Gamification for Measuring Cyber Security Situational Awareness

Glenn Fink[1], Daniel Best[1], David Manz[1], Viatcheslav Popovsky[2], and Barbara Endicott-Popovsky[3]

[1] Pacific Northwest National Laboratory, Richland, Washington, USA
[2] University of Idaho, Moscow, Idaho, USA
[3] University of Washington, Seattle, Washington, USA
{glenn.fink,daniel.best,david.manz}@pnl.gov,
dr_popovsky@hotmail.com, endicott@uw.edu

Abstract. Cyber defense competitions arising from U.S. service academy exercises, offer a platform for collecting data that can inform research that ranges from characterizing the ideal cyber warrior to describing behaviors during certain challenging cyber defense situations. This knowledge could lead to better preparation of cyber defenders in both military and civilian settings. This paper describes how one regional competition, the PRCCDC, a participant in the national CCDC program, conducted proof of concept experimentation to collect data during the annual competition for later analysis. The intent is to create an ongoing research agenda that expands on this current work and incorporates augmented cognition and gamification methods for measuring cybersecurity situational awareness under the stress of cyber attack.

Keywords: Cyber Defense Competitions, CCDC, cyber defender, cyberwarrior.

1 Introduction

The Pacific Rim Collegiate Cyber Defense Competition (PRCCDC) represents a unique opportunity for observational experiments. While there are many types of observational experiments, in computer security they mostly fall into two classes: laboratory experiments and field studies. Laboratory experiments can be highly controlled and enable researchers to test a hypothesis and quantify the contribution of each of several factors with confidence. With good experimental design, the results may be generalized safely. Unfortunately, the very controls required to obtain certainty cause results to be much less realistic, and potentially less relevant to real life. In contrast, field studies are used in situations where interesting behavior is to be observed, but it is impractical to compare a control group to an experimental group. In field studies, data collected can be highly relevant to real life, but the power of the conclusions that we can draw from these observations is greatly limited because of high variability and contamination from uncontrolled factors. Field studies are typically difficult to replicate, and results may be hard to quantify and merely anecdotal.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 656–665, 2013.

These researchers believe that the PRCCDC, and similar competitions, represent a venue for conducting experiments that are a hybrid of laboratory experiments and field studies. The nature of the competition introduces constraints that (with care) can be adopted as experimental controls while the range of activities available to measure are nearly as unlimited as those that happen in the real world. And possibly just as importantly, the data that can be collected could be published, shared, and reused much more easily without destructive anonymization, unlike that collected in real-world situations. Further, gamification methodologies can be applied that can expand on the purely observational experimentation described in this proof of concept.

2 History of the Collegiate Cyber Defense Competitions

Cyber defense competitions arose out of a military educational requirement for the U.S. service academies [1]. The competition was fierce and the result was so successful that civilian universities began to follow suit. Beginning in 2004, the US Military Academy at West Point adapted their 'capture the flag' exercise to a civilian scenario and introduced the competition at several universities across the country, including the University of Washington which incorporated the event into the Information Assurance and Cybersecurity Certificate program as an annual capstone experience. On February 27 and 28, 2004, a group of educators, students, government and industry representatives gathered in San Antonio, Texas, to discuss the feasibility and desirability of establishing a post-secondary level, national program for cyber security exercises. The outcome of these discussions was 1) a competition architecture with a clear set of rules and roles, 2) a fair and impartial scoring system that provides a level playing field for competitors, 3) an IT infrastructure designed to eliminate possible advantages due to hardware and bandwidth differences at different regional locations, and 4) resolution of possible legal concerns.

The resulting Collegiate Cyber Defense Competition (CCDC) system provides institutions teaching information assurance or computer security a controlled, competitive environment that can assess students' depth of understanding and operational competency in managing and protecting a corporate network [2]. The CCDC helps participating institutions of higher education evaluate their educational programs, provides an educational venue for students to apply the theory and practical skills they learn in their course work, fosters teamwork and ethical behavior, and creates interest and awareness among participating institutions and students. In 2006, the University of Texas at San Antonio agreed to host the first national CCDC. In 2007, the University of Washington opened up their internal competition to outside institutions, establishing the regional PRCCDC as an entrant into the national competition. 2013 is the sixth year of PRCCDC participation in Nationals. There are now ten regional venues: At-Large (virtual) Regional, Mid-Atlantic, Midwest, North Central, Northeast, Pacific Rim, Rocky Mountain, Southeast, Southwest, Western.

During competition, 8-10 student teams comprised of eight students each defend identical networks. The competition lasts 2-3 days. Teams are scored based on ability

to protect and defend against outside threats, maintain availability of web services, respond to business requests, and balance security needs against business needs.

A Red Team of external attackers, often professional penetration testers from local industry, relentlessly attack student networks throughout the competition. Students are expected to resist attack, or recognize and recover from attack, if penetrated. A White Team of judges—in the case of the PRCCDC a team of graduate students from Idaho State University's NIATEC program—issue a series of 'injects,' or administrative chores, that must be accomplished in an orderly and timely fashion in the face of attack. The entire process is designed to simulate the stress and intensity of managing networks in today's hostile Internet environment. These CCDC exercises employ controls designed to preserve fairness and safety among teams from participating schools. These same controls may be used as the foundation of high-quality experimental controls as long as fairness and safety are preserved. For instance, each team begins with a small, pre-configured, operational network they must secure and maintain located on a dedicated internal network. This also allows tight control over competition traffic. Each team is given the same set of business objectives and injects at the same time during the course of the competition.

Each student team is composed mostly of undergraduates, although two at most could be graduate students. No professionals are allowed, and the students may not be currently employed in an IT industry job. Students must be enrolled in a minimum number of class hours to qualify. Faculty advisors are not allowed to be with the team during competition. These restrictions double as experimental controls. The White Team enforces the competition's controls and employs an automated scoring engine that periodically tests availability and function of each student team service and network component during the competition. They also administer and grade responses to injects. Allowing only students and White Team members inside competition rooms eliminates potential variability from the influence of coaches. Running scores are not announced during the competition, eliminating potential stress factors.

The Red Team is the aggressor seeking to disrupt services and business objectives of the student teams. They are non-biased, commercially experienced, and comprised of volunteers. Loose controls are placed on Red Team activities that enforce objectives of fairness and safety. Within these controls, Red Team members employ any attack techniques at their disposal, including non-cyber attacks like social engineering. After the competition, the Red Team usually provides feedback to the student teams on their defenses and how the Red Team attacked them.

3 Data Collection

In this paper, the authors discuss how data that described the effectiveness of collaboration was collected at the PRCCDC. Future studies will include injecting collaboration-enhancing technologies to show the effectiveness of these treatments and augmented cognition methodologies designed to measure participant biological reactions to stress. Data collected was analyzed in a separate publication [3]. In this paper, we discuss experience gained in collecting the data to show the effort required, as well as the benefits this data will be to future studies. An observational experiment

was designed to collect baseline (control) information on collaborative practices in cyber security teams. Collecting full packet traces is common practice at these competitions, but it was felt much more data was needed to tell the stories behind the collaborative interactions that the competition fostered. This section discusses each of the kinds of data collected and how it was collected. During the competition, the following was gathered:

1. Data from the team scoring process,
2. Situational awareness data from team members,
3. Network packets and machine log files,
4. Video and audio of the competition,
5. Stress resilience characteristics of one of the teams.

3.1 Performance Data Capture

Having well defined and fair performance scoring built into the CCDC makes it an excellent source of regular data with a ground truth. Performance and timing data were gathered from the teams' execution of business requirements (injects) that were delivered by email as part of the competition. A HotMail web client was used to record the time when an email instruction was received, opened, and replied to. This timing data was integrated with situational awareness data discussed below. Scoring data gathered included evaluation rubrics for each inject (twenty per team) that guided scoring of student team performance when executing each inject. Computation was done by White Team volunteers and is somewhat subjective. Scoring data was also generated for each successful attack levied against the student teams. Whenever the Red Team infiltrated a student machine successfully, that student team lost points. If the attacked team filed a detailed incident report, they would salvage some portion of their loss. These incident reports helped assess collaborative behavior. Final scores accumulated by each team were gathered from the White Team as an ultimate measure of success. This scoring was partly objective, partly subjective. The subjective part came from humans grading the "goodness" of inject response. The more objective source of data came from the scoring engine which periodically tests the state of all the services teams must maintain. The scoring engine results provided an important source of ground truth when assessing situational awareness.

3.2 Situational Awareness Data Capture

Team situational awareness was measured as a way to infer team performance independently from the competition performance scoring. Researchers, armed with digital audio recorders, were assigned to occasionally ask situational awareness questions of student and Red Team members. Timing and accuracy data were used from their responses and from the injects to conduct an assessment of team situational awareness using Durso's Situation Present Assessment Method (SPAM) [4].

The Questions. The questions used for assessing situational awareness were binary choices (yes/no, A/B) designed to assess the team's cognition of their situation without interrupting their tasks. Reducing interruptions was one of the reasons Durso's model was chosen over interruption-based protocols like Endsley's Situation Awareness Global Assessment Technique (SAGAT) [5]. Additionally, the research team kept questions simple to answer using known, ready-to-hand, materials.

There were seven student teams and one Red Team in the competition. Four researchers gathered data. Each student team was queried every 20 minutes. The Red Team was also queried periodically, but the objective here was to inform the questions of the research team rather than to measure situational awareness. One researcher stayed with the Red Team, the remainder queried student teams. A question matrix was designed for the student teams with one-third of the questions, each, concentrating on concerns of the past 20 minutes, the present, or future 20 minutes, respectively. Durso's work shows that future-oriented questions were most indicative of expertise, so the tense of a question was controlled carefully. The following taxonomic breakdown of question types was used:

1. Defense-related
   a. Policies: What defensive actions should happen?
   b. Priorities: What defensive actions are most important?
   c. Events: What defensive actions were taken?
   d. Causes: What caused or would cause defensive action X?
2. Threat-related
   a. Policies: What offensive actions should happen?
   b. Priorities: From an attacker's view, what is the most important action?
   c. Events: What offensive actions happened or will happen?
   d. Causes: What caused or would cause attackers to take offensive action X?

From this taxonomy, a list of 48 questions was generated. The research team met approximately every 20 minutes and randomly selected one of these questions and applied it to the current situation, filling in information as needed. For example, one question was, "Do you expect your X service to be a likely attack vector in the next 20 minutes?" Before using this question, researchers had to replace X with the name of a service (e.g., email, web, ftp, etc.) thought most fitting at the.It was important to administer the same question to all the teams so as not tip a team off and provide an advantage. For instance, asking whether or not a team had changed the default router password might inform them that they should do this when they had not known to do so on their own.

The Querying Protocol. Each researcher was given the task of querying 2-3 student teams, selected at random, during the remainder of the 20-minute segment. Researchers were instructed to try to approach a team member they had not approached. This induced as much variability into picking the subject as possible. Some teams chose a spokesperson to handle all queries. In that case, the researcher noted the policy and always approached the spokesperson, honoring the team's wishes. The intent is to infer *team* situational awareness from these queries.

To ask a question, a researcher would approach a participant and place a green question card face down on the table in plain view. He then would start his audio recorder and say, "Excuse me, I have a question when you are available." When the participant was ready to answer, he or she would turn over the question card, and the researcher would ask him/her to read the question aloud and answer it. The audio recorder was left running from the initial "excuse me" until either the participant finished answering or five minutes of silence elapsed. At a maximum time of five minutes, the researcher would stop the recorder, pick up the question card, and move on.

The Analysis Plan. Durso's method [4] was employed to measure both situational awareness and workload. In Durso, the time from when the researcher says, "excuse me" until the interviewee reads the question is a measure of workload. Similarly, the time from when the question is read to when the participant answers is a measure of situational awareness. Durso never made any claims about this method working for assessment of team situational awareness. The authors believe effectiveness can be inferred from individual situational awareness, but teams are another matter. There are also difficulties that arise because the venue is not a tightly controlled experiment. Many uncontrolled distractions are happening in a competition whose effects may be larger than the situational awareness effects being measured. This lack of control is inherent to the venue, but the authors believe that quality data can be extracted and generalized, keeping in mind these limitations. Increased statistical power is gained by running the experiment simultaneously on seven teams; however, it must be kept in mind that these are not and cannot be true replications because they are different subjects and are not truly independent.

3.3 Network and Log File Collection

To provide the most information available about network activity, full packet traffic was captured in several key locations. The network topology consisted of a core router connecting all teams to the scoring server and the Red Team (see Figure 1).

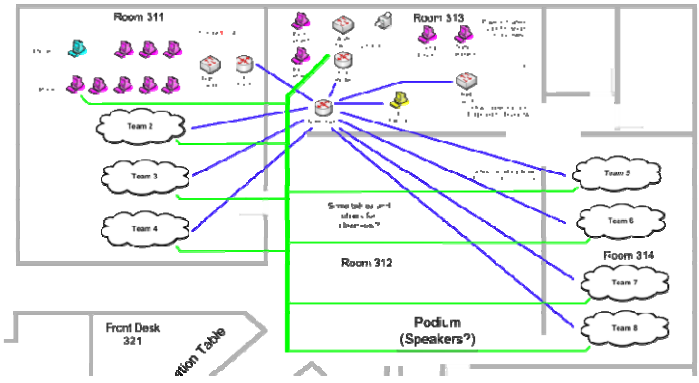

Fig. 1. Competition Network

Connected to the core router, each team's router defined the team's local network. Because Red Team activity could disable a team's router, there was no guarantee that each team's traffic would always reach the core router throughout the event. The aim was to gather as much data from the network, given configuration limitations.

To be as unobtrusive as possible, the core router and team routers were configured to mirror a set of ports to an available port (the "span port"). The associated network interface controller (NIC) of packet-capture laptops connected to the span port were configured to **not** have an IP address—making them essentially invisible. *tcpdump* was configured to capture full packets (headers and all data) by setting the *snaplen (-s)* parameter to 0 (no size limitation). Packet data was output to files of 100 million bytes (-C 100) to ease processing later. A startup script was installed to initialize *tcpdump* and ensure existing packet capture files were not overwritten when the program started. Each machine ran 32-bit Ubuntu Server 9.10 OS, configured with no optional services, in order to minimize attack vectors.

The core router was configured to capture packet data, and because of resource limitations, only three other packet-capture machines were provided on other routers. To allow for possible correlation of network data with captured video, the router of the single team who agreed to be filmed during the competition was one of those. Other packet capture locations were some of the other student teams, the Red Team, and machines teams used to access the Internet for patch downloads. After the event, log data was harvested from all available machines.

3.4 Video and Audio Data Capture

In addition to performance, situational awareness, and network data, video and audio were captured from the competition. City University of Seattle filmed the entire event and provided access to their raw footage. This footage was particularly useful to record the Red team's brief-back at the end of the competition; however, during the body of the competition, the coverage was too uneven as a reliable data source. Not all teams consented to recording which would have been prohibitively expensive in both equipment and time to analyze, so resources were concentrated on the one team from the UW iSchool which graciously agreed to allow video and audio capture.

Eight Logitech 600 webcams were placed strategically within the iSchool team's area to capture interactions and collaboration among participants. The cameras were pointed across the table to capture several subjects at once, allowing a clearer view of team interactions. The team sat in two circular pods with cameras mounted to the table and tops of equipment, facing back across the tables. Camera orientation was periodically checked to make sure they were still aimed correctly.

A single workstation streamed video from all eight webcams using the Logitech camera software and Debut video capture software to capture multiple streams, simultaneously. Eight simultaneous streams of 15fps video were captured at 1280x1024 pixel resolution. While not high quality, this was sufficient to identify whether people were collaborating and a little about their gestures and activities. Since webcams were unable to record clear audio, extra voice recorders were used on each table. During

analysis, a single audio track was used to simplify reviewing the video. To facilitate time synchronization, a sync signal was used to start recording and periodically throughout the competition: a researcher clapped his hands in front of the camera.

3.5 Stress Resilience Characteristics

The student team filmed also consented to being tested, individually, prior to the competition. This was done in order to characterize their psycho-physiological profile as an indicator of their nervous system type. Four tests were given that measured stress resilience, the ability to context-switch, and the ability to maintain balance in their psychological processes under stress. Results led to individualized profiles that, in a business setting, could be useful in managing performance.

This suite of tests was developed by E.P. Ilyin and has proven effective in assessing a subject's ability to handle stress in a variety of occupational settings for particular professions [6,7,8,9,10,11]. Application of this methodology has been helpful in optimizing individual performance in a range of competitive professional environments, including world class sports venues. The authors are adapting this approach to cyber defense competitions. It is believed it could have relevance for developing profiles of effective cyberwarriors, as well as stratagems for identifying and preventing burnout of cyberdefenders stressed by managing networks under constant attack.

3.6 Dry Run

Two dry runs of the data collection technology were conducted to determine feasibility. There were multiple area dependencies where data collection could be derailed. Although some data was lost, the research team was satisfied that a great quantity of useful data was captured. Due to equipment costs and space constraints, the researchers were unable to provide much duplication of collection.

4 Potential Uses of the PRCCDC Data

This data is a "gold mine" of potential research benefits. First, obtaining a realistic set of network data that does not have to be anonymized meets a crying need of the cyber security research community. (In previous research, unavailability of strong anonymization techniques was an important reason why organizations did not share their cyber data and learn from one another's mistakes [12]). Further, research groups at PNNL have long expressed interest in a data set where cyber and video data could be correlated to evaluation of levels of fatigue and stress related to cyber operator error. These authors anticipate using this data to evaluate key characteristics of effective cyber defense teams and individuals. It is expected that the team will return to this data set, again and again, as research matures.

5 Hindrances in Using PRCCDC as a Data Collection Venue

There are some problems discovered in using PRCCDC events as data sources. This is a high stress venue that allows students to impress potential employers and earn a berth to compete at the national CCDC in San Antonio, Texas. Some participants might feel some anxiety knowing that they are being monitored during the competition and not perform optimally.

Since these events are competitions in their own right, not simply experiments, the research team was constrained by the official competition rules. Additionally, the researchers were constrained to ensure that they did not disadvantage, or advantage, any single team by introducing a treatment.

While extremely helpful, those who set up and ran the competition had other jobs and priorities, making it difficult to impose the rigor needed to collect quality data when it impacted people who were not given any incentive to help. Despite these hindrances, the PRCCDC and similar CCDC events remain extremely valuable sources of data.

6 Future Work and Conclusions

This was a pilot study that provided a baseline for future work. The authors plan to interpose collaborative enhancement technology such as Vulcan, designed to improve analyst performance across competing teams, taking care not to (dis)advantage any team. Additionally, different interview techniques and different methods of query delivery and notification are planned to measure the effectiveness of collaboration. Further, semi-structured interviews, or other data sources such as physiological stress measurements, could be introduced to enrich the data set, facilitating the development of a useful profile of an effective cyber warrior.

The contributions of data collection and experimentation with this current work are:

6. Made available a source of de-identified cyber data for publication and sharing.
7. Put forth data-collection practices that may contribute toward a future standard.
8. Identified a new venue for profitable data collection.
9. Contributed towards better quality scientific methods in cyber security research.

These efforts will help researchers for years to come. Benefits of this study are expected to accrue to cyber security workers and researchers into the future.

Acknowledgements. This work was supported by the I4 Initiative of the Pacific Northwest National Laboratory, Richland, WA, managed for the US Department of Energy by Battelle Memorial Institute under Contract DEAC05-76RL01830. The authors wish to thank the organizations who sponsored the PRCCDC, without whom this event would not have been possible: the University of Washington Center for

Information Assurance and Cybersecurity, Idaho State University, Highline and Whatcom Community Colleges, DeVry University, Black Hat, The Boeing Company, Cisco, and Microsoft—in addition to tireless volunteers and student teams from all of the participating schools.

References

1. Schepens, W., James, J.: Architecture of a cyber defense competition. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 4300–4305. IEEE Press, New York (2003)
2. National Collegiate Cyber Defense Competition, http://www.nationalccdc.org/
3. Malviya, A., Fink, G., Sego, L., Endicott-Popovsky, B.: Situational awareness as a measure of performance in cyber security collaborative work. In: IEEE 8th International Conference on Information Technology, pp. 937–942. IEEE Press, New York (2011)
4. Durso, F., Dattel, A.: SPAM: The real-time assessment of SA. In: Banbury, S., Tremblay, S. (eds.) A Cognitive Approach to Situational Awareness, pp. 137–154. Ashgate Publishing, Burlington (2004)
5. Endsley, M.: A methodology for the objective measurement of pilot situational awareness. In: AGARD Symposium on Situational Awareness in Aerospace Operations, pp. 1–9. Neuilly Sur Seine, France ((1989)
6. Il'in, E.P.: Strength of nervous system and methods for this research. In: Psycho-Physiological Fundamentals of Physical Education and Sports, pp. 5–9. Leningrad (1972)
7. Il'in, E.P.: Instant-method for defining the degree of expressiveness of mobility/rigidity of acceleration/deceleration. In: Psycho-Physiological Fundamentals of Physical Education and Sports, pp. 16–21. Leningrad (1972)
8. Dimitrov, A., Popovsky, V.: Influence of several psychological indexes in the performance of fundamental game moves. Coaching Thought, No. 12 Bulgaria (1987)
9. Kuramshin, U., Popovsky, V.: Find your talent. Leningrad, Lenizdat (1987)
10. Kuramshin, U., Popovsky, V.: Prediction of sports abilities in a system of sports orientation for children and adolescents at their residences. Lesgaft University, Leningrad (1985)
11. Popovsky, V., Endicott-Popovsky, B.: Physical culture pedagogical system. In: III International Congress: People, Sport and Health, St. Petersburg, pp. 19–21 (2007)
12. Fink, G., McKinnon, A., Clements, S., Frincke, D.: Tensions in collaborative cyber security and how they affect incident detection and response. In: Seigneur, S., Slagell, A. (eds.) Collaborative Computer Security and Trust Management, pp. 34–63. Hershey, IGI Global (2009)