Visualization is Better! A Comparative Evaluation

John R. Goodall *

Secure Decisions division of Applied Visions Inc.

ABSTRACT

User testing is an integral component of user-centered design, but has only rarely been applied to visualization for cyber security applications. This paper describes a comparative evaluation of a visualization application and a traditional interface for analyzing network packet captures, that was conducted as part of the user-centered design process. Structured, well-defined tasks and exploratory, open-ended tasks were completed with both tools. Accuracy and efficiency were measured for the well-defined tasks, number of insights was measured for exploratory tasks and user perceptions were recorded for each tool. The results of this evaluation demonstrated that users performed significantly more accurately in the well-defined tasks, discovered a higher number of insights and demonstrated a clear preference for the visualization tool. The study presented here may be useful for future visualization for network security visualization evaluation designers. Some of the challenges and lessons learned are described.

KEYWORDS: User testing, comparative evaluation, security visualization, user-centered design.

INDEX TERMS: H.5.2 [Information Interfaces and Presentation] User Interfaces; I.3.8 [Computer Graphics] Applications

1 INTRODUCTION

Visualization for Cyber Security (VizSec), has rapidly matured over the past several years and there are now many techniques and tools applying information visualization to the problems of cyber security, particularly in network traffic analysis [1-7]. However, while the design of several of these tools are grounded in the tasks that real world users face, these tools are rarely tested empirically. This paper attempts to define a study design that is applicable to user testing for network analysis applications for VizSec by presenting a within subject comparative evaluation of a VizSec application with a commonly used network traffic analysis tool.

TNV, shown in Figure 1, is a visualization tool to facilitate the analysis of network packet capture data [8]. Usability guidelines were followed during design work and usability problems were fleshed out through multiple formative evaluations, including two heuristic reviews and one round of usability testing. The evaluation described in this paper tested the performance and perception of users with TNV in comparison with a common tool used for network analysis.

User testing can help to determine the utility and limitations of systems. Information visualization evaluation practices vary, and can be summarized into four areas [9]:

- *Controlled experiments comparing design elements*: a comparison of specific widgets or information mappings.
- *Usability evaluation of a tool*: an evaluation of problems users encounter when using a tool as part of the design process.
- *Controlled experiments comparing two or more tools*: a comparison of multiple visualizations or the state of the art with a novel visualization.
- *Case studies of tools in realistic settings*: an evaluation of a visualization tool in a natural setting with users using the tool to accomplish real tasks.

The evaluation described in this paper falls into the third category, controlled experiments comparing two tools.

2 RELATED WORK

User testing is an essential component of user-centered design. Despite an increasing body of research, user testing is still atypical in VizSec research. The following is a representative sample of studies from the information visualization community that empirically evaluated two or more tools.

Sebrechts, et al. [10] used a between-subject study design in a comparison of text with both 2D and 3D visual representations of search results. Subjects were quasi-randomly assigned to each of the three visual conditions. Sixteen timed tasks were completed using a think aloud protocol followed by a satisfaction questionnaire. Results were examined qualitatively and quantitatively.

Stasko, et al. [11] used a within-subject study design in a comparison of two space-filling methods for visualizing hierarchical data. Sixteen subjects performed sixteen timed tasks on both tools using different hierarchies to avoid learning effects due to working with the same data twice. The hierarchies were approximately the same size, depth, and overall structure. The ordering and conditions varied across participants. A subjective evaluation followed the experiment and results were examined qualitatively and quantitatively.

Plaisant, Grosjean, and Bederson [12] also compared methods for viewing hierarchical data. They used a within-subject design to compare a traditional interface (Windows Explorer), the Hyberbolic tree browser, and SpaceTree with seven tasks. The order of the interface and task sets was counterbalanced. The three different task sets used different branches of the hierarchy that were similar in size and complexity. A subjective evaluation followed the experiment and results were examined qualitatively and quantitatively.

Risden, et al. [13] used a within-subject study design in a comparison of 3D and traditional browsers for directory management typical of a web content developer. Subjects performed directory management tasks using the 3D visualization and one of the two traditional interfaces, with ordering split evenly. The subjects timed tasks themselves by activating and then stopping a timer. Results were examined quantitatively.

Although there are many other evaluations comparing a visualization tool with a traditional interface or another visualization in the literature, these are representative of the study

* email: johng@securedecisions.avi.com

6th International Workshop on Visualization for Cyber Security 2009
11 October, Atlantic City, New Jersey, USA
978-1-4244-5413-6/09/$25.00 ©2009 IEEE

designs and methodologies used. The study design used here draws on these existing studies.

3 Study Design

The goal of this study was to compare user performance on TNV and the current state of the art tool for network analysis. This comparative evaluation followed a repeated measure within-subject design where each participant performed the same series of tasks with both of the tools. Tasks measured performance of typical network analysis tasks using both TNV and a commonly used tool for packet capture and analysis, Ethereal. The same data sets were used for each of the tools, but the order of tool usage was counterbalanced. Results were examined both quantitatively and qualitatively (based on observations of the strategy used to answer questions and exploratory tasks). The data sets and tasks were pilot tested with an undergraduate student who was familiar with both tools. The data set size was reduced as a result of this pilot testing, and the tasks were made more specific (e.g., instead of asking "which host," wording was changed to "what is the IP address of the host") to prevent possibly ambiguous questions.

3.1 Participants

Eight Information Systems undergraduate and graduate students participated in this study. Participants consisted of two females and six males, with a mean age of 28.1 (standard deviation: 3.7). All participants were familiar with the basics of computer networking and had taken at least one class in networking (mean: 2.1 classes, standard deviation: 1.4). The mean self-reported knowledge of computer networking, on a scale of 1 (lowest) to 10 was 4.6 (standard deviation: 2.3). Although participants were not domain experts in networking or intrusion detection, three participants had some experience with Ethereal.

The study tested with novice users for several reasons. First, the problems of using current tools for learning were at the core of the field study results. Novices used various strategies, but to learn the basics of networking, many of the participants discussed "playing" with various tools, including Ethereal, the tool compared in this evaluation. Because TNV was specifically designed to facilitate learning – both domain-level learning and situated learning – using novices in the evaluation targeted one of the targeted populations of TNV. Second, expert users would have extensive experience with Ethereal coming into the study, with no exposure to TNV. This experience with one tool would likely skew the results of the study, since the population would be a threat to validity, as differences in results may be due to previous tool experience rather than to the differences in the tools and tasks in the study. The last factor is a practical issue: repeated solicitations of expert users to come to the lab to participate in the study were initially met with interest, but no commitments.

3.2 Tools

This study compared two tools for network packet analysis, a visualization tool and a traditional tool. The visualization, TNV, presents packet capture data in a visually compact display that emphasizes 'local' networks, the IP space that users are most concerned with. The display is essentially split between three areas. To the left is a narrow area that displays remote hosts, in the center is the area that displays links between hosts, and the large area to the right displays local hosts (those defined as being "local" to the user), which is divided into a grid where each row represents a unique local host and each column represents a time interval, with each resulting cell color coded to the number of packets to and from that host within that time period. Bisecting the display to separately show local and remote hosts increased the scalability of the visual display, so that many more hosts can be displayed at once by dividing the available screen real estate between local and remote hosts. In addition to being able to display more hosts at a time, this partitioning also fits well with analysts' perceptions of what they deem to be important. Because local hosts are of primary concern in most analysis tasks [14], the majority of the display space is devoted to the local hosts. A time slider is the primary navigation mechanism, and there are controls for filtering and highlighting packets, drawn as triangles within each cell.

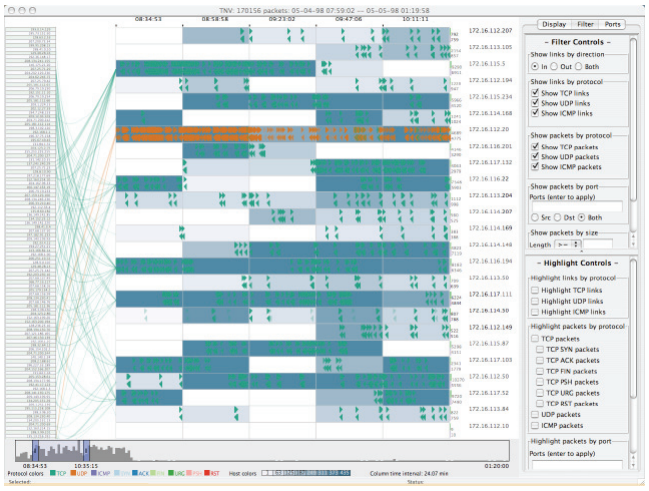

Figure 1. TNV: visual network packet analysis tool.

Ethereal, now called Wireshark, presents packet capture summary data as rows in a sortable table. Ethereal, shown in Figure 2, was chosen because of its popularity for packet capture analysis: 62% of survey respondents reported using Ethereal frequently, and another 26% reported using the tool occasionally [15].

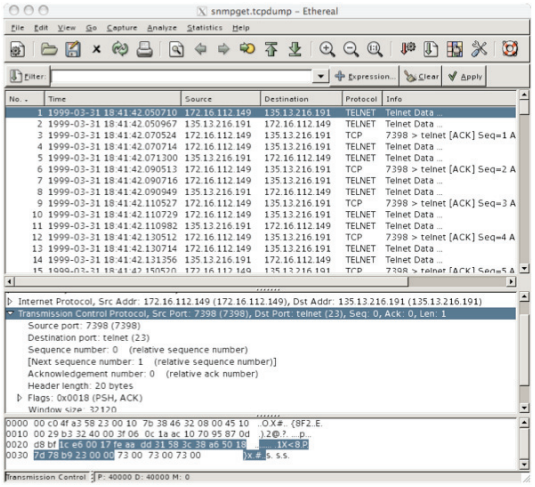

Figure 2. Ethereal: traditional network packet analysis tool.

3.3 Data Sets

Three different data sets were used, each of which was used for the same set of tasks for each tool. A very small data set was used for training. The other two data sets were subsets of the HoneyNet

Project's Scan of the Month data. The first of these consisted of 210 packets over a sixteen-hour period with eight local hosts and thirteen remote hosts. The second consisted of 762 packets over a nine-hour period with nine local hosts and eighteen remote hosts. These data sets were kept intentionally small due to results from pilot testing, in which the participant was completely overwhelmed with more than 1,000 packets using Ethereal.

3.4 Procedure

The participants each followed the following format: a) a brief introduction to the study and each of the tools and then requested to sign a consent form, b) training using either TNV or Ethereal, c) a series of timed tasks using that tool, d) training using the second tool, e) a series of timed tasks using the second tool, f) a satisfaction questionnaire was given. Half of the participants used TNV first, and the other half used Ethereal first.

During the training period the participant was first introduced to the tool. For Ethereal, the participant was given a "cheat sheet" of commonly used filters and an explanation of three statistical aggregation functions. Ethereal has a rich, but complex, filtering syntax, so providing commonly used filters was a way to minimize frustration and aid the novice users. Three of Ethereal's aggregation functions – Summary, Conversations, and Endpoints – were introduced. Summary presents an overview of the data set. Conversations lists traffic between two endpoints, and Endpoints lists traffic to and from each IP address. The aggregation functions were first briefly described, and each of these functions was then used during the training tasks. For TNV, the participant read the "Quick Start" which contained a screenshot of TNV with labels identifying each of the major functions.

The participant and the evaluator then walked through a series of tasks and associated questions for a very small data set. Answers to those questions were printed on the evaluation script, which was shared with the participant. Where there were multiple possible methods to arrive at the same answer, each of these methods was shown to the participant even if they arrived at the correct answer through a different method. For example, the first training question asked the participant to report the number of packets in the data set; using TNV there are multiple ways to answer this question, such as looking at the title bar, counting the number of packets, summing the histogram totals, or looking at the packet details. Regardless of whether the answer is correct or incorrect, all of the possible methods were introduced, with the participant driving the tool at all times. Participants could ask questions and take as much time as they needed during this training period. During this training period, participants were encouraged to think-aloud as they interacted with the system. Notes were taken on the strategies used to answer the questions and the users' think-aloud remarks.

After the training period the analyst performed a series of well-defined tasks using two different data sets for each tool. These tasks were timed and answers to these specific were recorded and scored for correctness. Tasks were timed out after five minutes, and noted as such.

Following these directed tasks, participants were asked to explore and describe the data sets in detail for each of the tools. Participants had five minutes to describe what they found interesting in the data. The data set used for this open-ended task was different for each tool depending on which tool the analysts used first; the smaller data set was used for the first tool; for example, if the analysts started with Ethereal, then they described the smaller data set using Ethereal and the larger using TNV. This was a largely arbitrary decision, but examining the same data set in detail twice would have introduced learning effects. The time to each of the insights described by participants was recorded, as were the insights themselves and the total time used (up to five minutes).

Following the training and tasks for each of the two tools, a questionnaire was administered to the participants to measure satisfaction and user perceptions. Each of the close-ended, Likert scale questions in the questionnaire was repeated for both tools.

3.5 Tasks

Typical tasks were derived from the field study to be representative of those performed during the course of ID analysis. The tasks can be divided into two broad categories:

- Well-defined tasks are directed with one possible correct answer;
- Exploratory tasks that ask subjects to draw open-ended conclusions from the data.

Each of these will be discussed in turn.

3.5.1 Well-Defined Tasks

Each participant for each tool completed ten tasks consisting of a total of sixteen questions (some tasks had multiple, related questions). The first five tasks consisting of seven questions were asked regarding the first data set, followed by five tasks consisting of nine questions regarding the second, larger data set. The tasks were of varying complexity and were chosen to be representative of the kinds of typical tasks a user would perform during network analysis. Wehrend and Lewis [16] defined multiple categories of operational tasks for information visualization tools: identify, locate, distinguish, categorize, cluster, distribution, rank, compare, within and between relationships, associate, and correlate. To simplify analysis and avoid ambiguous assignments of tasks to categories, only two of these categories were used in this study. Tasks were divided into the following high-level categories: Compare and Identify. Comparison tasks refer to those that require the user to make a judgment as to which of two or more entities are larger. These tasks tended to ask higher-level questions about the data and required the participant to use the entire data set to make a decision. Identification tasks require the user to locate and identify an entity based on its attributes. These tasks were at a lower-level and participants could often focus their attention on a small subset of the data to answer the questions. There were five comparison and five identification tasks, although some of these had multiple sub-questions.

3.5.2 Exploratory Tasks

The ability to explore and interact with data to draw meaningful conclusions is an essential activity in information visualization applications. Information visualization is "sometimes described as a way to answer questions you didn't know you had" [17]. However, it is impossible to measure the ability of a visualization tool to answer these kinds of unknown questions with predefined, directed tasks. However, exploratory tasks are very difficult to measure quantitatively. The few attempts at doing so in previous lab-based visualization evaluations are described below.

Mark, Kobsa, and Gonzalez [18], in a comparison of collaborative versus individual discovery using information visualization, described an experiment in which subjects were asked to discover as many findings in population survey data as they could. This task required no specific background. The number of findings was counted and the correctness of the findings was verified. The proportion of meaningful results within this set was determined by two independent coders judging whether the results constituted a meaningful finding, defined as

one that included a comparison between variables, indicated a minimum or maximum, and/or had a "surprise" value.

Juarez, Hendrickson, and Garrett [19] argued for the importance of measuring solution quality in their visualization evaluation framework. They acknowledged the difficulty of measuring something subjective such as quality and recommended that quality be defined for specific domains by a group of experts for each task solution. The experts in that study used a 1-10 rating scale.

In the former, each result is examined to determine if it is correct and if it is "meaningful" based on a coarse criterion of meaningfulness; no domain expertise is required because the data was generic. In the latter, domain experts rated each result on a scale of "quality." Neither of these is a perfect fit for this study. Rather, incorrect and any insights that were part of the well-defined task section were discarded, and the number of correct insights found was measured. The exploratory tasks were also examined qualitatively.

3.6 Independent and Dependent Variables

The independent variables were tool (TNV vs. Ethereal) and task type (Comparison vs. Identification). The dependent measures were accuracy (whether or not a given task answer was correct), completion time (time taken to perform a successful task), and user perceptions. User perceptions were measured through a post-test questionnaire with a 7-point Likert scale of seven questions, each of which was repeated for both tools, related to users' perceptions of satisfaction, confidence, and performance.

3.7 Hypotheses

Because the visual paradigm used is expected to be more intuitive to novice users, performance using TNV is anticipated to result in more accurate and faster responses overall than Ethereal, especially in comparison tasks. Specifically, this study examined the following hypotheses:

Accuracy:

Hypothesis 1: TNV will result in fewer errors as compared to Ethereal.

Hypothesis 1a: The advantage of fewer errors using TNV will be more pronounced in comparison tasks as compared to Ethereal.

Hypothesis 1b: There will not be a significant advantage of fewer errors using TNV in identification tasks as compared to Ethereal.

Participants were expected to be more accurate and have fewer errors across all tasks using TNV than Ethereal. Hypothesis 1 is testing for a main effect of tool, while 1a and 1b are testing the interaction effects between tool and task type. Accuracy was expected to be more pronounced for comparison tasks, but relatively comparable for identification tasks. This is expected because of Ethereal's powerful searching capability. The expected performance differences across the types of tasks are shown in Figure 3.

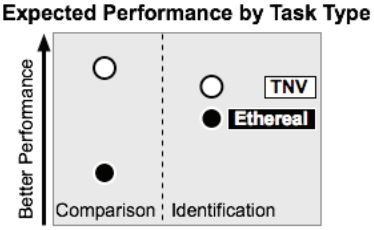

Figure 3. Expected performance by tool and task type.

Efficiency:

Hypothesis 2: TNV will result in shorter task completion times as compared to Ethereal.

Hypothesis 2a: The advantage of shorter task completion times using TNV will be more pronounced in comparison tasks as compared to Ethereal.

Hypothesis 2b: There will not be a significant advantage of shorter task completion times using TNV in identification tasks as compared to Ethereal.

Participants were expected to complete tasks faster using TNV than Ethereal. Hypothesis 2 is testing for a main effect of tool, while 2a and 2b are testing the interaction effects between tool and task type. Efficiency was expected to be more pronounced for comparison tasks, but reasonably similar for identification tasks.

Exploration:

Hypothesis 3: TNV will result in a greater number of insights during data exploration as compared to Ethereal.

During the exploratory tasks, participants are expected to perform better, measured by the number of insights discovered, using TNV.

User Perceptions:

Hypothesis 4: TNV will result in more positive user perceptions as compared to Ethereal.

Finally, participants are expected to give higher satisfaction ratings to TNV.

The order of tool usage is not expected to affect performance.

4 Results and Discussion

Participants completed the training tasks in an average of 18 minutes for TNV and 10 minutes for Ethereal. This difference could be due to greater interest in the visual tool over the textual tool. Observationally, participants were quick to learn and eager to explore TNV, and many expressed enthusiasm about using the tool. While they grasped the concepts of Ethereal, most participants had trouble understanding the tool's aggregation functions.

4.1 Well-Defined Tasks (Hypotheses 1 & 2)

The primary performance measures for the well-defined tasks were the number of correctly answered questions and the time to complete questions that were answered correctly as a function of tool and task. Each task was measured for accuracy, with a 1 indicating correct and a 0 indicating incorrect, unable to answer, or timed out. (For purposes of accuracy, each of these is treated the same.) There were a total 10 tasks consisting of 16 questions (some tasks had subtasks); subtasks were summed and divided by the number of subtasks to derive an accuracy score for tasks that consisted of multiple subtasks. All tasks were timed and analysis was conducted on time to completion for successful tasks.

Table 1 summarizes the mean and standard deviation of correct responses for each of the tools, broken down by task type, implying a trend towards more accurate performance using TNV for both types of tasks. The performance difference for comparison tasks was more pronounced than for identification tasks, but the mean for both was higher using TNV than using Ethereal.

Table 1. Mean of total number of successfully completed tasks by task type (maximum = 5), and for all individual tasks (max. = 10), and tasks including individual subtasks (max. = 16). Standard deviations are shown in parentheses.

	TNV	Ethereal
Comparison Tasks (max: 5)	4.250 (0.886)	2.750 (0.463)
Identification Tasks (max: 5)	4.458 (0.478)	3.646 (1.255)
Individual Tasks (max: 10)	8.708 (1.171)	6.396 (1.563)
Individual Subtasks (max: 16)	14.00 (1.604)	11.50 (3.251)

A repeated measures analysis of variance (RMANOVA) with repeated measures for tool (TNV, Ethereal) and task type (Comparison, Identification) was conducted. To ensure that counterbalancing the tool order usage had no effect on performance, order was treated as a between subject variable. The between subject variable of tool order was not significant in any of the tests.

As expected, there was a significant main effect of tool, $F(1,6) = 14.72$, $p = 0.009$, with the mean number of correct responses suggesting more accurate performance overall using TNV. Figure 4 plots the mean number of accurate responses across all tasks, showing a higher average number of correct responses overall using TNV. Hypothesis 1, that there would be a main effect of tool for accuracy of responses, was supported by the data; participants had significantly fewer errors using TNV than using Ethereal.

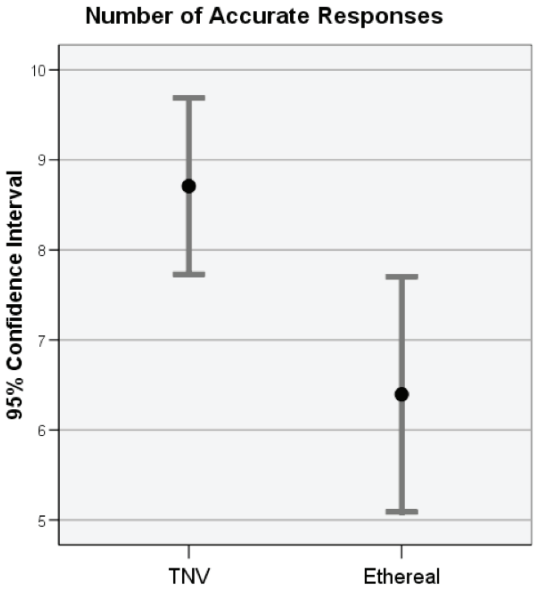

Figure 4. Mean and 95% confidence interval of accurate responses by tool. (maximum = 10)

Figure 5 plots the mean number of accurate responses by tool and task type, graphically showing a marked difference between tools for comparison tasks. Although there was no interaction effect between tool and task type – $F(1,6) = 2.139$, $p = 0.194$ – the graphical depiction of the data suggests that the differences between tools was more pronounced for comparison tasks.

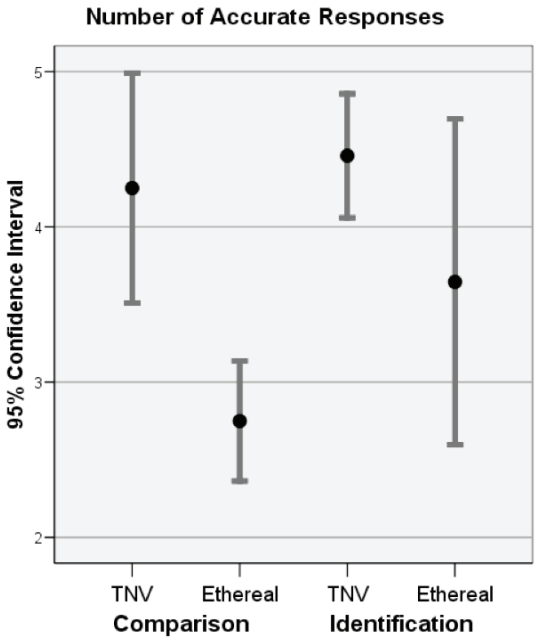

Figure 5. Mean and 95% confidence interval of accurate responses by tool and task type. (max. = 5)

To further clarify the effect of task type on performance accuracy, a paired sample t-test was conducted for the two tool/task type pairs:

- TNV Comparison vs. Ethereal Comparison: $t = 5.612$, $p = 0.001$
- TNV Identification vs. Ethereal Identification: $t = 1.860$, $p = 0.105$

Hypothesis 1a, that TNV users would be much more accurate for comparison, was not supported. However, further examination of the data stemming from the graph showing a marked difference between the tools for comparison tasks, the t-test for comparison tasks across tools was significant. Hypothesis 1b, that TNV users and Ethereal users would perform about equally in identification tasks, was supported. These results show significantly more accurate performance using TNV than Ethereal, and suggest that users performed much more accurately for comparison tasks than identification tasks across the tools.

In addition to task performance, time to complete *successful* tasks was also measured and evaluated. (A successful task is one in which all questions were answered correctly, partially correct answers were not considered successful.) Only successfully completed tasks were analyzed because incorrect responses could have been quick guesses or based on confusion of the tool. Additionally, tasks that were timed out (after five minutes) or tasks that participants gave up on were also not included in the analysis, as these could have skewed the results.

Because the tasks were of varying levels of difficulty and the average time for each task varied greatly, a standardized task completion time was computed for each task. This standardization permitted the analysis of all tasks regardless of the wide-ranging levels of difficulty and differing average times. The standardized time was computed by subtracting the average time for a task (collapsed across both tools) from the participant's time on that task and dividing the result by the standard deviation. Thus, for

each successful task for each participant the following equation was used:

Standardized_Time = (Participant_Time – Mean_Task_Time) / Task_Standard_Deviation

For example, a participant who successfully completed a task taking exactly the average time for that task would have a standardized time of 0. A participant who took two standard deviations longer than the average would have a standardized time of +2.0. Negative number indicated that a participant completed a task faster than the average, while positive numbers represent slower than average completion times. This approach is similar to the method used to compute a standardized time in an experiment evaluating two visualization described in Stasko, et al. [11]. Table 2 lists the computed average standardized task times by tool and by task type, showing a trend towards faster performance with TNV.

Table 2. Standardized average total task completion times for successfully completed tasks. Standard deviations are shown in parentheses.

	TNV	Ethereal
Avg. Comparison Time	–0.61 (0.63)	0.61 (0.94)
Avg. Identification Time	–0.03 (0.84)	0.03 (1.20)
Avg. Time	–0.19 (0.88)	0.19 (1.13)

A repeated measure ANOVA (tool and task type as repeated measure variables) with a between subject variable of tool order was conducted for standardized task completion time. There was a main effect of tool approaching significance, $F(1,6) = 5.581$, $p = 0.056$ on task performance time. (With a larger sample size, it seems likely that there would be an effect of tool for task time.) Hypothesis 2, that there would be a main effect of tool for successful task completion time, was not supported by the data, but the mean trend suggests that participants performed faster using TNV, as shown in Figure 6.

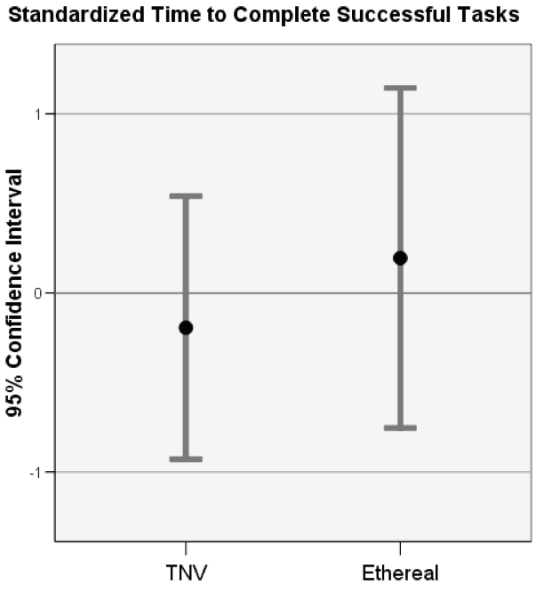

Figure 6. Mean and 95% confidence interval of standardized time to successful tasks by tool.

Figure 7 plots the mean standardized time of successfully completed responses by tool and task type, graphically showing a much more sizeable difference between tools for comparison tasks than for identification tasks, which were nearly identical. Although there was no interaction effect between tool and task type – $F(1,6) = 2.558$, $p = 0.161$– the graph suggests that the differences between tools was more striking for comparison tasks.

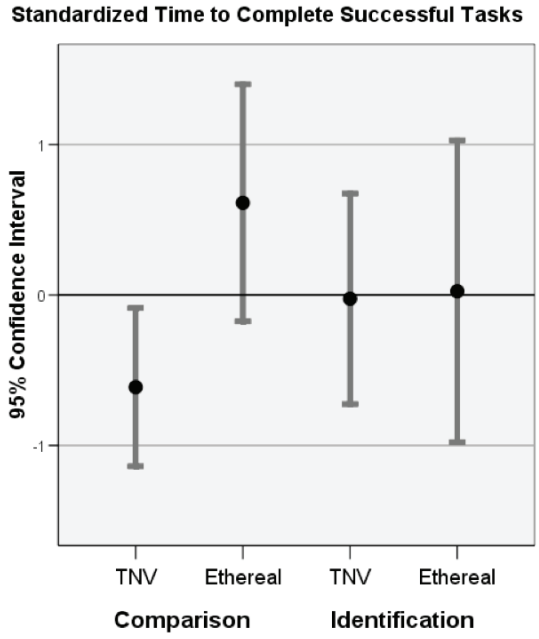

Figure 7. Mean and 95% confidence interval of standardized time to successful tasks by tool and task type.

Because of the graphical depiction of the data and to further clarify the effect of task type on performance in terms of time to complete successful tasks, a paired sample t-test was conducted for the two tool/task type pairs:

- TNV Comparison vs. Ethereal Comparison: t = –4.615, p = 0.002
- TNV Identification vs. Ethereal Identification: t = –0.085, p = 0.934

Hypothesis 2a, that TNV users would have much faster task completion times in comparison tasks, was not supported. However, graphing the data suggests a marked difference between the tools for comparison tasks, and the t-test for comparison tasks across tools was significant. Hypothesis 2b, that TNV users and Ethereal users would have similar task completion times in identification tasks, was supported. These results suggest faster performance using TNV than Ethereal, and suggest that users performed much faster for comparison tasks than identification tasks across the tools. The effect of task type on performance warrants further explanation.

While task accuracy and completion time was pronounced across tools for comparison tasks, performance – particularly completion time – was nearly identical for identification tasks. This was expected because of the sophisticated filtering functionality of Ethereal, which participants frequently used for identification tasks to filter out the noise to answer the questions. Because the data sets were relatively small, the filters removed nearly *all* non-relevant results, allowing participants to quickly answer these questions using simple filters. For example, the

average task time for all three questions making up Task 5 was 94.5 seconds (standard deviation: 75.94) across the six correct responses with Ethereal. This question was relatively easy to answer when applying a filter, which is how most of the participants answered the question quickly. The same average standard task time for TNV was 193.5 (standard deviation: 228.51) for the four correct responses with TNV, as participants could highlight the relevant traffic easily, but many times the participants did not see exactly where the highlighted packets were on the cluttered screen. This indicates that a similar filtering function that removes irrelevant traffic – as opposed to only highlighting it – may improve task times for identification tasks in TNV. This filtering functionality, mirroring the highlighting functionality present at the time of this evaluation, was added to TNV as a product of these results (as shown in Figure 1).

Unlike identification tasks, there was a marked contrast in task performance times on comparison tasks across tools, which may be explained by the strategy used by participants. The comparison tasks generally required users to make judgments of proportions of the data across the entire data set. This type of overview comparison was one of the driving factors in the design of TNV, and is generally one of the advantages of information visualization techniques. Ethereal has several statistical functions that can aid in aggregating the data to make these comparisons, but participants' who used this strategy for Ethereal comparison tasks held values in their head and mentally did math to perform the necessary further aggregation. For example, participants who were asked to determine the largest source port value for incoming traffic in the data used Ethereal's Endpoints function to see which port had the most packets, but would then switch between the TCP and UDP screens to see if there were any duplicate ports and mentally add the numbers together.

The other common strategy for comparison tasks using Ethereal was to sort the entire data set and then scroll through the data. For example, when asked which local host (the host on the predefined "home" network) had the most packets, the participants would sort by source address and estimate the rows associated with each host, then do the same when sorted by destination and try to put those two mental estimations together to answer the question. Comparison tasks were thus less accurate and also took longer using Ethereal. Using TNV, however, many of the participants would simply glance at the screen and be able to answer the question in a few seconds. Some comparison tasks, particularly those that were port related, required that participants examine a visualization panel in TNV other than the main display, which would often result in a large time increase while searching for the functionality. This caused the two comparison tasks that required port information to be answered slightly slower than other comparison tasks within TNV.

Examining tasks – and the strategies that participants used to answer the tasks – such as these in more detail can help explicate the results. The percentage of correct answers by task for each tool is shown in Figure 8. Except for Task 5, users were more or equally accurate with TNV for every task. The marked differences in the accuracy of Tasks 2, 3, and 4 require further explanation.

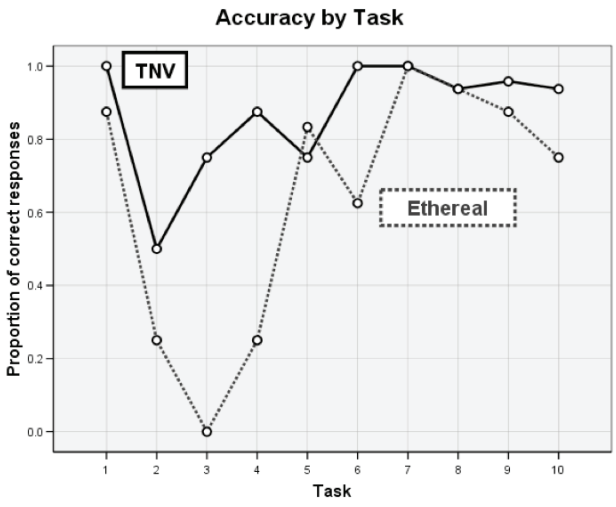

Figure 8. Percentage of correct responses by tool and task.

Tasks 2 and 3 required that the participant compare the port numbers of all packets to judge which port numbers were most prevalent. Since both tasks were asking different versions of a similar question, it would be expected that participants would perform in the same way for each tool. This was problematic with TNV, however. The problem appeared to have been an issue of visibility [20].

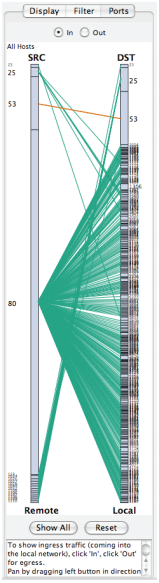

Figure 9. Port activity panel in TNV.

The port visualization, shown in Figure 9, is, by default, hidden from the user's view in TNV due to lack of screen space. The highlighting panel and the port visualization panel are collocated in the same tabbed panel, but the highlighting panel is shown by default. The user must thus remember to either choose the menu item to view all port activity for the data set or to highlight a local host and right click to choose to see the port activity associated with the host; in both cases the port visualization will automatically be given focus. The problem was that neither of these functions (main menu or popup menu) was visible, and the

port panel itself was likewise not visible. So participants ended up looking on what was displayed for something related to ports and were either unable to find it (four of eight in Task 2) or took a long time to find it (mean: 81 seconds; standard deviation: 48.85 in Task 2).

The next task required the same type of information, and participants learned from the previous experience. For Task 3, two additional participants answered correctly and the mean time was reduced from 81 to 22 seconds (four of the six correct answers were under 20 seconds). This learning process was not demonstrated using Ethereal for the same tasks.

Task 4 is defined as an Identification task, asking the user to identify the remote host that communicated with every local host. In TNV, this can be quickly accomplished through a visual inspection of the links, although the large number of links and resulting clutter make this difficult to detect. Participants generally would look for remote hosts that had multiple links, and select those hosts (thus highlighting the links and corresponding local hosts) in turn to identify the remote host that had links going to all of the local hosts. Using Ethereal, participants generally used the scrolling method, where they would sort by source address and try to identify the remote hosts that had packets going to multiple local hosts; once found, they would repeat the process. An easier method was to use the statistical aggregation functions, but the two participants who attempted this approach were still unable to answer. One participant gave up, another timed out after five minutes. This was the most difficult identification task for participants using Ethereal. Even the two correct answers took a long time, both using the scrolling method. This highlights one of the advantages of the visual search and identification capability of TNV versus the mental note and comparison strategy used for Ethereal: recognition is more accurate than recall. To minimize the user's memory load, computer interfaces should emphasize recognition of information objects rather than require users to remember them [21]. This is a recurring theme in information visualization, which performance results for this particular task highlight.

In addition to the contrasts between the accuracy in a few tasks, there are also striking differences in the time to completion for some tasks. In particular, participants performed Tasks 7 (compare directionality – incoming or outgoing – of all traffic) and 8 (compare local hosts to determine which had the most traffic) much faster when using TNV than when using Ethereal. The full details of each task and the corresponding task accuracy and completion time for each task are shown in Table 3, in the appendix.

Using TNV, all participants were successful in Task 7, a comparison task that asked participants to judge the most prevalent direction of traffic. Participants completed this task using TNV in an average of 5.63 seconds (standard deviation: 5.37). Using Ethereal, five participants were successful in Task 7, which was completed in an average of 59.2 seconds (standard deviation: 38.13). For Task 8, TNV average completion time for all eight successful responses was 5.38 seconds (standard deviation: 9.77), while average completion time for all eight successful responses using Ethereal was 68.13 seconds (standard deviation: 61.44). The strategy employed by participants for these tasks using TNV was a visual search; no manipulation of the data or the tool was required. The visual patterns were simply compared; in the case of Task 7, participants compared the number of packets pointing in both directions or the size of the incoming and outgoing histograms for all local hosts, and in Task 8, participants compared either the density of packets or the histograms for each local host. These tasks were much more difficult for participants using Ethereal. Participants either used the statistical aggregation functions or sorted the table of packets and scrolled though making mental comparisons, both of which participants had trouble performing. The visual interface made these comparison tasks, which were complicated using textual tools, relatively trivial.

4.2 Exploratory Tasks (Hypothesis 3)

In addition to performing the series of timed, well-defined tasks, participants were also asked to spend five minutes exploring to describe any insights they had into the data. This exploratory task was repeated for both tools on separate data sets. The results were mixed. One source of confusion was that the participants began explaining their perception of the tools, rather than the data itself. Once corrected, the participants had a difficult time describing what they found interesting in the data, particularly with Ethereal. Many participants would simply scroll up and down looking at the packets, but had a difficult time articulating their interpretation. Several of the participants gave up before the allotted time; one participant gave up on Ethereal after 75 seconds saying "I can't get much information from this."

In spite of this, there were several interesting trends in the exploratory tasks. Each of the unique insights was recorded and measured for correctness based on two criteria: first, was the insight technically correct, and second, was the insight one that was not derived from the previously conducted well-defined tasks.

On average, participants discovered more insights into the data and spent a longer time exploring using TNV. This was expected, as information visualization tools encourage exploration and have the ability to yield insights into the data that would not be found otherwise. A two-tailed paired sample t-test demonstrated that this difference in the number of insights was significant, $t = 2.986$, $p = 0.020$. Figure 10 plots the number of insights found by tool. Hypothesis 3, that participants would discover more insights with TNV than Ethereal, was supported. In addition to the number of insights, the kind of insights participants reported revealed a pattern.

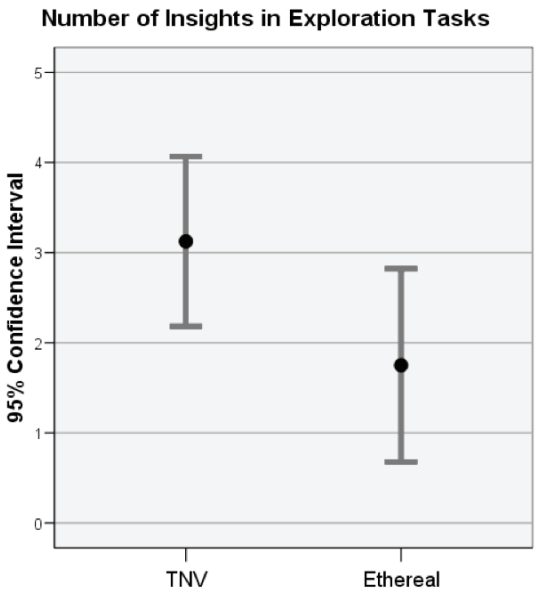

Figure 10. Mean and 95% confidence interval of the number of insights discovered.

Not surprisingly, exploration of the data using Ethereal tended to be at a lower level, while using TNV tended to be at a higher-level. For example, one participant noticed that a certain remote host was trying and failing to login over FTP to multiple hosts using Ethereal. The same participant using TNV reported that there was a substantial gap for a certain period of time in the data. Another participant using Ethereal noticed that there were clear text passwords (i.e., passwords that were not encrypted, but sent over the network "in the clear") in the FTP and Telnet packet data, and noted the most commonly used port for incoming traffic using TNV.

These different levels of insights reflect both the strengths and the visibility of functionality of each of the tools. TNV was designed to provide a "big picture" view of the data, while allowing for low-level detailed analysis, but the main screen does not provide these details. This was a change from earlier iterations of TNV, in which the details were integrated into the display. This was changed, however, to increase the amount of available screen space for the visualization, thus moving the details into a new window. This allowed for the maximum possible space to be used by both the visual overview and the textual details, but meant that the user had to flip back and forth between them when using a single monitor, as was used for the evaluation. On the other hand, this separation into two separate windows permits using dual monitors to enable both the details and the visualization to be present simultaneously with no overlapping. Additionally, the user must actively choose to examine the details, which only one participant did while exploring with TNV. By contrast, Ethereal excels at row-by-row detailed analysis and the tabular structure of the main screen reflects this. There are several useful aggregation functions that help provide an overview of the data, but these require the user actively choose the menu item. In both cases, participants generally adhered to the level of analysis that is most visible by default instead of choosing the more hidden, but more appropriate, functionality. It is expected that as the users became more familiar with each of the tools' functionalities, they would be able to make better progress in exploratory type tasks.

Lacking domain expertise probably also contributed to the small number of insights participants reported about the data; particularly with Ethereal, they were simply unsure of what they were looking at. Using TNV, at least, they could see visual patterns and anomalies, even if they were unable to articulate the exact meanings of those trends.

4.3 User Perceptions (Hypothesis 4)

The tendency for greater success and faster task completion times for TNV suggests that TNV is easier to learn than Ethereal. Self-reported user perception data related to ease of learning supports this, the mean score for TNV was 6.0 compared to 3.1 for Ethereal on a 7-point Likert scale. (For consistency, the scales were reversed here on one question that was worded negatively; all scores have a minimum of 1, strongly disagree, and a maximum of 7, strongly agree.) Figure 11 plots the mean response for each of the questionnaire results, clearly showing that participants favored TNV over Ethereal.

Figure 11. Mean satisfaction ratings on 7-point scale by tool.

A two-tailed paired samples t-test was conducted for each of the pairs of ratings for TNV and Ethereal. All pairs were significantly different:

- Overall satisfaction: $t = -5.333$, $p = 0.001$
- Ease of information processing: $t = -6.565$, $p < 0.000$
- Ease of searching: $t = -3.365$, $p = 0.012$
- Ease of learning: $t = -5.578$, $p = 0.001$
- Ease of seeing patterns: $t = -20.579$, $p < 0.000$
- Perceived level of confidence: $t = -3.813$, $p = 0.007$
- Perceived level of performance: $t = -5.118$, $p = 0.001$

Hypothesis 4, that TNV will result in more positive user satisfaction perceptions as compared to Ethereal, was supported for all given measures of satisfaction. This was expected due to the generally intuitive nature of the visualization techniques used in TNV, as compared to the more arcane interface presented by Ethereal. However, there may have been social pressure to respond positively to TNV, since the participants knew that the evaluator was also the designer of the tool.

The user perceptions questionnaire included open-ended questions asking participants to share their thoughts on what they liked and did not like about using each of the tools. The quantitative results described above are supported by these responses. Related to searching, one participant noted that "graphics help searching and analyzing task." The ability to process information and the ease of learning TNV as compared to Ethereal was emphasized in responses such as: "TNV was intuitive and kind of fun to play with" and "TNV is more intuitive." One participant responded that: "I think TNV probably has a shorter learning curve because you can see the results of actions as soon as you click on something and most of the tools are not buried in the menu items." Conversely, one participant responded that when using Ethereal "it's extremely difficult for the novice to answer analyzing question or detect abnormal behavior."

The participants' perception of being able to see patterns was the most marked contrast between the two tools, and the open-ended responses reflected this. Responses included "TNV is definitely more helpful to find patterns" and using TNV it was "easy to recognize patterns." One participant wrote that when using TNV "I felt I was more aware of the environment – the tool made exploring activity very easy." That statement summarizes

the design goals of TNV: to provide context to enhance awareness and to encourage and support exploration. It was also hoped that TNV would support novices in learning about their environment. Although not the purpose of the evaluation, one participant commented that "I learned a lot about networks" during the study, which seems to support this design goal.

5 Lessons Learned

Future researchers who would like to include user testing as part of their design and evaluation methodology may benefit from the study design presented here. While new tools will present their own unique challenges and data sets, there are some common threads that most network security visualization tools will follow. This section highlights some of the challenges network security visualization evaluation designers may face related to users, training, data and tasks.

One problem is the difficulty in finding domain experts to serve as users. For this evaluation, one of the design aspects that was tested was the ability of novice users to grasp patterns and anomalies in networking data without having vast domain knowledge. This was an a priori study design decision because one class of the target user population is junior security analysts who are unlikely to have much domain experience. Tools that target more advanced users should include representative users in the evaluations. One way to encourage domain experts to participate in user testing is to include them in the design process early. This approach builds their stake in the design and can get potential evaluation participants excited by tool and make them eager to participate in the design.

Training is another challenging issue that evaluation designers need to be solve. Network security visualizations do not need to be intuitive, they need to be effective at supporting tasks that network security analysts must accomplish. A powerful, complex visual paradigm that can facilitate the discovery of novel patterns may require some training to understand. This training may be self-directed, as is often the case in network security [22]. But time must be set aside for training both the visualization tool and the baseline tool, if evaluation participants do not already know it. This raises a problem that is intertwined with the domain expert issue, domain experts probably already know the baseline tool, while they are likely to be completely unfamiliar with the visualization tool. This intimate knowledge of one of the independent variables can skew the results – another reason that novices were used in the study presented here – so care must be taken when selecting participants, and the results of experts may need to be analyzed separate from novices.

Another challenge is finding appropriate data sets. While there are many data sets available, many are unlabeled. That is, they do not provide ground truth and it can be difficult for the evaluation designer to identify targets of specific tasks, which leads to a related problem. Defining realistic tasks that can be answered in a pre-determined, relatively brief time period is challenging on its own. To compound the problem, tasks and data sets are intrinsically linked; any given data set may not include the data appropriate to a task. While synthetic data sets can be used, this decreases the ecological validity of the evaluation. Determining tasks and identifying appropriate data sets is an iterative process.

Identifying realistic tasks that can be solved with available data sets is a problem in evaluations that attempt to quantify traditional usability metrics, typically accuracy (correctness of a task) and efficiency (speed to complete correct tasks). Another approach, which was somewhat addressed in this evaluation, is to use open-ended tasks. In this evaluation, we found that novices had trouble in finding interesting patterns or anomalies in the data, but there were some unexpected insights that were discovered. Insight-based studies are becoming a popular technique for information visualization evaluation [23, 24]. This is based on the acknowledgement that information visualization can be very effective at exploration. However, the metrics for quantifying insights are still being developed. The approach taken in the evaluation described in this paper was to combine traditional user testing metrics with basic insight-based methods. This may be the best approach in network security visualizations, since the goals of the tool likely include both increasing the accuracy and efficiency of performing of common tasks as well as increasing users' abilities to find novel insights in the data.

Finally, especially for open-ended tasks, instructions should be very clear. One problem we found in the exploration tasks is that the participants were focused on describing the tool, not the data. This was not discovered in pilot testing, but several of the participants used this approach, which was counter to the intent of the study design.

6 Conclusion

This paper has presented a comparative evaluation between an information visualization tool and a textual tool for network packet analysis. In support of Hypothesis 1, participants performed significantly more accurately using TNV than Ethereal in the well-defined tasks. Concerning Hypothesis 2, time to successful completion across tools was approaching significance, but the means suggest faster performance using TNV. Although there was no interaction effect between tool and task type for accuracy or task completion time, graphing the data suggests better performance for comparison tasks. In support of Hypothesis 3, participants discovered a higher number of insights using TNV than Ethereal in exploratory tasks. In support of Hypothesis 4, participants clearly preferred TNV to Ethereal, most strikingly in the perceived ease of seeing patterns and anomalies in the data.

Especially for novice users attempting to learn the domain and situated knowledge needed to accomplish intrusion detection analysis, this evaluation has been in a first step in validating the visual approach used in TNV as compared to the near ubiquitous network analysis tool used today. This was especially true for making accurate and timely judgments of proportion for given attributes. Those who already have expertise in Ethereal may also benefit from the added context that TNV provides, but this requires further study.

While there are some limitations in the study – the relatively small sample size, although this is consistent with much of the information visualization literature, and the use of novices, although several reasons for this were enumerated – it is important that the visualization for cyber security community begin to incorporate user testing into the design process. This is a first step towards encouraging that in future work by outlining a potential methodology for testing.

Acknowledgements

The author wishes to thank Liewei Dai and Medha Umarji for their suggestions and feedback on this research.

References

[1] G. Conti, and K. Abdullah, "Passive visual fingerprinting of network attack tools." pp. 45-54.

[2] R. F. Erbacher, K. L. Walker, and D. A. Frincke, “Intrusion and Misuse Detection in Large-Scale Systems,” *IEEE Computer Graphics and Applications,* vol. 22, no. 1, pp. 38-48, 2002.

[3] K. Lakkaraju, R. Bearavolu, A. Slagell *et al*., "Closing-the-Loop: Discovery and Search in Security Visualizations." pp. 58-63.

[4] C. P. Lee, and J. A. Copeland, "Flowtag: a collaborative attack-analysis, reporting, and sharing tool for security researchers." pp. 103-108.

[5] J. McPherson, K.-L. Ma, P. Krystosk *et al*., "PortVis: a tool for port-based detection of security events." pp. 73-81.

[6] D. Phan, J. Gerth, M. Lee *et al*., "Visual Analysis of Network Flow Data with Timelines and Event Plots," *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*, J. R. Goodall, G. Conti and K. L. Ma, eds., pp. 85-99: Springer, 2008.

[7] W. A. Pike, C. Scherrer, and S. Zabriskie, "Putting Security in Context: Visual Correlation of Network Activity with Real-World Information," *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*, J. R. Goodall, G. Conti and K. L. Ma, eds., pp. 203-220: Springer, 2008.

[8] J. R. Goodall, W. G. Lutters, P. Rheingans *et al*., “Focusing on Context in Network Traffic Analysis,” *IEEE Computer Graphics and Applications,* vol. 26, no. 2, pp. 72-80, 2006.

[9] A. Komlodi, A. Sears, and E. Stanziola, *Information Visualization Evaluation Review,* UMBC-ISRC-2004-1, 2004.

[10] M. M. Sebrechts, J. V. Cugini, S. J. Laskowski *et al*., "Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces." pp. 3-10.

[11] J. Stasko, R. Catrambone, M. Guzdial *et al*., “An evaluation of space-filling information visualizations for depicting hierarchical structures,” *International Journal of Human Computer Studies,* vol. 53, no. 5, pp. 663-694, 2000.

[12] C. Plaisant, J. Grosjean, and B. B. Bederson, "Space Tree: Supporting Exploration in Large Node Link tree, Design Evolution and Empirical Evaluation." pp. 57-64.

[13] K. Risden, M. P. Czerwinski, T. Munzner *et al*., “An initial examination of ease of use for 2D and 3D information visualizations of web content,” *International Journal of Human Computer Studies,* vol. 53, no. 5, pp. 695-714, 2000.

[14] R. Ball, G. A. Fink, and C. North, "Home-centric visualization of network traffic for security administration." pp. 55-64.

[15] J. R. Goodall, “Defending the Network: Visualizing Network Traffic for Intrusion Detection Analysis,” Ph. D. Dissertation, University of Maryland-Baltimore County, Department of Information Systems, 2007.

[16] S. Wehrend, and C. Lewis, "A problem-oriented classification of visualization techniques." pp. 139-143.

[17] C. Plaisant, "The challenge of information visualization evaluation." pp. 109-116.

[18] G. Mark, A. Kobsa, and V. Gonzalez, "Do four eyes see better than two? Collaborative versus individual discovery in data visualization systems." pp. 249-255.

[19] O. Juarez, C. Hendrickson, and J. H. Garrett, Jr., "Evaluating visualizations based on the performed task." pp. 135-142.

[20] D. A. Norman, *The design of everyday things*, New York, NY: Doubleday, 1988.

[21] J. Nielsen, *Usability Engineering*, San Francisco, CA, USA: Morgan Kaufmann, 1993.

[22] J. R. Goodall, W. G. Lutters, and A. Komlodi, “Developing Expertise for Network Intrusion Detection,” *Information Technology & People,* vol. 22, no. 2, pp. 92-108, 2009.

[23] P. Saraiya, C. North, and K. Duca, "An Evaluation of Microarray Visualization Tools for Biological Insight." pp. 1-8.

[24] P. Saraiya, C. North, V. Lam *et al*., “An Insight-Based Longitudinal Study of Visual Analytics,” *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, no. 6, pp. 1511-1522, 2006.

APPENDIX

Table 3. Average task completion times in seconds for successfully completed tasks. Total number of correct responses per task indicated in parentheses (maximum = 8).

	Specific Task	TNV	Ethereal
Data Set 1 (210 Packets)	1. What protocol (ICMP, TCP, UDP) is most predominant in the data?	16.5 (8)	40.9 (7)
	2. Which source port number is the most active for incoming (ingress) traffic in this data set?	81.0 (4)	167.5 (2)
	3. Which source port number is the most active for outgoing (egress) traffic in this data set?	20.8 (6)	- (0)
	4. Several of the remote hosts scanned more than one of the hosts on the local network, but which remote host scanned every local host?	45.6 (7)	166.5 (2)
	5. There is a burst of FTP (ports 20 and 21) traffic in this data set.		
	5a. About what time did this burst of activity begin?	106.5 (4)	43.5 (6)
	5b. During this burst there are a number of large packets (a length of 1500). What local host is associated with these large packets?	50.7 (7)	54.9 (7)
	5c. What remote IP address is associated with this traffic?	59.0 (7)	3.0 (7)
Data Set 2 (762 Packets)	7. What direction (incoming or outgoing) is most of the traffic going?	5.6 (8)	59.2 (5)
	8. Which local host has the most traffic?	5.4 (8)	68.1 (8)
	9. Several remote hosts scanned the entire local network (communicating with all local hosts).		
	9a. Of those remote hosts that scanned the entire local network, which remote host scanned the entire network using ICMP packets?	21.9 (8)	16.9 (8)
	9b. About what time of day did this ICMP scan start?	12.1 (7)	5.3 (7)
	10. Only three local hosts sent outgoing packets to remote hosts.		
	10a. Which three local hosts sent outgoing packets?	16.0 (8)	55.1 (7)
	10b. Of those, which was the only local host that sent out UDP packets?	9.7 (7)	29.7 (7)
	10c. Which remote host were these packets sent to?	19.8 (8)	16.4 (7)
	11. At the end of this data set a number of UDP packets came into the local network to a group of local hosts.		
	11a. Did they all originate from the same remote host? (yes or no)	30.0 (8)	17.8 (5)
	11b. All of these packets had the same data payload – what was it?	15.1 (7)	24.3 (7)