing approaches, two attacks used in the experiments, and three testbed designs. The discussion will focus on how orienting toward the scientific method and science practices informed the research.

A. *Persistent Homology for Network Sensing*

Cyber networks are often monitored and summarized with Netflow traffic data. The records consist of a time stamp, sending and receiving IP (Internet Protocols) addresses, sending and receiving ports, and packet counts. These data are amenable to analysis with graph methods using various definitions of nodes and edges depending on the research question. However, graphs defined in this manner often evolve over time in unpredictable ways, due to noise and residual traffic that is of little interest. One strategy for contending with the unwanted variations is to work with topological representations of the graphs. Topology is a field of mathematics that studies geometric properties of spaces that are preserved when the space is continuously deformed.

Persistent homology [15] is a topological method that investigates features in spatial data that are preserved as influence spheres around each point expand and merge, much like the human brain merges the pixels on a computer screen to make a smooth image. In our application, baseline signatures of IP addresses are constructed from packet traffic with other addresses over discrete time intervals and a topological representation of the baseline is produced. New data beyond the baseline are collected into time intervals. The baseline plus each additional interval is used to generate a new topological representation. The persistent homology of this updated topology is compared to that of the baseline topology with a dissimilarity metric [16]. The metric scores can then be used to develop a classifier for determining whether an instance of an updated topology is different from the baseline.

One of the challenges in analyzing network traffic with this approach is that graphs and topologies are abstractions of the raw data. If we collected data in the wild and applied persistent homology and a distance metric to it, we would not know if the metric is responding to interesting or spurious events. This problem motivates the need for a testbed environment with known ground truth events to test the method.

B. *Testbed and Attack Design*

The initial attack was a lateral movement worm that increased the number of talking IP addresses exponentially on each iteration. The testbed was a simple model of an enterprise network. The focus was traffic flow with a network complexity and scale to generate noise to hide the attack signal. Two networks were created to represent the segmented structure of typical enterprise networks. The first network included user workstations and web application servers providing enterprise services. The second network provided the connectivity between the web application servers and the databases that support their operation. The web application servers all had two network connections and there was no direct route between user workstations and the database servers. The enterprise model included 2,000 user workstations, 10 web servers, and 20 database servers. Each network was tapped and packets were continuously captured.

The experiments were designed to investigate lateral movement of an attacker through the network. The first experimental trial was to assume an aggressive, timed lateral movement that would mimic a worm progression to take over the network. To instantiate this model, a Python script was created that exponentially jumps each time interval. The script was set to wait one hour and five minutes between jumps. The multicast domain name system was configured so that the worm could find a target list of IPs for its next jump series. As the progression of the lateral movement was the subject of study and not the method, secure shell was used with preset credentials to enable the script to progress through the virtual machines. The script started on one user workstation in the network. The script was set to multiply indefinitely.

The Netflow data captured during the network emulation were analyzed with the persistent homology method using a total density vectorization scheme. That is, for a given time window, a vector was constructed whose length is the number of IP addresses in the network and the kth component of the vector is the number of packets sent/received by the kth IP address. The initial 26 minutes of data were used as the baseline and 30-second overlapping time intervals that advanced by 10 seconds in each step were used sequentially to update the baseline and compute a dissimilarity score. The baseline was held constant. Figure 3 shows the trace of the dissimilarity score against the window sequence. The red line at the beginning identifies the start of the lateral movement.

The hypothesis was that we would see the advance of the attack and specifically at some point the signal of the attack would be strong enough that the one hour and five minute periods would be evident in the dissimilarity scores. The plot is substantially different from our expectations. The score elevates to around 700 then drops to zero at 200, 1000, and 1400 windows. There is also a plateauing behavior from about 200 to 1000 windows. Finally, the score drops to zero at around 2000 windows, recovers slightly, and dies out at around 3500 windows.

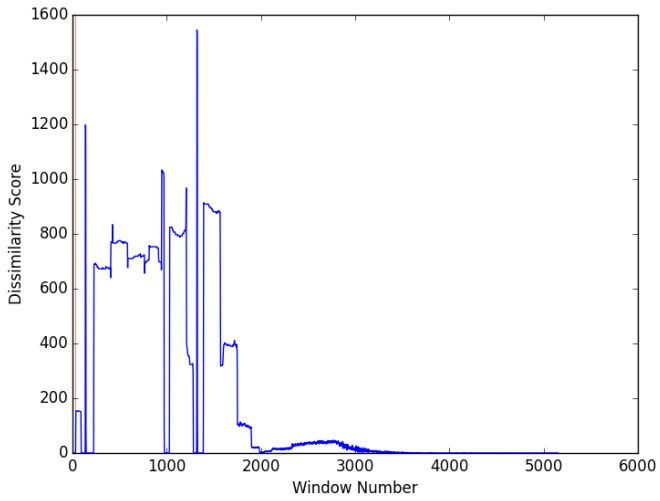

Fig. 3. First experiment, dissimilarity metric scores over time as window number.

Several explanations for these behaviors were proposed but none of them were supported by the data. The realization was that the testbed complexity prevented us from generating an expected phenomenology from the attack, as represented by the dissimilarity score generated from the persistent homology. As

discussed in the science practices, these results indicate the need for a simpler experiment.

The next iteration of the experiment was to generate a much simpler testbed such that the dissimilarity score exhibited predictable behavior. The testbed design consisted of:

1. A single subnet (10.0.0.0/24) of 16 hosts (IP addresses 10.0.0.1 to 10.0.0.16).

2. 10.0.0.1 is running a web server on port 80.

3. Each of the other hosts (10.0.0.2 to 10.0.0.16) contact the web server using the following Markov chain.

   a. 50% of the time, a "user" will browse to the home page of the web site (index.html).

   b. 30% of the time, the "user" will then "click through" to a subsequent page after requesting the home page (./primary/[color].html or ./secondary/[color].html).

   c. 20% of the time, the "user" will go directly to one of the subsequent pages (bypassing the home page), simulating a "bookmark".

4. There is a random time offset (delta T) that each client uses so the traffic appears "random" but is really based on a 60-second loop, with each client requesting a web page at 60 – delta T.

The experiment started by simulating normal traffic HTTP without attacks for 15 minutes. After 15 minutes (1:00 PM local time), two attacks were launched sequentially from host 12 (10.0.0.12) that lasted just over one minute. A ping sweep of the entire network was conducted that took ~30 seconds and a port scan of the web server on 10.0.0.1 that took ~40 seconds. Finally, there was another 15 minutes of normal traffic.

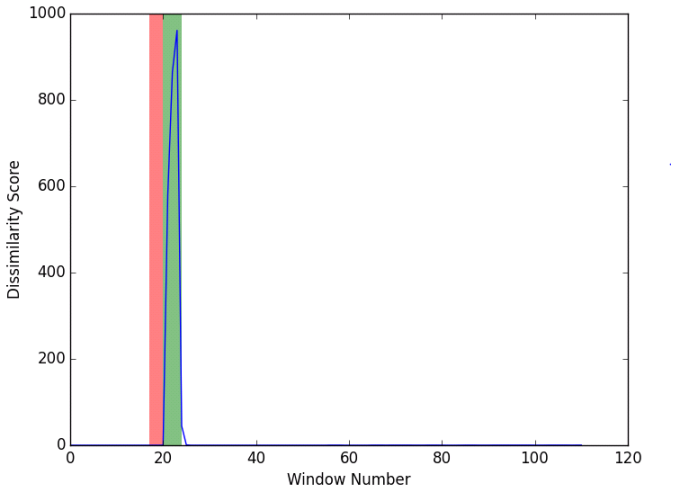

Fig. 4. Second experiment, dissimilarity metric scores over time as window number for a ping sweep and a port scan.

Figure 4 shows the trace of the dissimilarity score against the window sequence. The baseline of the metric was developed using the first 12 minutes of data. The windows are 30 seconds in duration and advance 10 seconds at each time step. In this much simpler experiment it is obvious that the ping sweep attack (salmon-colored bar) was missed and that the score was sensitive to the port scan attack (green bar). The overall phenomenology is much easier to comprehend and investigate, indicating that this experiment's level of complexity is better for the proof of concept that dissimilarity scores based on the persistent homology representation of the data could show a response to at least one of the attacks. Another desirable outcome is that the windows outside of the attacks are quiet and show very little difference from the baseline, relative to score response to the port scan. The method detected the attack and returned to the baseline, implying that it could detect a subsequent attack.

The lack of response to the ping sweep was investigated and was found to be an artifact of the experimental setup and not an issue with the scoring metric. The phenomenology of the ping sweep is not captured by the Netflow data because the ping sweep was executed using a protocol that is not captured by Netflow.

Figure 5 shows the next experiment with the same testbed and two identical port scans separated by five minutes of normal traffic. The attacks were separated in time because with adjacent attacks there could be latencies in the score responses that could be missed. This plot shows that the scoring metric is responsive to the port scans at their inceptions and returns to baseline at the completion of the scans without latency. The metric also appears to be stable in its response to the scans with very similar maximum values. A future experiment will be to change the rate of the port scans to assess changes in the response of the metric at different attack intensities.

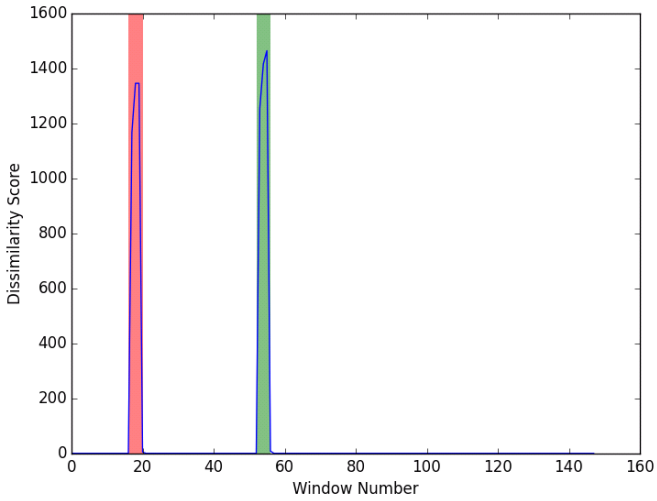

Fig. 5. Third experiment, dissimilarity scores over time as window number for two port scans separated by five minutes.

V. Discussion and Conclusions

Given the inherent complexity of cyber environments and the difficulty of conducting repeatable experiments in the wild, the Science Council engaged with the ARC projects to develop simpler proof of concept experiments on virtual networks. The value of the experiments was that they have ground truth for when the network traffic is benign, when the network is being

attacked, the network post attack, and what that attack is. These simpler experiments include attributes consistent with more complex and realistic experiments.

Developing a simple network and attack system for evaluating the performance of sensing metrics required several iterations. The first testbed had sufficient complexity that changes in the performance metric could not be unambiguously related to changes in the network traffic. The simpler network design allowed a much more straightforward relationship between the performance metric and the network attacks to emerge, providing confidence in the metric. These experiments are repeatable; the specifications for the network, the attacks, and the performance metrics can be shared with and implemented by other researchers. The results can be verified and the research extended by the community.

A critical question that remains is whether the results from simpler simulation experiments are helpful to addressing the larger question of maintaining mission functionality of a network during cyber-attacks. Assessing the resilience of a cyber network to attacks implies that there are methods for determining that an attack or attacks are ongoing along with some measure of the relative severity of the threat that informs a decision between consequences of the attack and any penalties incurred by the response. These initial experiments demonstrate that a metric based on persistent homology can sense the change in network traffic caused by the attacks. Given these results, additional experiments are being designed to evaluate the sensitivity of the metric to the magnitude of the attack signal. Other experiments will be conducted in which the complexity of the testbed is increased to learn about developing a classifier based on the metric and whether classes of attacks can be distinguished from more complicated benign traffic. Other metrics have been developed and will undergo equivalent testing.

Implementing the Science Council raised two questions: Are science practices from other domains relevant to cyber research, and is there value in sub-setting large and complex cyber problems into more tractable experiments? The work discussed above can be repeated by others, and there is a path to making progress toward increased realism, as represented in Figure 2, while still having reproducibility. The simple experiments were necessary to gain confidence about the viability of the prototypical metrics. Significant work remains to be done before the ARC research can be applied directly to deployed cyber environments. The initial indications are that the Science Council and the science practices have been beneficial to developing results that form the basis for added complexity and realism. It remains to be seen whether a "theory and science of cybersecurity" can emerge from these efforts.

REFERENCES

[1] http://www.computerweekly.com/news/1280095471/RSA-hit-by-advanced-persistent-threat-attacks

[2] http://www.telegraph.co.uk/technology/google/8553131/Google-Gmail-cyber-attack-Chinese-spies-had-months-of-access.html

[3] JASON, MITRE. Science of Cyber-Security. Nov. 2010. fas.org/irp/agency/dod/jason/cyber.pdf

[4] Schneider, F. B., Blueprint for a science of cybersecurity. The Next Wave. Vol 19 No. 2. 2012

[5] https://ecommons.cornell.edu/bitstream/handle/1813/22943/SoS%20blueprint.pdf?sequence=2

[6] https://www.nsa.gov/research/tnw/tnw194/articles/article13.shtml

[7] Spinola, S. Addressing the challenges of cybersecurity R&D. 21CT. 2013. http://www.21ct.com/blog/addressing-the-challenges-of-cybersecurity-rd/

[8] Popper, K. R., The Logic of Scientific Discovery. Routledge Classics, 2nd edition 544 pp. 2002.

[9] Kuhn, T. S., The Structure of Scientific Revolutions. The University of Chicago, 4th edition, 264 pp. 2012.

[10] Feyerabend, P. Against Method. 3rd edition. 279 pp. Verso, 1993.

[11] Crawford S and L Stucki. 1990. Peer review and the changing research record. Journal of the American Society for Information Science 41:223–228.

[12] Lohr, K. N., 2004. Rating the strength of scientific evidence: relevance for quality improvement programs, International Journal for Quality in Health Care, vol. 16, no. 1, pp. 9–18, Feb. 2004.

[13] Druckman, J. N., D. P. Green, J. H. Kuklinski, and A. Lupia, Eds., 2011, Cambridge Handbook of experimental political science. Cambridge University Press,

[14] Zar, J. H. Biostatistical Analysis. 5th edition, Prentice Hall. 944 pp. 2010.

[15] Box, G. E. P., J. S. Hunter, and W. G. Hunter. Statistics for Experimenters: Design, Innovaton, and Discovery, 2nd edition. John Wiley and Sons. 2005.

[16] Herbert Edelsbrunner, H. and J. Harer. 2008. Persistent homology — a survey, chapter Surveys on Discrete and Computational Geometry: Twenty Years Later, pages 257–282. American Mathematical Society, Providence, RI.

[17] Mileyko, Y. S. Mukherjee, and J. Harer. 2011. Probability measures on the space of persistence diagrams. Inverse Problems, 27(12):124007, 2011.