2011 Eighth International Conference on Information Technology: New Generations

Situational Awareness as a Measure of Performance in Cyber Security Collaborative Work

Ashish Malviya, Glenn A. Fink, Ph.D., and Landon Sego, Ph.D.
National Security Directorate
Pacific Northwest National Laboratory
{Ashish.Malviya|Glenn.Fink|Landon.Sego}@pnl.gov

Barbara Endicott-Popovsky, Ph.D.,
Information School and UWIT Tacoma,
University of Washington
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 in turn could lead to better preparation of cyber defenders in both military and civilian settings. We conducted proof-of-concept experimentation to collect data during the Pacific-rim Regional Collegiate Cyber Defense Competition (PRCCDC) and analyzed it to study the behavior of cyber defenders. We propose that situational awareness predicts performance of cyber security professionals, and in this paper we focus on our collection and analysis of competition data to determine whether it supports our hypothesis. In addition to normal cyber data, we collected situational awareness and workload data and compared it against the performance of cyber defenders as indicated by their competition score. We conclude that there is a weak correlation between our measure of situational awareness and performance. We hope to refine and exploit this correlation in further studies.**

Keywords-Cyber Defense Competitions; CCDC; cyber defender; cyberwarrior; situational awareness; situation present assessment method; SPAM

I. INTRODUCTION

Cyber security is an essentially collaborative activity, but collaboration is stifled by the sensitivity of the data cyber analysts sift through. Our hypothesis is that computer-mediated collaborative technologies that honor the sensitivity of cyber data can help cyber security professionals keep their systems and networks safer without compromising sensitive data. But to determine whether a particular collaborative technology improves performance, we must first be able to measure performance. We posit that situational awareness, which can be independently measured, may be a predictor of performance in cyber security just as it has been shown to be in other disciplines [Durso, Endsley, etc.] [1] [2].

The Collegiate Cyber Defense Competitions (CCDCs) present a unique venue, midway between controlled laboratory experiments and situated studies, for observational experiments in cyber security.

Laboratory experiments can be highly controlled, enabling researchers to test an hypothesis and quantify the contribution of each of several factors with confidence. Unfortunately, they can test only small features of larger processes, making their results, while generalizable, far less relevant to real life. In contrast, situated studies can be highly relevant to real life, but the generality of their conclusions is greatly limited because of high variability and contamination from uncontrolled factors. The results are typically impossible to replicate, and may be hard to quantify or merely anecdotal. But CCDC competitions provide an objective score and introduce constraints that may serve as control measures for experiments. The range of activities at these competitions is very realistic, but unlike with real-world studies, the collected data can be published, shared, and reused without destructive anonymization.

In our study, we instrumented the Pacific-rim Regional CCDC (PRCCDC) to measure indicators of situational awareness and compared them to team scores. In this paper, we discuss the data collection, our situational awareness methodology, and the analysis of the results from the PRCCDC.

The competition employed a team structure with a neutral administration and exercise-control team (the White Team) in the lead. The White Team was responsible for running the competition, scoring, enforcing the rules, and making policy decisions. A team of cyber penetration testers, the Red Team, was recruited to attack the student teams and attempt to disrupt the services they were tasked to protect. Both the White and Red teams were comprised of volunteers.

The remaining seven teams were competitors comprised of two to eight students. These Blue Teams each had a faculty advisor who was present at the exercise, but not with the team during the competition portion. The teams were primarily full-time undergraduates who were allowed to have only a limited amount of professional experience in system administration or cyber security. Teams were allowed to have one or two graduate members, but the same experience restrictions applied.

II. DATA COLLECTION

Our principal objective for this study was to determine whether situational awareness of team members participating in the competition could predict the overall team's score. During the competition, the following was gathered:

1. Data from the team scoring process
2. Network packets, e-mail records, and machine logs
3. Video and audio of the competition.
4. Situational awareness data from team members

978-0-7695-4367-3/11 $26.00 © 2011 IEEE
DOI 10.1109/ITNG.2011.161

Hindrances encountered in the data collection processes will be discussed in a later section. Data gathered from audio, video, network, log files and situational awareness queries were later analyzed at Pacific Northwest National Laboratory to determine what happened during the competition, to estimate overall team workload and situational awareness, and to determine whether our situational awareness metrics predicted performance as indicated by team scores.

A. *Performance Data Capture*

Performance and timing data were gathered from the teams' execution of tasks arising from simulated emerging 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.

Scoring data gathered included evaluation rubrics for each inject (twenty per team) and the output of the competition's automated scoring engine. These rubrics guided scoring of student team performance when executing each inject. Computation was done by White Team volunteers and is somewhat subjective. Scoring data were 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 that were recorded in the rubrics. If the attacked team filed a detailed incident report, they would salvage some portion of their loss. Additionally, these incident reports helped reveal collaborative behavior stemming from intrusions detected by the teams.

The automated scoring engine periodically tests the state of all the services the student teams are supposed to maintain. For instance, it may send an email to one of the fictitious users that the student team is supporting. Then it will check the inbox of this user to see if the mail server is working properly. The scoring engine results provided an important source of ground truth when assessing situational awareness, and were combined with the inject results to produce the final scores.

B. *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 that connects all teams to the scoring server and the Red Team (see Figure 1). Connected to the core router, each team's router defined the team's local network. Because Red Team activity could disable a student 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 and machines, given resource and configuration limitations.

To be as unobtrusive for packet capture as possible, the core router and Team 7's router were selected and configured to mirror a set of ports to an available port (called the "span port"). A packet-capture laptop computer was then connected to the span port and the associated network interface controller (NIC) was configured to not have an Internet protocol (IP) address. The lack of an IP address made the packet-capture computer essentially invisible to the other users of the network. Upon startup of each capture machine, the NIC was activated, and the *tcpdump* packet-capture program was initiated.

The core router was already configured to capture packet data, and because of resource limitations, only three other

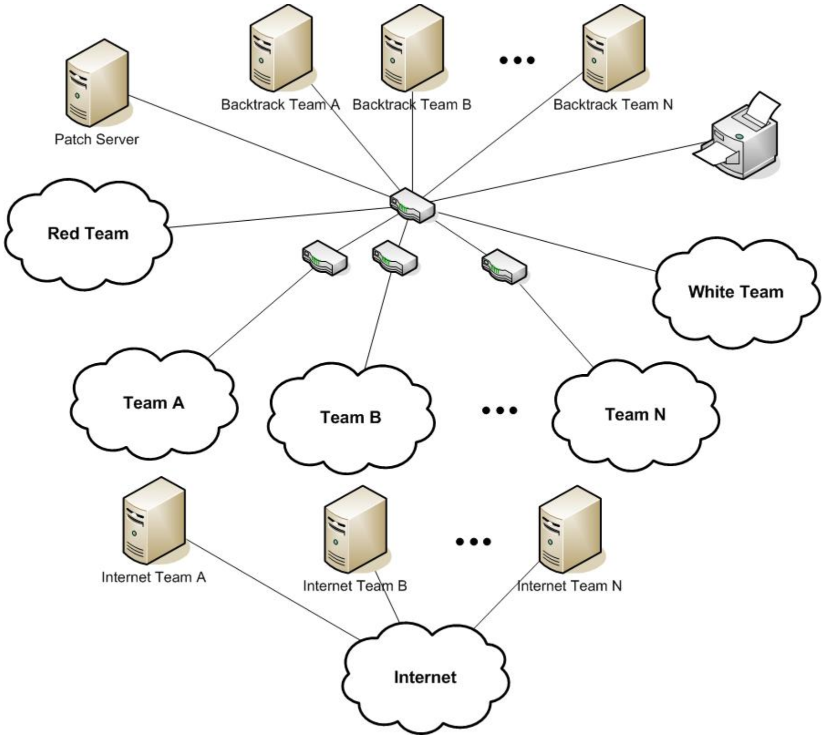

Figure 1. Network layout for the competition.

packet-capture machines were provided on other routers. To allow for possible correlation of network data with captured video, one of the routers chosen was that of University of Washington's iSchool Team (Team 7) who agreed to be monitored during the competition. Monitoring programs, such as key loggers, packet capture, etc., were intentionally left off the systems, since they may have been discovered and regarded as malicious injections by the contestants. To avoid interrupting the competition, we accepted the possibility of data loss due to the Red Team attacks.

C. Video and Audio Data Capture

Video and audio were captured from a limited portion of the competition. Video and audio data were not captured on all teams because recording all the teams would have been prohibitively expensive in both equipment and the time it would take to analyze all the data. Thus, we concentrated collection on Team 7, whose members consented to allow video and audio capture at their tables during the entire competition. Six Logitech webcams were placed strategically within the iSchool team's area to capture video from which to note interactions and collaboration among participants. With this data researchers aimed to observe the collaborative efforts and interaction between the team members at any critical event marked by a time stamp in network logs. Additionally, we captured audio via a digital audio recorded placed at the Red Team table. The City University of Seattle also filmed the entire event and agreed to provide the access to their raw footage. This footage was particularly useful to record the Red team's brief-back to the student teams at the end of the competition; however, the data were not available to the researchers at the time of analysis.

D. Situational Awareness Data Capture

Team situational awareness was measured as a way to infer team performance independently from the competition performance scoring. Additionally, timing and performance data were collected from the business injects as an indicator of situational awareness.

Four 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 to conduct an assessment of team situational awareness using Durso's Situation Present Assessment Method (SPAM) [1].

1) The Questions

A matrix of questions was designed for the student teams. Questions were categorized in three sections concentrating on (1) concerns of the past hour, (2) those of the present or (3) predicted concerns spanning one hour into the future. Durso's work shows that future-oriented questions were most indicative of expertise, so the tense of the 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 actually occur?
2. Threat-related
   a. Policies: What offensive actions should the attacker take to gain access?
   b. Priorities: From an attacker's perspective, what is the most important action to take?
   c. Events: What offensive actions actually (will) happen?

From this taxonomy, a list of 48 general questions was generated and arranged in a table. Every half hour, the research team met and randomly selected one of these questions, and together, they asked the same question to all seven teams during that 30-minute time period. The research team was able to ask 22 of these questions over the entire period of competition. Researchers adapted questions to the current situation, filling in information as needed.

2) B. The Querying Protocol

Each researcher was given the task of querying 2-3 student teams, selected at random, during the 30-minute segment. Since team members were not seated in constant places during the competition, researchers assigned identification numbers to each member within a team and used random generator to decide which team member they would approach with a question. This reduced bias when picking the subject. However Team-1 chose a spokesperson to handle all queries. In that case, the researcher honored the team's policy and always approached the spokesperson. Our intent is to infer *team* situational awareness from these queries and compare it to team performance. Durso's method was intended to measure individual situational awareness, but in this collaborative exercise, team members did not receive individual scores that would allow us to compare situational awareness to an individual's score. By randomly choosing a new team member each round, we attempted to control for variation caused by differences in individuals' levels of situational awareness.

To ask a question, a researcher would approach a participant and place a question card face down on the table in the view of the interviewee. He then would start his audio recorder and would 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.

3) C. Additional considerations regarding the querying process

Durso's method [1] was employed to measure both situational awareness and workload. According to 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 completely read to when the participant answers is believed to be a measure of situational awareness. Researchers noted the time when the questions were placed on the table and the time when team member started reading the question by saying, "picked up the question" in the recorder.

Researchers took great care while designing the data-collection protocol, but variation associated with individuals' responses and time stamps did occur. Some participants were very concise in answering the questions and adhered to the yes/no answer format leading to short response times. However, other participants elaborated and gave lots of information about the situation. To address these inconsistencies, we did not take into account the amount of information researchers collected from these questions. Rather, we calculated the response time as the length of time for the respondent to make their initial statement in response to the question posed by the researcher. In this experiment, use of SPAM is somewhat different from that of Durso. SPAM is a secondary task method of evaluating situational awareness. To our knowledge, such a method has never been used to evaluate characteristics of groups. The accuracy of this method when used to assess team situational awareness remains a matter for future research.

E. Hindrances in Using PRCCDC as a Data Collection Venue

There are some problems discovered in using CCDC events as data sources. These events are often high-stress venues that allow students to demonstrate their abilities to potential employers who are observing the competition. Thus, some participants might feel some anxiety knowing that they are being monitored during the competition and may not perform optimally. Several teams expressed chagrin when they were asked to be video recorded; therefore only one team was filmed.

There were also data collection difficulties because the venue was not a tightly controlled experiment. Since the competition was a high-profile event for the students where potential employers could evaluate their capabilities, every effort was made to enable students to do their best. Experimental controls had to be of secondary importance. Thus, the teams were informed that they were not to be scored on the basis of their responses to the situational awareness questions. As a result, sometimes teams didn't take questions seriously and thus we had as many as 4-5 missing values in the response data of each team.

Many uncontrolled distractions in the competition setting had effects that may be larger than the situational awareness effects being measured. For instance, sometimes a participant did not respond immediately because he was in a conversation, not truly busy. Another competition-borne control problem occurred if a participant picked up the question, but was then interrupted before he/she could answer. Some participants may not have answered even though they possibly knew the answer. One of the researchers observed this when he asked a question about the performance of a system to a participant who was monitoring that system and the participant answered, "I don't know." Notwithstanding these challenges, we did observe associations between our measures of situational awareness and the performance score earned by the team, which we discuss below.

Since these events are competitions in their own right, not simply experiments, the research team was constrained by the official competition rules. For instance, situational awareness queries were not made part of the scoring criteria, which might have better ensured that student participants would take them seriously. This definitely affected the data quality of the situational awareness. Additionally, the researchers were constrained to ensure that they did not disadvantage, or advantage, any single team by informing them or otherwise influencing them to take a particular course of action. Thus, researchers were restricted to asking more dynamic situational awareness questions based on Red Team plan of attacks and current activity rather than scripted questions where ground truth was known. Despite the hindrances that this venue has for conducting controlled experiments, the PRCCDC and similar CCDC events can be valuable sources of data for cyber researchers.

III. Analysis of Situational Awareness and Performance

We posit that one of the key indicators of performance is the time a team would take to respond to a question posed by the researchers. However, it is important to note that in a number of instances, five minutes of silence did elapse after the researcher's petition for the team to answer a question. In these situations, no response time could be recorded. These results are summarized in **Error! Reference source not found.**2. Note that the highest scoring team, Team 2, did not answer half of their questions, whereas the other teams answered all (or all but one) of their questions. Perhaps this occurred because Team 2 was overly busy, or perhaps they placed higher priority on responding to the injects than responding to researcher questions.

Consequently, the association between response time and team score (shown in **Error! Reference source not found.**3) could only be considered for the set of questions that a given team actually answered. The correlation coefficient for team score and mean response time was -0.600, a somewhat weak correlation, suggesting that teams which took longer to answer questions (which we presume to be indicative of lower situational awareness) tended to score lower than those teams that responded more quickly. Note that teams 3, 5, and 7 did answer most of their questions

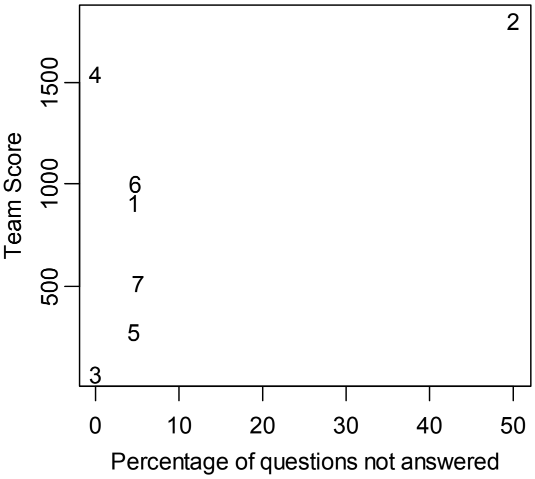

Figure 2. Team score versus the percentage of questions not answered by the team. Numbers in the plot represent the team label.

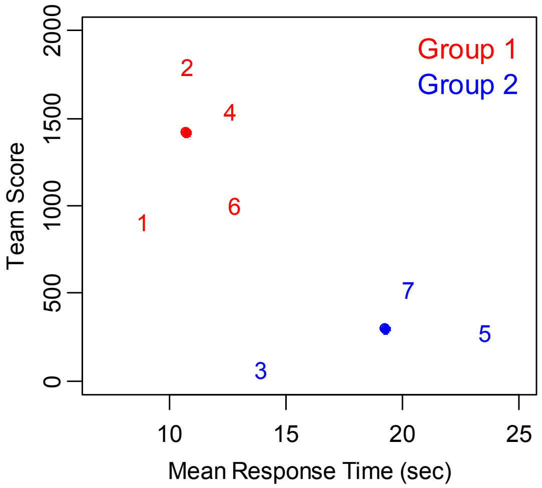

Figure 3. Team score versus the mean response time to the questions posed to each team. Numerals represent the team labels. The solid circles represent the two-dimensional means of Groups 1 and 2.

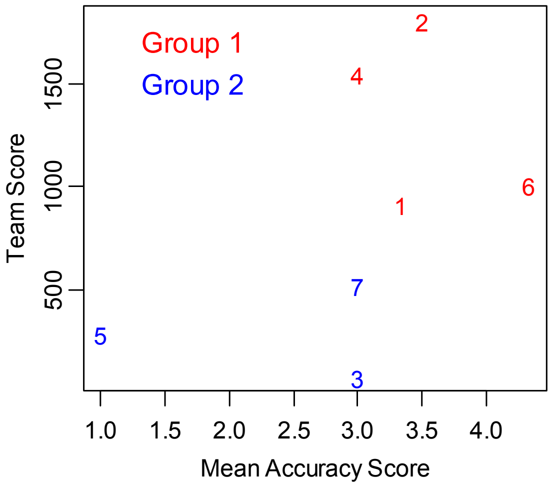

Figure 4. Team score versus the mean accuracy score. For comparison, the same groupings shown in **Error! Reference source not found.** are repeated here.

fairly quickly, but their average response time was increased substantially by a handful of questions where they took a relatively long time to answer. Visual inspection of Figure 3 reveals two natural groupings, or clusters, which are corroborated by both k-means and hierarchical clustering [3, 4]. A Hotelling T^2 test of the null hypothesis that the two-dimensional group means are equal [3, 4] results in a p-value less than 0.017, providing evidence that the two-dimensional population means of these two groups are, in fact, distinct. This suggests that Group 1 was indeed higher scoring with faster response times, while Group 2 scored lower with slower response times.

In addition to measuring response time we graded the accuracy of the responses to three of the questions using a scale of 1 to 5, where 5 indicated the most accurate, or most correct answer. Only three of the 22 questions posed to each team were graded because the other 19 questions did not lend themselves to accuracy assessment. (For Team 2, we only could grade two questions, since one of these graded questions was not answered by Team 2). We presume these accuracy scores are also a measure of situational awareness. A plot of the team score versus the mean accuracy score is given in Figure 4. The correlation coefficient between the team score and the mean accuracy score was only 0.474. However, it is interesting to note that **Error! Reference source not found.**4 roughly resembles a reflection (over the vertical axis) of the pattern shown in **Error! Reference source not found.**3, suggesting the potential that teams which have a higher situational awareness (as reflected by the accuracy of their responses to questions) tend to score higher in the competition. However, our conclusions are necessarily tentative due to the fact that, at most, only three questions from each team were graded for accuracy.

III. Conclusions and Future Work

We found that measures of team situational awareness (time and accuracy in responding to questions) is weakly correlated with overall performance, as measured by the team competition score. Because our investigation of the PRCCDC was an observational study with a small sample size (only seven teams), the extent to which we may generalize our conclusions is limited. Nonetheless, our analysis suggests that further study of measures of team situational awareness and their correlation to cyber warrior effectiveness is warranted.

This study helped the researchers devise ways to collect situational awareness data for cyber events, and determine its efficacy at predicting performance. In the future, the authors hope to repeat a similar experiment, but interpose Vulcan, a collaborative enhancement technology, as a treatment. Vulcan is designed to improve analyst performance across competing teams, so as not to (dis)advantage any team.

In the future, we will use different interview techniques for the situational awareness queries and different methods of query delivery and notification. Questions to be asked will measure the effectiveness of collaboration. It may be desirable to integrate situational awareness queries with the scoring mechanisms. In addition, semi-structured interviews, or other data sources such as physiological stress measurements, could be introduced to enrich the data set. In the future, we also hope to assess the effectiveness of variations of the SPAM approach for team measurement. Such measurements could allow the development of a useful profile of the effective cyber warrior.

Among the contributions of this work are:

- Lessons learned in applying situational awareness measurement methods to cyber analytic situations,
- Adapting a secondary-task situational awareness assessment method to team assessment, and
- Experience gained from instrumenting a competition event in order to conduct scientific experimentation.

We expect that our early work in applying situational awareness measurement to cyber analysis situations will help researchers design and hone treatments that will improve collaboration among analysts. We also envision our data and experimentation results could be useful to design the situational based decision making process for machine learning. The end result will be safer networks and computers for the society.

IV. Acknowledgements

This work has been supported by the Information and Infrastructure Integrity Initiative, an internal (Laboratory Directed Research and Development) investment of the Pacific Northwest National Laboratory in Richland, WA. The Pacific Northwest National Laboratory is managed for the US Department of Energy by Battelle Memorial Institute under Contract DEAC05-76RL01830.

The authors also wish to thank the organizations who sponsored the PRCCDC, without whom this event would not have been possible—Black Hat, The Boeing Company, Cisco, and Microsoft—in addition to the tireless volunteers and student teams from all of the participating schools.

References

[1] Durso, F., Dattel, A, Banbury, S. and Tremblay, S. (2004). SPAM: The real-time assessment of SA. Burlington, VT: Ashgate Publishing Company, pp. 137-154.

[2] Endsley, M. (1989). a Methodology for the objective measurement of pilot situational awareness. Copenhagen, Denmark: Neuilly SurSeine, France, pp. 1-9.

[3] Rencher AC. (2002). *Methods of Multivariate Analysis,* 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.

[4] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.