Why Is There No Science in Cyber Science?

[A panel discussion at NSPW 2010]

Roy A. Maxion
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213
maxion@cs.cmu.edu

Thomas A. Longstaff
Applied Physics Laboratory
The Johns Hopkins University
Laurel, MD 20723
Thomas.Longstaff@jhuapl.edu

John McHugh
RedJack, LLC
Silver Spring, MD
John.McHugh@cs.unc.edu

ABSTRACT

As researchers with scientific training in fields that depend on experimental results to make progress, we have long been puzzled by the resistance of the experimental computer science community in general, and computer security research in particular, to the use of the methods of experimentation and reporting that are commonplace in most scientific undertakings. To bring our concerns to a broader audience, we proposed a discussion topic for NSPW 2010 that covers the history and practicality of experimental information security with an emphasis on exposing the pros and cons of the application of rigorous scientific experimental methodology in our work. We focused on discussion points that explore the challenges we face as scientists, and we tried to identify a set of concrete steps to resolve the apparent conflict between desire and practice. We hoped that the application of these steps to the papers accepted at NSPW could be an early opportunity to begin a journey toward putting more science into cyber science. The discussion, as expected, was wide ranging, interesting, and often frustrating. This paper is a slight modification of the discussion proposal that was accepted by NSPW with the addition of a brief summary of the discussion.

Categories and Subject Descriptors

A.m. [**General Literature**]: Miscellaneous; K.7.m [**The Computing Profession**]: Codes of Good Practice.

General Terms

Experimentation, Measurement, Reliability, Security

Keywords

Experimentation, Research Methodology, Science of Security

NSPW'10, September 21–23, 2010, Concord, Massachusetts, USA.

The Justification

This is a topic for which no discussion should be necessary. The proposers have backgrounds in areas where experimentation is well established: cognitive science and physics. Several years ago, we became disturbed at the lack of proper scientific methodology in the experimental areas of computer science/security. In 2007, Maxion managed to insert the following language into the call for papers for a workshop devoted to security measurement, Quality of Protection.

> Experimental papers are required (1) to explicitly state the hypothesis being tested or the problem being solved and (2) to have a methodology section. The methodology section should contain enough details that a reader could reproduce the work, at least as a thought-experiment. Where appropriate, this section should include information like: materials, apparatus & stimuli used, a description of the subjects or data sets used, the experimental design, and the procedure followed.
>
> Theoretical papers should succinctly state the hypothesis that results from the theory, and describe an experiment for its validation.

During the following three years, the language remained unchanged and almost completely without effect, because the organizers had to choose between enforcing the language and holding a workshop without papers. Last year, the IFIP WG 10.4 devoted a workshop to the issue, and we were surprised at the resistance from respected researchers in the dependable systems community. There was a consensus that forcing their students (who, after all, do most of the research in their laboratories) to provide even the modest level of rigor implied by the QoP language would slow their pace of publication and thus diminish their job prospects. We meet similar resistance from Government Research Program Managers who view their jobs as fostering technology transfer rather than contributing to a body of scientific knowledge. As reviewers of proposals and papers, we can try to impose higher standards, but we are often overruled by our colleagues. Maxion teaches a course in experimental methods at CMU; but McHugh's attempt to get such a course added to the graduate curriculum at Dalhousie was rebuffed by colleagues bent on reducing requirements in the hopes of graduating more students with less faculty effort.

We believe that addressing this issue is important for both security research and for computer science in general where

we have been unable to provide solid evidence for such assertions as "Development method X reduces coding errors" or "Testing method Y finds more errors than method Z." We thought that the discussion would be lively and hoped that, even if we were "preaching to the choir," it would produce some useful ideas on how to improve the quality of Cyber Security Research.

1. THE TOPIC

As researchers with scientific training in fields that depend on experimental results to make progress (Cognitive Science and Physics), we have long been puzzled by the resistance of the experimental computer science community in general, and computer security researchers in particular, to the use of the methods of experimentation and reporting that are commonplace in most scientific undertakings. It may be a consequence of the bastard origins of the field – an exogamous union of engineering and mathematics. Mathematicians, in general, do not need experimentation to make progress because logic and reasoning show the way, while engineers are often satisfied with a demonstrably useful result rather than a predictively useful one. As a consequence, much of computer science either has sound mathematical underpinnings, but results of limited applicability to broad problems, or anecdotal experiences that are neither repeatable nor foundational and provide a limited basis for a systematic growth in knowledge and understanding. Prediction is the key result that we want from scientific research.

> Science (from the Latin *scientia*, meaning "knowledge"), refers to any systematic knowledge or prescriptive practice that is capable of resulting in a **correct prediction**[1].

Research in Cyber Security can be roughly divided into two areas, theoretical and experimental:

Theoretical Cyber uses pure mathematics or logical syllogisms to demonstrate relationships (examples: crypto strength calculations, lattice theory, information entropy, access control matrix). As in other mathematical research areas, results are highly reusable and can be systematically improved.

Experimental Cyber is far more problematic. It requires the use of the scientific method to demonstrate causal relationships. Once these relationships have been scientifically demonstrated, the relationship can be presumed true and used to build further knowledge.

Current standards for research are appallingly low. Consider a typical Cyber "Experiment," a process on which many a paper (and more than one academic degree) has been based:

1. Have an idea for a "new" tool that would "help" security
2. Program/assemble the tool (the majority of the work)
3. Put it on your local net
4. Attack your system
5. Show the tool repels the attack
6. Write up "the results" and open-source the tool
7. (optional) Start up a company which *might* succeed

An alternate paradigm begins with finding an attack that gets by an existing tool of this sort, and developing an idea for an improved tool. The process then picks up with step 2, above. These processes can be (and are) repeated over and over. This is known as standing on the toes of "Giants" but is a far cry from an application of the Scientific Method.

Have an idea $\neq$ Form hypothesis.
Build & deploy tool, attack system $\neq$ Perform experiment, collect & analyze data.
Show tool repels the attack $\simeq$ interpret data and draw conclusions.
Write up results and open source tool $\simeq$ add results to body of knowledge.

Sometimes we see processes that, on the surface, appear to be an improvement:

1. Form a hypothesis that appears to be falsifiable like "my new mathematical technique will detect attacks previously unseen"
2. Design an experiment using sampled network data, injected with attacks of interest against a tool that implements your design
3. Set up a test lab with background traffic, the system to be tested, and the injected attacks. Add points to collect both network and system data for analysis
4. Run the experiment several times and collect data for analysis
5. Perform ROC analysis on the collected data and use this to determine the performance of the technique
6. Write up a conference paper on the results of the experiment

This may result in a follow-on process, similar to the alternate paradigm discussed above, that begins with contacting the author to get copies of the tool and the data (if it can be made available) used for the experiment, followed by rebuilding the lab to verify the experiment. At this point, you compare the results obtained to some other tool or technique you think might be better, perform a comparative ROC analysis, and publish your new results, referencing the previous conference paper.

Unfortunately, there is still quite a bit that is lacking. The original hypothesis may have been testable and falsifiable (if it was well formed), but the experiment was probably not adequately controlled and the experiment was probably not reproducible. There was little or no quantitative analysis (a hallmark of a physical science!) and more importantly, the hypothesis did not state a causal relationship that could be used to really advance the field. We are still standing on the toes of the previous work (although perhaps with platform shoes).

We need to return to basics. By focusing on the Scientific Method, we may be able to create a body of scientific knowledge in the Cyber world that consists of causal relationships demonstrated through experimentation. This

[1]From Wikipedia (the source of sources!)

knowledge cannot be merely a collection of tools. A good start is the experimental verification of theoretical results – *e.g.*, does the theoretical result actually apply in real situations?

Most existing Cyber observations are not designed to stimulate questions about causality; casual observations such as "I see this attack against my system," "I observe port scans against my system," or "This tool seems to improve security" do not lead directly to usable hypotheses. Before a good hypothesis can be formed, observations must drive the researcher to speculate on causality. Consider the following observations: "It seems like the more traffic I see, the more often I'm attacked," "Attacks against my system seem to degrade performance," or "Encrypting the data in transit does not seem to stop theft of information." Observations of this sort lead towards studies that produce measurements related to cause and effect.

There is an intermediate step between making observations and constructing scientific hypotheses. A good system model must be constructed from observations to lead to scientific hypothesis formation. Theoretical and accepted experimental results can be incorporated in a system model as "knowns." Deliberate building of system models helps to identify potential dependent and independent variables. Only then can you decide on a particular experiment design scheme (factorial, Taguchi, random, etc.), systematically control the independent variables, and observe the effects. With a system model in hand, we can form useful hypotheses that clearly state an expected causal relationship:

- It must be falsifiable, *i.e.*, it must be able to be disproved through the defined experiment
- Variables in the experiment must be controlled, *i.e.*, only variations that confirm or deny the causal relationship should be allowed in the experiment
- Results from the experiment must be reproducible, *i.e.*, external parties should be able to completely replicate the conditions of the experiment to achieve the same result

Now we can see the weaknesses in the "improved" experimental process. The hypothesis could not be disproved, the environment had many more variations than the one causal relationship we are investigating, and, since natural (uncontrolled) data were used, the experiment is not repeatable.

Even after forming a good hypothesis based on causal observations, there are significant challenges in Cyber experimentation, some of which are controlling the Cyber environment for variations in the causal factor alone, creating representative data for the experiment, and collecting (and maintaining) data that will confirm or refute the hypothesis, and documenting the experiment such that it can be precisely replicated. (See Richard Feynman's comments on reproducibility in his 1974 Cal Tech commencement address, Cargo Cult Science.)

The data are extremely important. We have a long and sorry history of using synthetic data for research in intrusion detection. This has been largely a failure, because intrusion detection devices operate on relationships among data characteristics, and the data generators don't manifest all of the data's characteristics: *e.g.*, transition probability, sequences, symbol frequency, gap consistency and composition, etc. Current practice is blind to the multiple characteristics of the data, and test data may in fact be generated in such a way as to correctly represent variety in some characteristics, but not others. If the generator and detector under test are not dimensionally matched, any test result is without merit. Even if they are matched, but the data are not representative of the real world, the results may not be useful.

Analysis of data from Cyber experiments is especially problematic. Events measured in Cyber experiments are rarely statistically independent, rendering many analysis techniques flawed. Cyber data tends to be categorical rather than continuous, making many traditional distance and other metrics problematic. The measurement apparatus rarely has precise performance characteristics (*e.g.*, error rates), so results of data collection cannot be assigned precision or accuracy. Many results of Cyber experimentation are a combination of human-centric and system-centric elements, making valid analysis techniques difficult to select and apply.

Once an analysis has been performed, it must be used to confirm or refute the hypothesis. The context of the conclusion may be restricted to the conditions of the test alone and not be generalizable. Experimental data may not have contained sufficient variation on the controlled variable to generalize the results beyond the selected dataset. Unexpected variation beyond those of the controlled variable may bias the results (*e.g.*, other processes running on the experimental data capture equipment may introduce noise into the measurements).

Once research hypotheses are confirmed, they must be added to the overall body of knowledge to be reused. Unfortunately, Cyber as a research discipline is not well structured. Many, many taxonomies and ontologies exist for Cyber vulnerabilities alone. There are no "natural laws" to structure results, and necessary and sufficient conditions applying to a given result are elusive. Adding to the existing body of knowledge in Cyber consists of creating a linear and mostly unconnected list of "facts." Direct connections between theoretical results and empirical results are rarely captured.

All of these issues make it very difficult to take experimental Cyber research at face value. Thus, we do not build on the results of the experimental research, but instead reproduce the results in a slightly different environment without taking the differences into account. Lack of structure on the body of knowledge leads to a LRU policy for research results *i.e.*, "If Google doesn't know it, it never happened." The few fundamental results we do have are mostly unapplied (*e.g.*, the reference monitor for assured policy enforcement).

What has to happen to improve the situation? We need to leverage the knowledge of physical scientists and Cyber scientists with "clue." This starts with better and deeper education; computer science students are frequently not trained to use the scientific method, and they usually lack skills in experiment design, data collection, statistics and data analysis. Conferences and Journals must promote (or even require) the use of the scientific method as a main acceptance criterion. Papers must demonstrate scientific rigor in their reporting of experimental activities and results. The "Body of Knowledge" must be structured and used to guide future work. Good data should be generated and cherished and shared. There needs to be an explicit separation between

scientific contributions and technological contributions; scientific contributions must be rewarded.

Roger Schell, in his 2001 essay "Information Security: The State of Science, Pseudoscience, and Flying Pigs," says,

> The state of the science of information security is astonishingly rich with solutions and tools to incrementally and selectively solve the hard problems. In contrast, the state of the actual application of science, and the general knowledge and understanding of the existing science, is lamentably poor. Still we face a dramatically growing dependence on information technology, *e.g.*, the Internet, that attracts a steadily emerging threat of well-planned, coordinated hostile attacks. A series of hard-won scientific advances gives us the ability to field systems having verifiable protection and an understanding of how to powerfully leverage verifiable protection to meet pressing system security needs. Yet, we as a community lack the discipline, tenacity and will to do the hard work to effectively deploy such systems. Instead, we pursue pseudoscience and flying pigs. In summary, the state of the science in computer and network security is strong, but it suffers unconscionable neglect.

It has been nearly a decade since these words were written. If anything, the situation is worse today than it was then. We hope that this discussion will aid us in moving forward.

2. DISCUSSION POINTS

The following are discussion points that we expected the panel and the audience to address.

1. Given the small minority of practitioners able to apply scientific method effectively in security research, how can the majority learn to see the benefits of this type of research?

2. How can the general history of scientific research (*e.g.*, the eventual acceptance of natural laws verified by experiment in other fields) provide a way forward towards mature security research?

3. Is security really not a scientific discipline at all, but rather a category of engineering technology? What, then, is the science on which this technology is based?

4. Can there be a separation of security experimentalists and technology producers, or are the two fundamentally intertwined?

5. What are the financial and incidental incentives to produce scientific results in security? Is the only path a tool that you can market?

3. THE DISCUSSION

After some negotiation with the NSPW organizers, the panel was split into two sessions. The first session took place at the beginning of the workshop; the three panelists made brief opening statements, and the floor was opened to discussion. One outcome of this portion of the discussion was a realization that using the physical sciences as an exemplar of a discipline in which the scientific method is well established is probably counterproductive. A substantial amount of time was spent on this topic and its consequences. Better models are probably the social sciences which deal with the same sorts of complex and uncontrollable environments as cyber security, but nonetheless have well established methodologies for performing experiments, analyzing data and reporting results so that others may build upon them. At the end of this session, the panel challenged the audience to consider several things as they participated in the remainder of the workshop:

1. Do the speculative papers presented have potentially testable hypotheses?

2. For papers that have "results", are those results reproducible? (Reproducible does not mean merely repeatable, a significant difference!)

3. Did this workshop do a good job at scientifically evaluating the papers presented at the conference?

4. Can we as a community recognize the above qualities?

The first session ended with a discussion concerning whether security was a sufficiently mature discipline (an often-heard opinion) to support a scientific inquiry and the comment from one of the panelists (Maxion):

> What we as a panel are trying to understand is whether lack of "science" is a good or a bad thing, and why might science not be needed now? But if not now, when; and if now, how can we fix the field?

For a variety of reasons, we were unable to focus the group's attention on the questions raised above, and the initial panel tried to redirect the discussion for the second panel session which closed the workshop. The original panel replaced itself with three new panelists: Richard Ford of the Florida Institute of Technology, Carrie Gates of CA Labs, and Lizzie Coles-Kemp of Royal Holloway. The session was moderated by Longstaff, and began with introductory remarks from the new panel. Ford made the point that science is not about boring experiments, but rather moving from opinion to knowledge, and that doing better science does not imply being less productive. Gates took a deliberately contrarian view, claiming that from an industry standpoint, rapid innovation was far more important to the bottom line than scientific advancement. Coles-Kemp tried to frame the issue in terms of her experiences involving the way that the social research community works with the mathematical research community, particularly with regard to the need to respect the duality of physical and social objects, and the need to consider the meaning of scientific methods with respect to a given new paradigm.

The discussion provoked by these statements was wide ranging and pointed. A number of participants suggested that the poor performance in the field of a number of security technologies, anti-virus detection in particular, was due to the short-term focus of industry in this area. The need for data and evidence to support marketing claims was also mentioned. As the discussion drew to a close, Longstaff asked for concrete suggestions for instilling some rigor in the

field. Several people suggested steps to improve the paper-reviewing process, including requiring an explicit methodology section in experimental papers, something that is standard practice in other disciplines. This might be accompanied by a rebuttal period in which authors have a chance to clarify the description in their paper. It was also suggested that both experimental code and data should be made available as a matter of course, with authors required to explain if this cannot be done.

Finally, it was noted that there really should be no tension between Ford and Gates since you can apply a systematic approach whether you seek truth or market share, quoting Bacon "Truth will sooner come out from error than from confusion."

"Science is your friend," as Richard Ford (Program Committee Co-Chair) noted, and contrary to being an enemy of innovation, "science enables innovation by giving us the tools to see where our current understanding is insufficient. Real knowledge in only generated when we confirm each other's understanding of the world."

4. THE PANEL

Roy Maxion is a Research Professor in the Computer Science Department at Carnegie Mellon University. His research covers several areas of computer science, including development and evaluation of highly reliable systems, machine-based learning, and human-computer interfaces. He has worked in intrusion detection, insider/masquerader detection, and two-factor authentication using keystroke dynamics. He has been a member of the dependable systems community since 1984, and is a Fellow of the IEEE.

Tom Longstaff is the Chief Scientist of the Cyber Missions Branch in the Applied Information Science Department of the Applied Physics Laboratory (APL) and is the Program Chair for the Computer Science, Information Assurance, and Information Systems Engineering within the Whiting School of Engineering at The Johns Hopkins University. APL is a University Affiliated Research Center, a division of the Johns Hopkins University founded in 1942 and located in Laurel, MD. Tom joined APL in 2007 to work with a wide variety of infocentric operations projects on behalf of the US Government to include information assurance, intelligence, and global information networks. Prior to coming to APL, Tom was the deputy director for technology for the CERT at Carnegie Mellon University's Software Engineering Institute.

John McHugh is the Senior Principal at RedJack, LLC and previously held the Canada Research Chair in Privacy and Security at Dalhousie University in Halifax, NS. Prior to that, he served as a Senior Member of the Technical Staff at the CERT at Carnegie Mellon University's Software Engineering Institute, as a Tektronix professor in the Computer Science Department at Portland State University in Oregon, and as a Vice President of Computational Logic, Inc. He has performed research is covert channel analysis, process models for building trusted systems, and intrusion detection. He is the author of one of the few critical reviews that has been published in the Cyber Security area. His recent research is in the area of large scale network data analysis.

5. ACKNOWLEDGMENTS

The first author was supported by National Science Foundation grant number CNS-0716677. The authors are grateful to Holly Hosmer for strengthening our resolve, and to Richard Ford for enabling it.