Results and Lessons Learned from a User Study of Display Effectiveness with Experienced Cyber Security Network Analysts

Christopher J. Garneau, Robert F. Erbacher, Renée E. Etoty, and Steve E. Hutchinson
U.S. Army Research Laboratory

Abstract

Background. Visualization tools have been developed for various network analysis tasks for Computer Network Defense (CND) analysts, yet there are few empirical studies in the domain of cyber security that validate the efficacy of various graphical constructions with respect to enhancing analysts' situation awareness.

Aim. The aim of this study is to empirically evaluate the utility of graphical tools for enhancing analysts' situation awareness of network alert data compared with traditional tabular/textual tools. This paper focuses on results of the study and lessons learned for future similar studies.

Method. A tabular display was presented along with two alternative graphical displays in a web-based environment to 24 experienced network analysts. Participants were asked to use the displays sequentially to identify intrusion attempts as quickly and accurately as possible. Data were fabricated by an experienced analyst and do not rely on alert data from a real network.

Results. Analysts performed well on the tabular (baseline) display and also preferred this display to others. However, they were slightly faster and similarly accurate using one of the graphical alternatives (node-link). Subjective feedback shows that many analysts are receptive to new tools while some are skeptical.

Conclusions. Graphical analysis tools have the capability of enhancing situation awareness by preprocessing and graphically arranging data for analysis. Real-world analysts bring a wealth of experience and insight to this sort of research, and the large number of expert responses included in this study is unique. Tempering analyst expectations for the study by clearly explaining the study environment and tasks to be completed would likely lead to more accurate results.

1. Introduction

Suspicious computer network activity identified by an Intrusion Detection System (IDS) requires Computer Network Defense (CND) analysts to make quick, accurate decisions about activity that warrants further investigation and possible remediation. In this initial triage phase of intrusion detection, details on any potential attacks are less important than overall situation awareness of suspicious activity as identified by the IDS configuration. Constant monitoring of textual log files is a difficult task for humans, even for analysts who are trained to quickly recognize abnormal patterns in the data. There exists an opportunity to develop and implement visualizations that preprocess and graphically arrange data to aid in the cyber security analysts' search activities, however graphical techniques have not seen wide implementation in analysts' operations [7]. This paper discusses a user study that investigated three interfaces for representing network alert data to gain insight on features and visual attributes that would be most effective for enhancing analyst situation awareness. The focus of this paper is on the design of the study and how subjective feedback from the study may inform and improve the design of future studies.

1.1. Background

Visualization tools have been developed for various network analysis tasks for CND analysts, including identifying salient features in datasets, tracking analyses, reusing effective workflows, testing hypotheses, and so on. However, in general, there are few available studies in the domain of cyber security that validate the efficacy of various constructions with respect to enhancing analysts' comprehension of alert data. Some studies have investigated analyst needs and have employed cognitive task analysis (CTA) [4, 6]. Requirements and characteristics of next-generation visualizations have resulted from these efforts. More research to better understand analyst needs and validate visual tools will benefit the state of the art in cyber security network analysis and the available tools used to support such analyses.

1.2. Goals of the Study

The overarching goal of this study is to evaluate the utility of graphical tools for enhancing analysts' situation awareness, compared with traditional tabular/textual tools. Specifically, questions posed by the study—relevant to the discussion in this paper—include:

- Do graphical displays enhance performance?

- What barriers limit the adoption of graphical displays?
- What types of graphical displays would be most effective?

In this paper, only results from experienced network analysts are considered, even though the study also included novices (university students) in the pool of participants.

2. Related Work

Evaluating scientific visualization techniques is a longstanding challenge [1, 2, 10]. Similarly, the field of information visualization has a strong tradition in pioneering research in evaluation techniques [3, 14, 17]. User studies often rely on timing and accuracy information collected during the study coupled with subjective user surveys given after the experiment is completed. This combination of empirical measurement with subjective questionnaire is designed to assess the efficacy of a visualization technique with respect to related methods. However, the analysis of user evaluation studies remains difficult. These challenges are often compounded by the limited empirical data acquired during the study. Beyond the specific details of the many user study experiments, they all share a common goal: to assess the strengths and weaknesses inherent to a visualization technique or system. Incorporating as many objective measures as possible into the experiment not only provides a more robust analysis, but also mitigates subjectivity often introduced by users' preferences, biases, and retrospection.

Due to the nature of today's complex scientific data, simply displaying all available information does not adequately meet the demands of domain scientists. A wide variety of visualizations for cyber security analysts have been proposed [16]. Determining the best use of visualization techniques is one of the goals of scientific visualization evaluations. The types of improvements offered by the method being studied dictate evaluation methods. Some evaluations are concerned primarily with technological improvements such as rendering speed or the management of large data. User studies have been used to evaluate everything from aircraft cockpits [15] and surgical environments [13] to visualization methods [11]. Evaluating visualization methods that focus on human factors often employ user studies or expert evaluations to determine their effects on interpretation and usability. An expert assessment takes advantage of knowledgeable users to enable more poignant analysis of use cases and these experts also bring with them their own preconceptions and preferences that can skew studies. Traditional evaluation methods provide mechanisms to gauge aspects of visualizations or environment. Unfortunately, experiments using surveys to measure user experience introduce subjectivity and bias from the users. Subjectivity in user responses may be partially mitigated using questionnaires developed with the Likert Scale [12]. Subjectivity in evaluation may provide important insights into how users interact with the systems being studied. However, subjective measures do not help answer questions regarding how effective a method is at eliciting insight from a dataset. This is a primary purpose of visualization. Our goal and purpose is to use this project as an empirical study to examine the cognitive aspects of visual displays with the goal of identifying components and representations that most effectively aid the computer network analyst in interpreting the underlying activity in a network sample. Results from the study are helpful to understand the potential and limitations of the suggested visual displays attempting to aid analysts' needs to better achieve their tasks.

3. Study Overview and Method

In the study, participants acted as analysts and their job was to identify as many network threats as possible within a limited set of IDS alert data. Because the goal of the study was to examine how situation awareness may be enhanced in the initial triage phase, no additional investigation of alerts was required or permitted and analysts were expected to discriminate between potentially malicious and benign alerts based strictly upon the data presented in the displays. Objective response variables include: (1) true-positive rate of identification of intrusion attempts for each type of display, (2) false-positive rate of identification of intrusion attempts for each type of display, and (3) time required for identification for each type of display. Subjective feedback was also collected and is the focus of the "lessons learned" presented in this paper.

3.1. Display Design

Three types of displays were chosen for inclusion in the study:

1. **Tabular display** (baseline): Basic functionality is similar to Microsoft Excel. Participants could sort and filter data by any parameter, and individual rows were selected for submission via checkboxes. See Fig. 1.

	ID	Time	Src Entity	Src Port	Dst Entity	Dst Port	Dst Country	Alert
☐	1	7/5/13 5:37	USA.3	52869	USA.12	445	USA	Fragmented IP Packet
☐	2	7/5/13 5:11	USA.113	2787	land of OZ.159	80	land of OZ	WEB-MISC Netscape Enterprise Server directory view
☐	3	7/5/13 5:30	USA.12	3377	Pellucidar.28	443	Pellucidar	Fragmented IP Packet
☐	4	7/5/13 5:15	USA.110	2986	land of OZ.99	80	land of OZ	WEB-MISC Netscape Enterprise Server directory view
☐	5	7/5/13 5:35	USA.12	3498	Neverland.49	443	Neverland	Fragmented IP Packet
☐	6	7/5/13 5:09	USA.3	2660	Dreamlands.106	80	Dreamlands	WEB-CGI php.cgi access
☐	7	7/5/13 5:38	USA.76	53249	Hogwarts.88	8080	Hogwarts	ET TROJAN Qhosts Trojan Check-in
☐	8	7/5/13 5:21	USA.12	3193	Wonderland.8	80	Wonderland	WEB-CGI php.cgi access
☐	9	7/5/13 5:11	USA.113	2800	Gullivers World.198	443	Gullivers World	Fragmented IP Packet
☐	10	7/5/13 5:12	USA.12	2852	USA.4	21	USA	FTP Satan Scan
☐	11	7/5/13 5:30	USA.12	3386	Dreamlands.106	80	Dreamlands	WEB-CGI php.cgi access
☐	12	7/5/13 5:07	USA.3	52593	Blefuscu.112	443	Blefuscu	Fragmented IP Packet
☐	13	7/5/13 5:04	USA.12	2587	Utopia.211	80	Utopia	Javascript Exploit CVE-2012-09-10a
☐	14	7/5/13 5:38	USA.12	52870	Pern.152	21	Pern	FTP STOR overflow attempt
☐	15	7/5/13 5:11	USA.113	2768	Neverland.49	443	Neverland	Fragmented IP Packet
☐	16	7/5/13 5:12	USA.12	2859	USA.1	21	USA	FTP Satan Scan
☐	17	7/5/13 5:11	USA.113	2737	Middle-Earth.64	80	Middle-Earth	INFO Connection Closed MSG from Port 80
☐	18	7/5/13 5:30	USA.12	3383	Pern.110	443	Pern	Fragmented IP Packet
☐	19	7/5/13 5:25	USA.96	65113	Atlantis.10	8000	Atlantis	ET TROJAN Sality Variant Downloader Activity
☐	20	7/5/13 5:11	USA.113	2731	Deltora.36	80	Deltora	INFO Connection Closed MSG from Port 80
☐	21	7/5/13 5:21	USA.12	20	USA.3	52644	USA	FTP - Suspicious MGET Command
☐	22	7/5/13 5:30	USA.12	3394	Tatooine.157	443	Tatooine	Fragmented IP Packet
☐	23	7/5/13 5:11	USA.113	2758	Pandora.116	80	Pandora	WEB-CGI webspeed access
☐	24	7/5/13 5:14	USA.12	2948	USA.11	21	USA	FTP Satan Scan
☐	25	7/5/13 5:12	USA.12	2853	USA.5	21	USA	FTP Satan Scan

Fig. 1 Tabular display, showing alerts with ID 1-24 (no alerts selected).

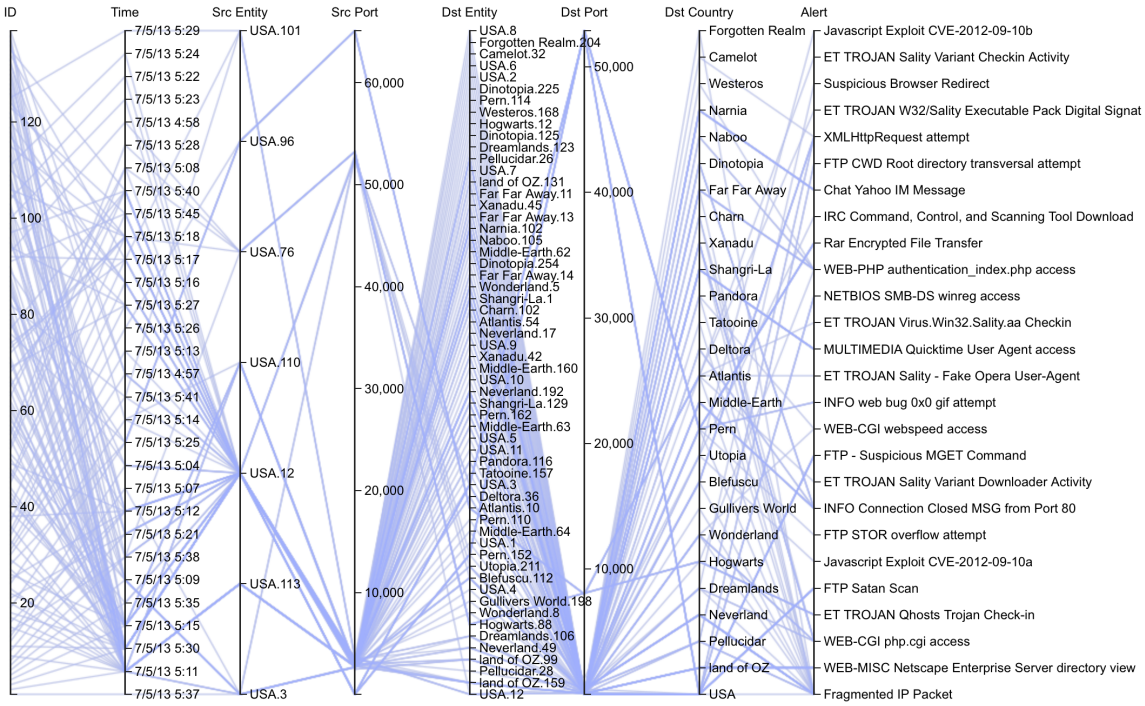

Fig. 2 Parallel coordinates display shown to participants (no alerts selected).

2. **Parallel coordinates display** (graphical alternative 1): This display is one of the most published multivariate visualization techniques [e.g., 8, 9]. Participants could highlight/filter a range of values on any of the parameters; to further refine the selection for submission, ranges on additional axes could be selected. See Fig. 2.

3. **Node-link display** (graphical alternative 2): This display has been tailored to the task of intrusion detection based on related visualization research [5]. As a participant moused over a

marker (red dot), a popup appeared showing details of the alert associated with the source-destination node pair. Participants could click any number of markers for submission. See Fig. 3. This display was selected by an ex-analyst who had switched to the R&D side of the house and had experience reviewing many different display formats.

The source data presented in each display were identical and were synthesized by an expert from an hour's worth of alert messages. For security reasons, it was not possible to use real data captured from an operational environment and so data were fabricated for this study. The data uses an attack scenario where many external nodes are attacking a smaller number of friendly peer nodes. The alert data contain three types of intrusion attempts of varying difficulty: (1) a three-stage intrusion that consisted of a web infection, scanning, and data exfiltration (32 alerts, "easy" difficulty), (2) periodic Trojan scanning (5 alerts, "moderate" difficulty), and (3) Sality Trojan infection (5 alerts, "hard" difficulty). 42 alerts of a total 139 alerts belonged to one of the three intrusion attempts. Eight parameters were associated with each alert message in each of the displays: (1) alert ID, (2) date/time stamp, (3) source entity/IP, (4) source port, (5) destination entity/IP, (6) destination port, (7) destination country, and (8) alert message (see Fig. 1 for a tabular representation of a subset of the data).

3.2. Study Design

The study collected background and demographic data, utilized pre- and post-task questionnaires, and also obtained objective and subjective feedback. The study was administered as a within-subjects design (i.e., every single participant subjected to every single treatment). Each participant completed the task independently while sitting at a computer workstation using the three displays sequentially, and the display presentation order was varied to minimize the effects of practice and order bias. In other words, each participant completed the task using their first assigned display (with a time limit of 20 minutes), then their second assigned display, then

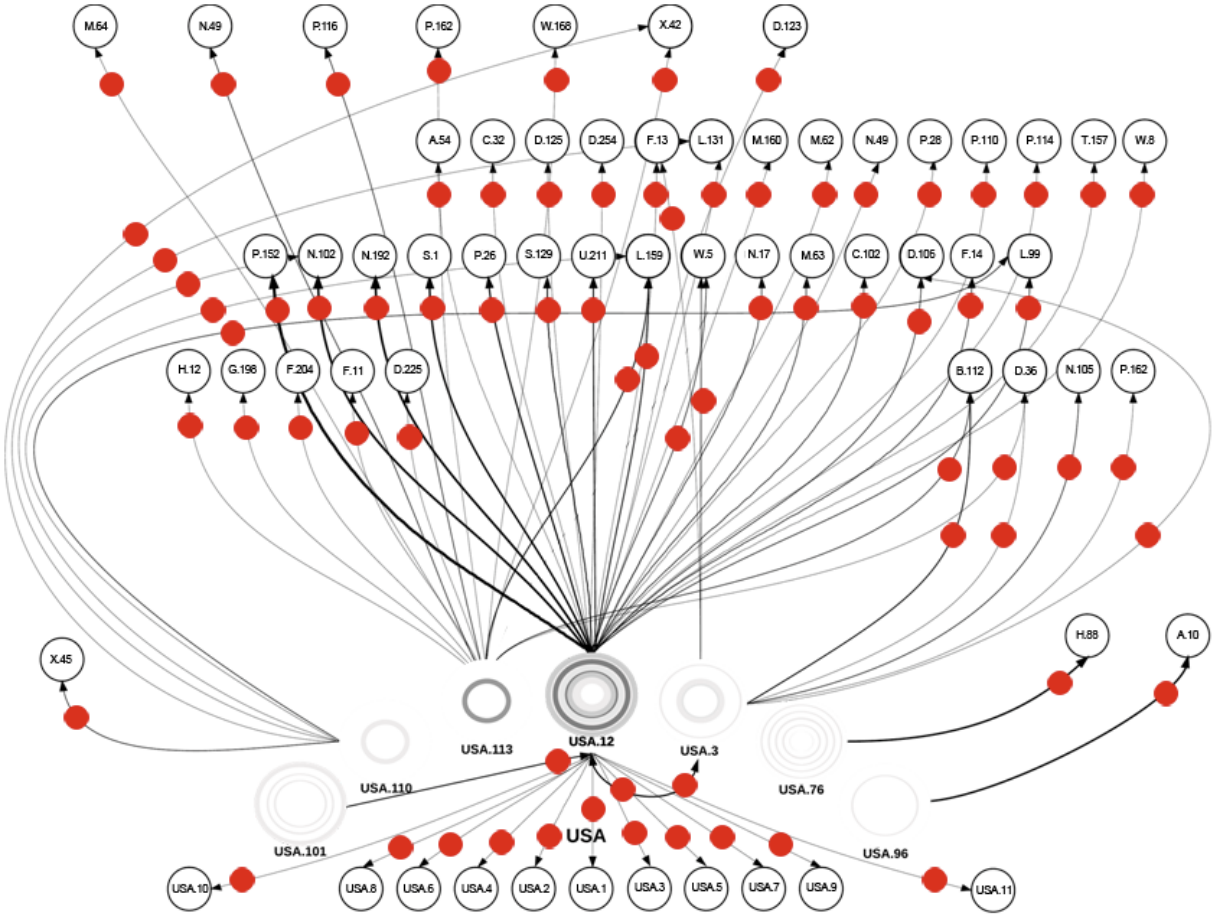

Fig. 3 Node-link display (no alerts selected).

their third assigned display. Since there are three unique displays there were six permutations of the ordering. The order of the displays was assigned such that a similar number of participants were assigned to each display order (i.e., roughly equal number of participants were assigned to the following orderings: TNP, TPN, NPT, NTP, PTN, PNT, where T=Tabular, P=Parallel Coordinates, and N=Node-link). The computer workstation consisted of a typical laptop computer, external monitor, keyboard, and mouse with a typical desk and office chair. The study was conducted in a web-based environment; survey questions were administered with the survey application LimeSurvey[1] and the main study task with the three displays was administered via custom HTML and JavaScript code. The study was approved by an appropriately constituted institutional review board (IRB) at ARL.

3.3. Procedure

The procedure used for the study is described in this section.

Step 1. Study began with a welcome followed by an introduction of the investigators.

Step 2. Investigators then briefed the participants on the study and obtain informed consent. Participants of this study were given a random identification number.

Step 3. The investigators explained each visual display and the techniques for representing a network system's attributes.

Step 4. The investigators conducted a demo of the participants' tasks in the web-based environment. This served as practice for the participants.

Step 5. Investigators provided time to entertain participants' questions concerning their tasks or any other aspects of the study.

Step 6. Participants completed background, demographic, and pre-task questionnaires.

Step 7. Participants completed main task of the study with the three visual representations as described in Section 3.2 (i.e., they were presented with the three displays according to the assigned ordering and were asked to identify as many intrusion attempts as possible in each).

Step 8. Participants completed their post-task questionnaires and provided the investigators with any final remarks or comments.

Step 9. Investigators lead a debrief session and provided the participants with a copy of the signed consent form.

Participants were given a maximum of three hours to complete the tasks described above as well as tasks for a similar related study. Most participants completed the tasks for this study only—including pre- and post-task questionnaires—in about 1.5 hours. Participants completed the tasks independently either alone with the investigator in the room or with the investigator plus one other participant in the room (but working separately at opposite sides of the room).

4. Results

Results of the study are presented next, including demographics of the participants, objective performance of participants on the analysis task, and a subset of comments provided by participants.

4.1. Participant Characteristics and Demographics

The participant population consisted of 24 experienced analysts from ARL who actively or previously had conducted CND analyses and were employed by ARL at the time of the study. These analysts were selected for inclusion in the study due to their unique skillset and availability. A majority of the analysts' full-time job is monitoring sensors for malicious activity—initially through generated alerts and subsequently through raw data logs files. While "expert" is a vague and potentially misleading label, the participants in this study are considered to be experts specifically at analyzing network data for malicious activity. All were assumed to be familiar with tabular tools and may or may not have been familiar with graphical tools. All were eighteen years of age or older, had 20/20 vision (or corrected 20/20 vision), all passed a test for colorblindness, and none reported having any other disabilities. Note that demographic questions did not force a response so some participants did not answer some questions. Table 1 summarizes participant demographics.

Table 1 Demographics for participants. Some participants did not answer one or more of these questions.

	Number of Analysts
Gender	
Female	0
Male	22
Race	
White	13
Black or African American	5

[1] https://www.limesurvey.org

Asian	1
Other	3
Age	
18-25 years	3
26-35 years	11
36-45 years	8
46-55 years	2
Highest level of education completed	
Some college but did not finish	5
Two-year college degree/A.A./A.S.	7
Four-year college degree/B.A./B.S.	7
Some graduate work	3
Completed Masters or professional degree	2
Experience as cyber analyst	
Less than 1 year	2
1-3 years	7
3-5 years	11
5-10 years	2
Greater than 10 years	1

Table 2 Objective performance parameters for each of the three displays. TP and FP give the average number of true positives and false positives identified, respectively. n indicates the number of responses considered for the given display—this metric varies because not all analysts completed the task using all displays.

	Tabular	Parallel Coordinates	Node-Link
n	23	21	23
TP	24.8	20.3	25.9
FP	17.1	30.4	15.7
Completion Time (min)	15.6	11.9	12.2
Accuracy	0.752	0.624	0.771
Precision	0.670	0.501	0.703
Recall	0.590	0.482	0.617

Table 3 Differences between means of various objective performance parameters are indicated for tabular vs. parallel coordinates (PC) and tabular vs. node-link. Positive values indicate higher values for the first of each pair. Significance is also indicated (* Significant at the 0.05 probability level, ** Significant at the 0.01 probability level, * Significant at the 0.001 probability level).**

	Tabular vs. PC	Tabular vs. Node-link	Node-link vs. PC
TP	+4.50	-1.09	+5.58
FP	-13.34	+1.35	-14.7*
Time (min)	+3.76*	+3.44*	+0.32
Accuracy	+0.13*	-0.019	+0.15**
Precision	+0.17	-0.033	+0.20*
Recall	+0.11	-0.028	+0.14

4.2. Objective Performance

Objective performance is measured in terms of: (1) true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) that analysts identified in the dataset; (2) measures derived from total TP, FP, TN, and FN (such as accuracy); and (3) time/duration to complete tasks. The derived measures are defined in terms of total TP, FP, TN, and FN as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Tables 2 and 3 provide summaries of these metrics across the displays. Fig. 4 plots the number of true positives identified against the number of false positives identified for each participant using each of the three

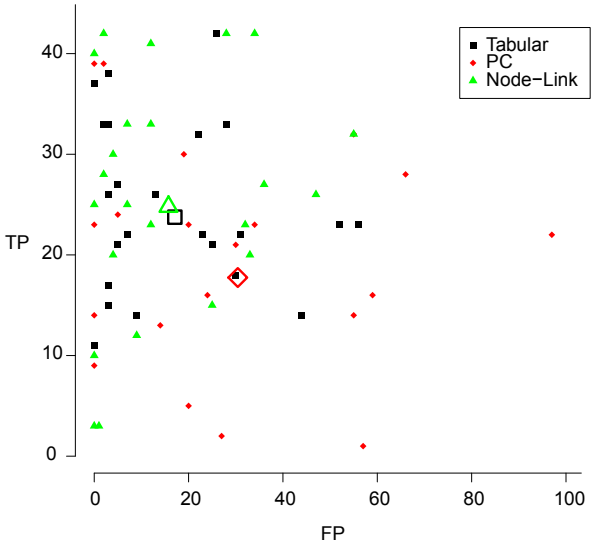

Fig. 4 Number of true positives (TP) vs. false positives (FP) for each participant using each display. Averages for each of the three displays are shown using the larger, open (unfilled) shapes.

displays.

4.3. Subjective Feedback/Comments

Subjective feedback was collected via questionnaires. This section presents a subset of participant comments to the indicated free response prompts that are representative of most responses (duplicate or very short responses are excluded, instead focusing on those providing a unique point of view). All responses could not be included here due to limited space.

"What components of the visual displays (table, parallel coordinates, and node link display) were most effective?"

- "Table is much better at identifying the actual alerts. Parallel is more involved with showing all the alerts that are on the node. Node Link is best at showing all the alerts that are attached to each ID"
- "The table was the most effective. It is easy, you don't need all these fancy visual tools to find issues. Make everything simpler. The Node display and the Parallel display looked like a lot of noise. I wasn't interested in using them."

"What aspects of the visualizations (table, parallel coordinates, and node link display) did you like best?"

- "I thought the node link display was interesting. It was nice to see a view of a network topology and the path the data travels."
- "Visually separating the internal hosts from external hosts to quickly see flow of data between internal only and internal to external."
- "Table is much better for seeing and lumping the alerts together. Parallel showed the flow a lot better. Node link showed the flow and other ID numbers to the alerts faster."
- "I liked being able to quickly identify the most active times of the day and the most common request domains of the parallel coordinates displays."

"What aspects of the visualizations (table, parallel coordinates, and node link display) did you not like?"

- "It takes a while to glance at everything when you can glance at percentages and numbers. Those are quicker to grasp sometimes."
- "The line display and node display were convoluted at best; they detracted from the information presented. In terms of investigative procedure, I believe they would only serve to stifle."
- "The parallel coordinates was a nightmare to visualize and the node link display was far too time consuming to hover through."
- "Connections were very hard to follow, information was displayed in a non-intuitive manner, correlations were very difficult to find without excessive work."
- "The parallel coordinates interface was not useful or intuitive and I could not sort the coordinates. Moreover, once I was finished with an alert it should have been removed from my view. Also I should be able to remove noise from my view with a filter. The node link display was slightly more useful but the node sizes were not intuitive nor were the most important piece of analysis data displayed up front: the alerts!"
- "Parallel coordinates is good for fine analysis but not for raw / bulk analysis"
- "The graphical representations were completely unusable to me. The table was fine but there needs to be drill down options to see more data. I look at an alert then I check some traffic if I see something suspicious I dump the traffic and do a thorough investigation."

"What did you learn from this study?"

- "That there are some great relationship tools for network intrusion, and some not so great ones."
- "Being able to see correlations is very important (time, src ip, dst ip, etc). All of this information is very difficult to fit into a graphical interface. Without being able to sort and filter, this makes the analysts job far more difficult. The graphical displays seem decent for a "birds eye view", but a nightmare for actual everyday analysis."
- "Aggregated data obfuscates signal!"
- "A simple table can be the better option sometimes."
- "Graphics helps make analysis easier"
- "With all due respect regarding this project, it seems like there is a long way to go before a very useful graph or node visualization would be effective or efficient to use."
- "I do not like graphical displays while doing analysis. I find them highly unnecessary, unless of course I was presented with ones that are more intuitive."

5. Discussion of Results and Lessons Learned

The results from the study presented in this paper provide insight into two areas of inquiry: (1) what types of displays and visual attributes of those displays are most effective for enhancing analyst performance?, and (2) what aspects of the study implementation worked well

and what aspects need improvement for future similar studies, based on analyst feedback?

5.1. Discussion of Performance

Analysts performed well on the tabular (baseline) display, and also preferred this display to others. Based on their familiarity with this representation—and its use in their existing workflow—this was to be expected. Compared with performance on the parallel coordinates display, analysts were slower but more accurate using the tabular display. Compared with the node-link display, analysts were slower and about as accurate using the tabular display. It is notable that analysts had a high rate of false positive identification for the parallel coordinates display (see Table 2), despite the display's advantages for representing many-dimensional data. This would translate into more time wasted in a real-world scenario investigating benign alerts.

Subjective feedback confirms the objective performance results. Some analysts could see the value in the parallel coordinates display, but most felt strongly against it ("parallel coordinates is a nightmare to visualize", "the parallel coordinates interface was not useful or intuitive"). In general, the baseline tabular display was preferred to graphical alternatives ("the graphical representations were completely unusable to me"). However, some analysts could see the value in the graphical displays ("it was nice to see a view of a network topology and the path the data travels", "node link showed the flow and other ID numbers to the alerts faster").

5.2. Lessons Learned

It is instructive to consider analyst performance and feedback so future studies can improve upon the current work.

To avoid inaccuracies in study results, it is important to take participants' expectations into account in the study methodology. While specific instructions were provided to the analysts, some wanted more data and could not understand how to use the displays for the task provided. One said, "Table data was fine but I need more than play data to do proper analysis. It isn't just the alert or a darker line of traffic that determines infection or compromise." It seems as though the intent of the study was not well communicated to this participant. The displays were not intended to replicate a complete analysis session, but rather provide a tool for rapidly identifying indicators of compromise for further investigation and enhance situation awareness, i.e., we were only investigating initial triage, initial indicators at this stage. Future studies will need to be performed for detailed analysis of these indicators for full analysis. Moreover, one way to enhance participants' experience and also gain further insight would be to ask participants what their next analysis steps would be, even though these are not considered in the experiment; this might give analysts a sense that they have completed the analysis.

Designing a complete simulated environment would be ideal for the most accurate results. Such a study environment should include features such as: (1) data of real-world appearance and scale, (2) the ability to contact other teams, such as a threat center or a remote target site through a fully scripted conversation, (3) ability to make simulated communications with the forensic analysis team, and so on. However, such a study would require years of research and design and may not provide results that justify such an undertaking, and numerous pilot studies (similar to the current work) would have to precede such a detailed study. Leveraging expectations by noting to analysts that they were participating in a scaled-down study should have been emphasized in the pre-study briefing.

It likely would have been beneficial to emphasize to analysts possible future improvements to their workflow and explicitly ask them to consider components of the display that they found useful or interesting (rather than dismissing a certain display altogether as a "nightmare"). While experienced analysts bring a wealth of knowledge and insight to research of this nature, they have a certain way they approach their work and may be critical of alternatives. Some were receptive to new tools ("It's good that I got a chance to see what type of tools can be deployed in future and felt very good to leave feedback about these tools"). However, others had more cynical viewpoints ("The table was the most effective. It is easy, you don't need all these fancy visual tools to find issues. Make everything simpler. The Node display and the Parallel display looked like a lot of noise. I wasn't interested in using them"). It is undesirable to eliminate contrary perspectives, but approaching the study with a "help me to help you" attitude may enhance results. While analysts possess extensive knowledge in the cyber security domain, they likely do not have much knowledge of the principles for effective display design and may instinctively react negatively to a display that is unfamiliar. Negative feedback should be considered seriously but ought not discourage future innovation in tools and displays for the field.

Other modifications to the implementation of the study would likely improve results and participants' perception of the displays. Including as much interactivity as possible in the displays to be evaluated would benefit the participant experience and thus enhance credibility

of the results. While the web environment in this study used lightly customized JavaScript libraries (e.g., D3.js[2]) and permitted some interactivity, more extensive interactivity likely would have mitigated analyst complaints about limitations of the displays ("The parallel coordinates interface was not useful or intuitive and I could not sort the coordinates. Moreover, once I was finished with an alert it should have been removed from my view.", "Connections were very hard to follow, information was displayed in a non-intuitive manner, correlations were very difficult to find without excessive work").

The study was invasive for participants, requiring them to leave their regular work site and adjust to an unfamiliar laboratory setting. Future studies should attempt to make the study as non-invasive as possible. Ideally, experimentation would occur in the regular work environment, but if this is not possible (e.g., due to security restrictions), future studies should attempt to replicate environmental conditions of the analysts' environment (lighting, temperature, computer hardware, etc.) as closely as possible. These adjustments would make it more likely that a participant would behave as they normally do, which should be a goal of any future studies. However, such modifications should be weighed against the benefit of further instrumentation and data collection capability; for instance, eye tracking would provide insight into the elements of the displays to which participants were paying attention.

There were also other limitations in the study implementation that should be noted. Using the same alert data across the three displays might introduce confounding effects; i.e., participant exposure to the underlying dataset in the first presented display might shape interactions in subsequent displays. While randomizing presentation order for the displays somewhat lessens this effect, generating distinct yet similar data sets for each display might be preferable. In an attempt to enhance participants' incentive to perform well (and lend a game-like quality to the test environment) an accuracy indicator was added to each of the displays. However, there are several drawbacks to its inclusion: it would not exist in a real world scenario, it may have influenced perception of the tools, and it may have altered performance in unexpected ways. Similarly, a 20-minute time limit (enforced by a countdown timer visible to the participant) perhaps added a certain sense of realism and time pressure to the task but the time limit was chosen arbitrarily and benefits and effects on results are unclear. There also likely differences in policies based on site, and interpreting and understanding such differences is an important analyst responsibility. Future studies should address this component.

[2] http://d3js.org

5.3. Future Work

Future studies might consider several changes and enhancements to the study implementation discussed in this paper. First, modifying the experiment design by asking participants to self-rate confidence in their answers for comparison with actual accuracy scores might yield insight about the user experience of the interface. To better align with analyst expectations, future studies might better contextualize the visualizations within other tasks that analysts perform (analyzing alerts is only a part of discriminating malicious network activity from benign activity). Thoroughly understanding analyst workflow and the current tools used by the analyst participants would be essential.

Future studies might also further investigate integrating elements of traditional/tabular and graphical displays. While the displays selected for inclusion in this study were intended to be representative of different types of multivariate displays indicated for the analyst tasks under consideration, they have not been optimized for usability (e.g., placement and size of elements, controls, and so on) and fully implemented with the necessary features for detailed analysis. Future studies might investigate more complete and perhaps alternate types of displays for representing alert information (incorporating interactivity as discussed previously). Finally, future work might investigate the use of different populations of participants. While unreported here, this study also gathered input from "novice" users (university students); future work might investigate including novices that possess domain knowledge but have little or no operational experience (i.e., a new hire) to assess how training varies among the different kinds of displays.

6. Conclusion

This study revealed that analysts are most comfortable using analysis tools with which they are already familiar (i.e., tabular/textual tools), yet are able to achieve similar accuracy in less time for an alert scanning task using some graphical alternatives (node-link). Such graphical displays have the capability of enhancing situation awareness by preprocessing and graphically arranging data for analysis. Real-world analysts bring a wealth of experience and insight to the research, but tempering analyst expectations for the study by clearly explaining the study environment and tasks to be completed will likely lead to more accurate results. Similar future studies validating proposed alternative graphical tools should also try to make the interfaces as interac-

tive as possible and should be constructed with a keen knowledge of existing analyst tools and workflow.

7. References

[1] Acevedo, D. and Laidlaw, D. "Subjective quantification of perceptual interactions among some 2D scientific visualization methods." *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006): 1133-1140.

[2] Acevedo, D., Jacson, C., Drury, F. and Laidlaw, D. "Using visual design experts in critique-based evaluation of 2D vector visualization methods." *IEEE Transactions on Visualization and Computer Graphics* 14.4 (2008): 877-884.

[3] Carpendale, S. "Evaluating information visualizations." *Information Visualization*. Ed. Kerren, A., Stasko, J. T., Fekete, J., and North, C. Berlin: Springer, 2008. 19-45. Print.

[4] D'Amico, A., Whitley, K., Tesone, D., O'Brien, B., and Roth, E. "Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts." *Proceedings of the 49th Human Factors and Ergonomics Society Annual Meeting, Orlando, FL, 26-30 September 2005*. Santa Monica, CA: Human Factors and Ergonomics Society, 2005. 229-233.

[5] Erbacher, R. F., Walker, K. L., and Frincke, D. A. Intrusion and misuse detection in large-scale systems. *IEEE Computer Graphics and Applications* 22.1 (2002): 38-47.

[6] Erbacher, R. F., Frincke, D. A., Moody, S. J., Fink, G. "A Multi-Phase Network Situational Awareness Cognitive Task Analysis." *Information Visualization* 9.3 (2010): 204-219.

[7] Etoty, R. E. and Erbacher, R. F. *A Survey of Visualization Tools Assessed for Anomoly-Based Intrusion Detection Analysis*. Adelphi, MD: U.S. Army Research Laboratory, 2014. Print. ARL-TR-6891.

[8] Giacobe, N. and Xu, S. "Geovisual analytics for cyber security: Adopting the Geoviz toolkit." *Proceedings from the 6th Visual Analytics Science and Technology (VAST) IEEE Conference, Providence, RI, 23-28 October 2011*. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2011. 315-316.

[9] Goodall, J. R., and Sowul, M. "VIAssist: Visual analytics for cyber defense." *Proceedings of the 2009 IEEE Conference on Technologies for Homeland Security (HST), Waltham, MA, 11-12 May 2009*. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2009. 143-150.

[10] Kosara, R., Healey, C. G., Interrante, V., Laidlaw, D. H., and Ware, C. "Thoughts on user studies: Why, how and when." *IEEE Computer Graphics and Applications* 23.4 (2003): 20-25.

[11] Laidlaw, D. H., Kirby, R. M., Jackson, C. D., Davidson, J.S., Miller, T. S., Da Silva, M., Warren, W. H., and Tarr, M. J. "Comparing 2D vector field visualization methods: A user study." *IEEE Transactions on Visualization and Computer Graphics* 11.1 (2005): 59-70.

[12] Likert, R. "A technique for the measurement of attitudes." *Archives of Psychology* 22.140 (1932): 1-55.

[13] Reitinger, B., Bornik, A., Beichel, R., and Schmalstieg, D. "Liver surgery planning using virtual reality." *IEEE Computer Graphics and Applications* 26.6 (2006): 36-47.

[14] Riche, N. "Beyond system logging: Human logging for evaluating information visualization." *Proceedings of BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV) 2010, a workshop of the ACM Conference on Human Factors in Computing Systems, Atlanta, GA, 10-11 April 2010*. New York City, NY: Association for Computing Machinery Special Interest Group on Computer-Human Interaction, 2010.

[15] Sarter, N. B. and Woods, D. D. "Pilot interaction with cockpit automation II: An experimental study of pilots' model and awareness of the flight management system." *The International Journal of Aviation Psychology* 4.1 (1994): 1-28.

[16] Shiravi, H., Shiravi, A. and Ghorbani, A. "A survey of visualization systems for network security." *IEEE Transactions on Visualization and Computer Graphics* 18.8 (2012): 1313–1329.

[17] Shneiderman, B. and Plaisant, C. "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies." *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI) 2006, Venezia, Italy, May 23–26, 2006*. New York City, NY: Association for Computing Machinery Special Interest Group on Computer-Human Interaction, 2006.