Original Article

# An evaluation of visual and textual network analysis tools

John R. Goodall

Oak Ridge National Laboratory, One Bethel Valley Road, PO Box 2008, Building 5300, MS-6418, Oak Ridge, Tennessee, USA
E-mail : jgoodall@ornl.gov

**Abstract** User testing is an integral component of user-centered design, but has only rarely been applied to visualization for cyber security applications. This article presents the results of a comparative evaluation between a visualization-based application and a more traditional, table-based application for analyzing computer network packet captures. We conducted this evaluation as part of the user-centered design process. Participants performed both structured, well-defined tasks and exploratory, open-ended tasks with both tools. We measured accuracy and efficiency 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 design presented may be useful for future researchers performing user testing on visualization for cyber security applications.
*Information Visualization* (2011) **00,** 1–13. doi:10.1057/ivs.2011.2

**Keywords:** user testing; comparative evaluation; security visualization; user-centered design

## 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 designs of a few of these tools are based in the challenges that real-world users face, these tools are rarely tested empirically. User testing for information visualization in general is difficult; for VizSec applications it is even more so. Domain users are tremendously busy, often do not want to (or are not allowed to) discuss their work, and are difficult to recruit. Defining real-world tasks and identifying appropriate data sets is challenging in other domains, but perhaps more so here; little work is published on the work practice of security experts and publicly available data sets are scarce. Cyber security experts use a myriad of tools, many of which are custom built or heavily customized, making comparative evaluations difficult.

This article attempts to define a study design that is applicable to user testing new visualization tools for network analysis. The study design is a within-subject comparative evaluation of a VizSec application with a commonly used network traffic analysis tool using both pre-defined and open-ended tasks. The contribution of this article is a first effort in evaluating the efficacy of VizSec applications and presenting a methodology for future VizSec researchers to follow.

TNV is a visualization tool designed to facilitate the analysis of network packet capture data.[8] The tool was designed to support the *ad hoc* analysis

Received: 27 May 2010
Revised: 14 January 2011
Accepted: 18 January 2011

of networking data captured as part of the intrusion detection process. We followed usability guidelines during design, 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 article 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.

Each of these evaluation categories can be useful in VizSec research. Comparing potential visualization techniques could be done early in the design process to determine the most useful mappings of networking data. Usability testing could be done throughout the development life cycle to identify workflow or other issues with the tool. Comparative evaluations can validate the effectiveness of a new visualization approach to commonly used tools. *In situ* field testing can help uncover disconnects between user requirements and implementation. The evaluation described in this article falls into the third category, controlled experiments comparing two tools.

## 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 two-dimensional and three-dimensional 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 16 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 *et al*[12] also compared methods for viewing hierarchical data. They used a within-subject design to compare a traditional interface (Windows Explorer), the Hyperbolic 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 three-dimensional and traditional browsers for directory management typical of a web content developer. Subjects performed directory management tasks using the three-dimensional 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 designs and methodologies used. The study design used here draws on these existing studies.

## 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. These tasks were derived from prior field work that included contextual interviews with security analysts.[14] 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).

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. Accuracy and efficiency 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. During the exploratory tasks, participants are expected to perform better, measured by the number of insights discovered, using TNV. Finally, participants are expected to give higher satisfaction ratings to TNV. The order of tool usage is not expected to affect performance.

The data sets and tasks were pilot tested by an undergraduate student familiar with both tools. The data set size was reduced as a result of this pilot testing, and the tasks were made more specific to prevent potentially ambiguous questions.

## Participants

Eight Information Systems undergraduate and graduate students participated in this study. Participants consisted of two women and six men, with a mean age of 28.1 (standard deviation, sd: 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, sd: 1.4). The mean self-reported knowledge of computer networking, on a scale of 1 (lowest) to 10, was 4.6 (sd: 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 use of 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 field study participants discussed 'playing' with various tools, including Ethereal, the tool compared in this evaluation. Because TNV was designed to facilitate learning – both domain-level learning and situated learning (that is, learning about the environment and how network patterns are manifested within that environment) – using novices in the evaluation targeted one of the targeted populations of TNV. However, while Ethereal is the standard for packet capture and analysis, it is packed with functionality and may not be the easiest tool to learn in the limited time given; because of this, participants were given a 'cheat sheet' of the common commands needed to accomplish the tasks. 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, because 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: network defense analysts work under tremendous time pressures. They simply have little free time.

## 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 interested in. Ethereal is an open source tool that has long been the standard for network packet analysis.

### *TNV*

The design of TNV, shown in Figure 1, is grounded in the work practice of security analysts, specifically in the practice of detecting intrusions into a network.[14] This section highlights some of the design rationale and describes the system.

The need for identifying and retaining context when performing analysis is a recurring theme with network security analysts. Analysts are rarely able to make a decision about a security event (for example, an Intrusion Detection System (IDS) alert) based solely on the data embedded in the event artifact (for example, text of IDS alert). Instead, analysts supplement this event data with data collected from other sources, often captured on an *ad hoc* basis. Analysts also rely heavily on their own knowledge and experience when doing analysis. This contextual information, based on supplemental data sources or analyst's knowledge, is needed to provide situational awareness and properly diagnose an event. This includes determining the root cause of the event, reconstructing the event's timeline and identifying any related outcomes. In order to make these determinations, analysts must fuse together the details of the event itself with the larger, surrounding context of activity. However, many of the current tools that analysts use to collect and investigate this important contextual data often take a micro-level approach and do not support developing a coherent understanding of the surrounding context. Other tools offer higher-level aggregations that can provide context for an event, but this requires analysts to continually refocus between these higher-level tools and other tools that provide details. Analysts repeatedly discussed how they would lose context when examining the details of packets and when leaving the displays that offered the high-level context. To overcome this lack of synchronization, analysts often rely on their short-term memory to integrate low-level and high-level data without external representational support. Security analysts need to be able to examine low-level network packet data within a broader view. TNV preserves context by presenting a big picture view of the data linked to other lower-level views, both visual and textual, down to packet-level details. This allows analysts to explore network traffic at high- and low-levels simultaneously by making the contextual data surrounding an event explicit and available on one screen.

In addition to providing analysts with context for the details that they are investigating, time was found to be crucial in the analysis process. The temporal aspects of network data allow for correlation of an event across multiple systems, are easily identifiable by all data sources and are essential to deconstructing an attack. TNV emphasizes the temporal aspects of networking data by explicitly showing temporal relationships.

The display is partitioned into three areas: remote hosts, local hosts and connections between them. To the far left is a narrow area that displays remote hosts, to the right of that is the area that displays links between hosts and the

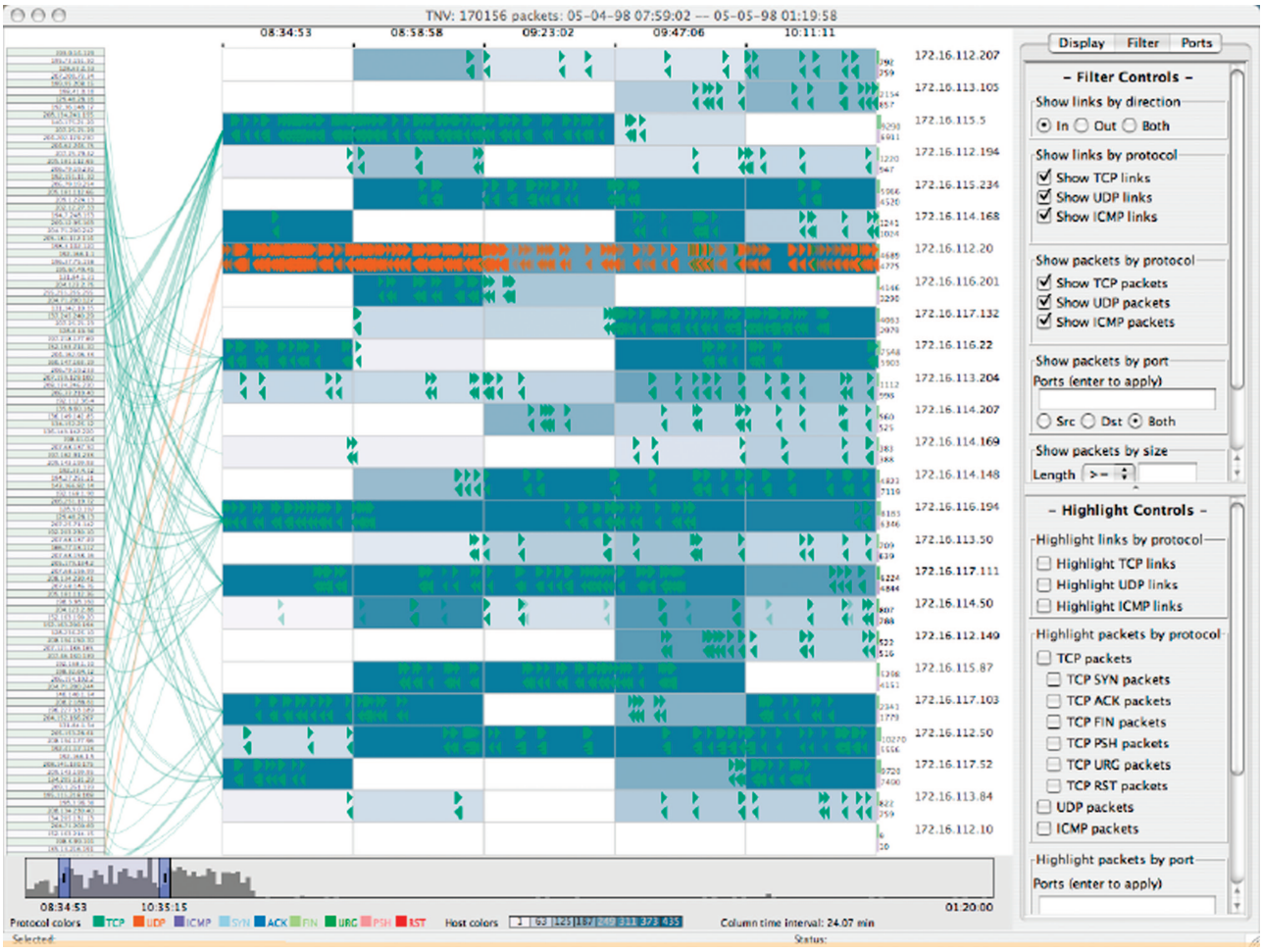


**Figure 1:** TNV: visual network packet analysis tool.

large area on the right displays local hosts (those defined as being 'local' to the user). The local hosts section 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,[15] 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. Within each local host row, individual packets are drawn as triangles, with incoming packets drawn on top of the cell and pointing to the right and outgoing packets drawn on the bottom and pointing to the left. Thus, an overview of traffic patterns can be quickly discerned by looking at the shadings within cells, and details can be seen by looking at the individual packets. Individual packet details can also be viewed in a tabular format.

*Ethereal*

Ethereal, now called Wireshark, presents packet capture summary data as rows in a sortable table. Ethereal was chosen because of its popularity for packet capture analysis: 62 per cent of respondents in a survey conducted as part of this research reported using Ethereal frequently, and another 26 per cent reported using the tool occasionally.[14]

## 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 (http://honeynet.org/challenges). The first of these

consisted of 210 packets over a 16-hour period with 8 local hosts and 13 remote hosts. The second consisted of 762 packets over a 9-hour period with 9 local hosts and 18 remote hosts. These data sets were kept intentionally small owing to results from pilot testing, in which the participant was completely overwhelmed with more than 1000 packets using Ethereal.

## 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, and therefore providing commonly used filters was a way to minimize frustration and aid the novice users. Three of Ethereal's aggregation functions were briefly described and each was then used during the training tasks. For TNV, participants read the 'Quick Start' that 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. 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, participants performed a series of well-defined, timed tasks using two different data sets for each tool. Tasks were timed out after 5 min, and noted as such. An exploration task followed, in which participants had 5 min to explore and describe any insights into the data with each of the tools. The time to each of the insights described by participants was recorded, as were the insights themselves and the total time used (up to 5 min).

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.

## Tasks

### *Well-defined tasks*

Each participant for each tool completed 10 tasks consisting of a total of 16 questions (some tasks had multiple, related questions). (See Table 1 for a full list of tasks.) 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: comparison and identification. 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.

### *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 tasks, but exploratory tasks are difficult to measure. Mark *et al*[18] described an experiment in which subjects were asked to discover as many findings in population survey data as they could. The number of findings was counted and the correctness 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. Results were examined to determine if it is correct and 'meaningful' based on a coarse criterion of meaningfulness; no domain expertise is required because the data was generic. Juarez *et al*[19] argued for the importance of measuring solution

**Table 1:** List of specific tasks and associated data sets with average task completion times (in seconds) for successfully completed tasks

| *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) |

Total number of correct responses per task indicated in parentheses (max. = 8).

quality. 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. Neither a coarse criterion of meaningfulness nor having domain experts score on quality was a perfect fit for this study. Rather, the author reviewed each reported insight, discarded incorrect ones and any insights that were part of the well-defined task section, and counted the number of correct insights. However, no attempt was made to quantify the meaningfulness or quality of the insights. The exploratory tasks were also examined qualitatively.

### Independent and dependent variables

The independent variables were tool (TNV versus Ethereal) and task type (Comparison versus 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.

## Results and Discussion

Participants completed the training tasks in an average of 18 min for TNV and 10 min for Ethereal. This difference could be owing 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. Although they grasped the concepts of Ethereal, most participants had trouble understanding the tool's aggregation functions.

### Well-defined tasks

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.

**Table 2:** Mean of total number of successfully completed tasks by task type (maximum (max.) = 5), and for all individual tasks (max. = 10), and tasks including individual subtasks (max. = 16)

| | *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) |

Standard deviations are shown in parentheses.

(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 lists each of the tasks, associated data sets, and the results for each.

*Task accuracy*

Table 2 summarizes the mean and standard deviation of correct responses for each of the tools, broken down by task type, implying a trend toward 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 when using TNV than using Ethereal.

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 2 plots the mean number of accurate responses across all tasks, showing a higher average number of correct responses overall using TNV. Participants had significantly fewer errors using TNV than using Ethereal.

Figure 3 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.

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 versus Ethereal Comparison: $t = 5.612$, $p = 0.001$

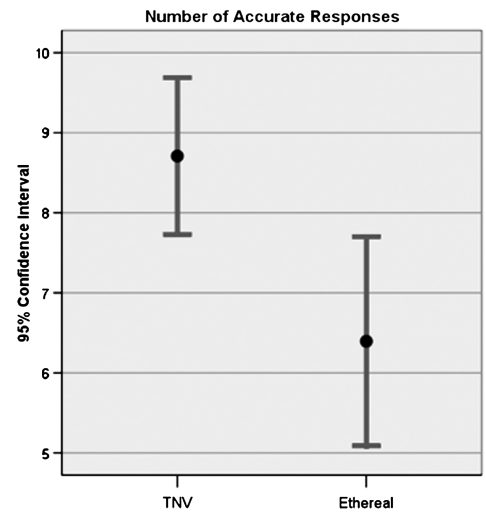


**Figure 2:** Mean and 95 per cent confidence interval of accurate responses by tool (max. = 10).

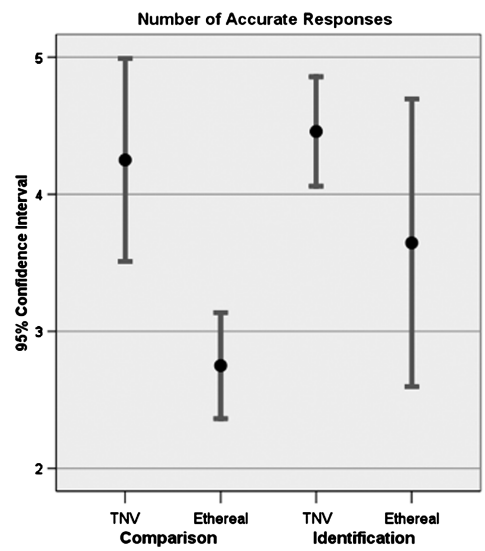


**Figure 3:** Mean and 95 per cent confidence interval of accurate responses by tool and task type (max. = 5).

- TNV Identification versus Ethereal Identification: $t = 1.860$, $p = 0.105$

These results show more accurate performance using TNV than Ethereal, and suggest that users performed much more accurately for comparison tasks than identification tasks across the tools.

*Task completion time*

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. In addition, tasks that were timed out 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:

$$\text{Standardized_Time} = (\text{Participant_Time} - \text{Mean_Task_Time}) / \text{Task_Standard_Deviation} \quad (1)$$

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, whereas 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 3 lists the computed average standardized task times by tool and by task type, showing a trend toward faster performance with TNV.

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. The mean trend suggests that participants performed faster using TNV, as shown in Figure 4.

Figure 5 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.

**Table 3:** Standardized average total task completion times for successfully completed tasks

| | *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) |

Standard deviations are shown in parentheses.

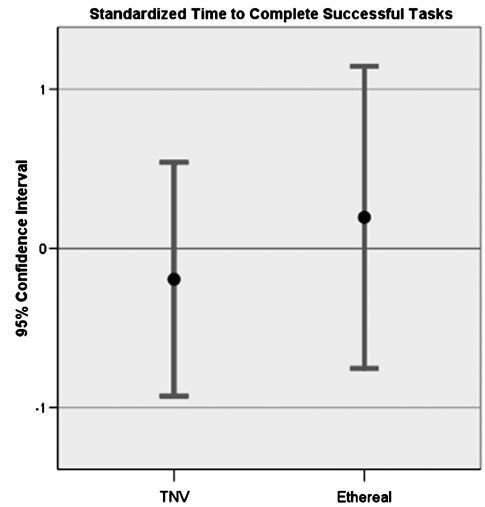


**Figure 4:** Mean and 95 per cent confidence interval of standardized time to successful tasks by tool.

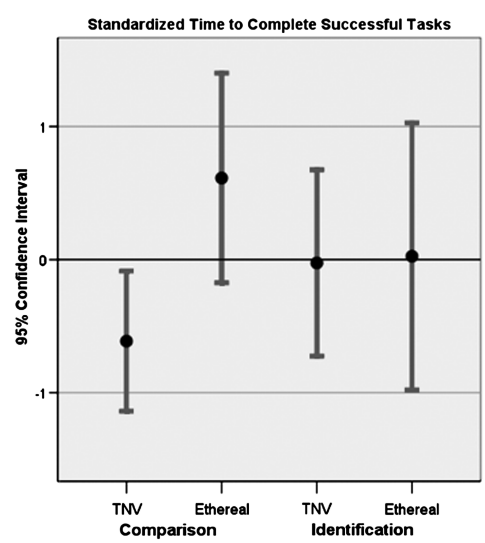


**Figure 5:** Mean and 95 per cent confidence interval of standardized time to successful tasks by tool and task type.

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 versus Ethereal Comparison: $t = -4.615$, $p = 0.002$
- TNV Identification versus Ethereal Identification: $t = -0.085$, $p = 0.934$

The *t*-test for comparison tasks across tools was significant. These results suggest faster performance using TNV than Ethereal, and suggest that participants performed faster for comparison tasks than identification tasks across

the tools. The effect of task type on performance warrants further explanation.

Although task accuracy and completion time was pronounced across tools for comparison tasks, performance – particularly completion time – was nearly identical for identification tasks. This is probably 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 (sd: 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 (sd: 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.

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 time 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.

*Selected task completion strategies*

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 6. Except for Task 5, users were more or equally accurate with TNV for every task. The differences in the accuracy of Tasks 2, 3, and 4 require further explanation.

Tasks 2 and 3 required that the participant compare the port numbers of all packets to judge which port numbers were most prevalent. As 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]

The port visualization is, by default, hidden from the user's view in TNV. The user must 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. Therefore, participants ended up looking on what was displayed for something related to ports and were either unable to find it (four

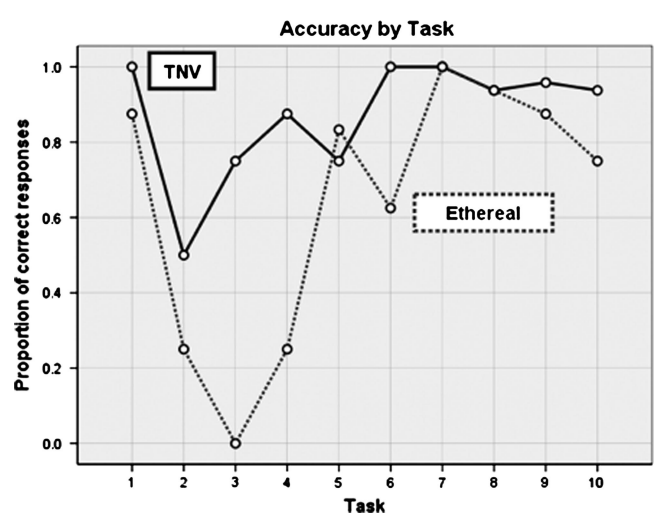


**Figure 6:** Percentage of correct responses by tool and task.

of eight in Task 2) or took a long time to find it (mean: 81 seconds; sd: 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, an Identification task, asks participants 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. 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.[21]

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. 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 (sd: 5.37). Using Ethereal, five participants were successful in Task 7, which was completed in an average of 59.2 seconds (sd: 38.13). For Task 8, TNV average completion time for all eight successful responses was 5.38 seconds (sd: 9.77), whereas average completion time for all eight successful responses using Ethereal was 68.13 seconds (sd: 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.

### Exploratory tasks

Participants were asked to spend 5 min exploring the data using a think-aloud approach 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 a higher number of 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 7 plots the number of insights found by tool. Participants discovered more insights with TNV than Ethereal. In addition to the number of insights, the kind of insights participants reported revealed a pattern.

Exploration of the data using Ethereal tended to be at a lower level, whereas 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 periods of time in the data. Another participant using Ethereal noticed that there were clear text passwords in the FTP and Telnet packets, 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, in which the details were integrated into the display. This was changed,

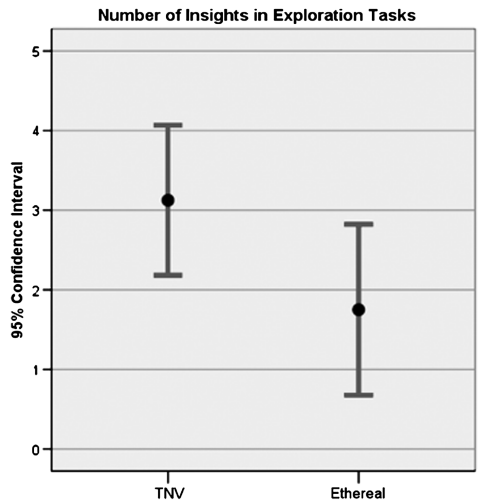


**Figure 7:** Mean and 95 per cent confidence interval of the number of insights discovered.

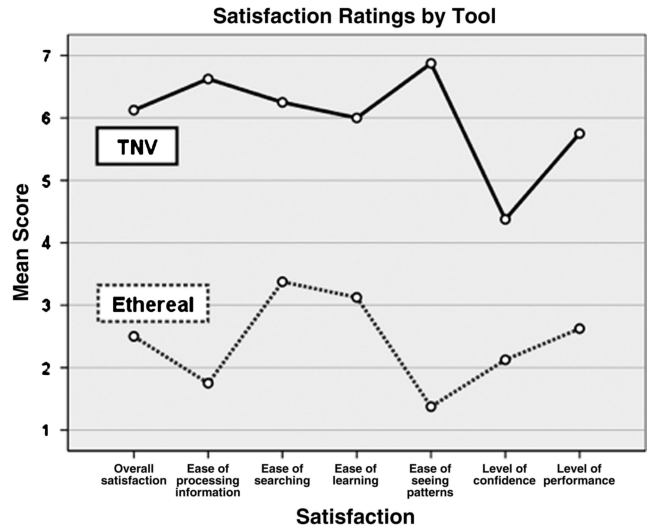


**Figure 8:** Mean satisfaction ratings on a 7-point scale by tool.

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. In addition, 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 aggregation functions to provide an overview of the data, but these require the user to 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.

### User perceptions

The tendency for greater success and faster task completion times for TNV suggests that it 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. Figure 8 plots the mean response for each of the questionnaire results, clearly showing that participants favored TNV.

A two-tailed paired samples *t*-test was conducted for each of the pairs of ratings for TNV and Ethereal:

- 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: $t = -5.118$, $p = 0.001$

Participants preferred TNV 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, as the participants knew that the evaluator was also the designer of the tool.

The 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; for example 'TNV is definitely more helpful to find 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.

## Conclusions

Researchers and developers creating new VizSec tools may benefit from the lessons learned in this effort. Although analysts are busy and can be difficult to recruit, they often are excited by technology. This is especially true for tools that are both novel and show potential for being useful. Future researchers may want to use their technology as a means of recruiting, such as giving away a license of the tool for anyone who participates in the study. Defining tasks and identifying data sets that are representative of actual practice is a challenge in all domains. Although this is certainly true for VizSec tool evaluations, researchers developing tools for network traffic analysis may be able to use the tasks and data described in this study; for other types of tools, the tasks may need to be modified substantially. Because network security often involves both searching for known patterns and exploring the data looking for things that are different from what analysts normally observe, we chose to incorporate both well-defined and exploratory tasks. However, because the participants were relatively inexperienced as network analysts, the insights identified in the exploratory task were mostly trivial. We expect that this kind of open-ended evaluation will be more effective with domain experts, but because network analysts have a tacit, situated knowledge of their own networks, this type of evaluation may be much more effective using the data from the analysts rather than giving them a data set from a network with which they have no knowledge. The need to test tools *in situ*, however, is at odds with the security policies of many organizations that allow only certain security products to interface with their data; there is no simple solution to this open issue.

This study revealed several limitations of TNV, some of which have already been incorporated into the latest version. This evaluation also revealed the strengths of both TNV and Ethereal. The ability to search and filter data easily with Ethereal led to nearly identical performance in identification tasks. Ethereal uses a robust filtering language that enables users to form complex expressions to weed out noise, it may be useful to incorporate an identical domain-specific query language domain experts already know. Taking advantage of the domain-specific query language is not just applicable to TNV, any network security visualization tool should incorporate this filtering capability. Additional user testing should also be performed. This work was a first step in demonstrating the utility of TNV to novice users, but an additional study should be performed with expert users. This may be difficult, as expert users are very difficult to solicit as participants. An additional challenge is that experts will almost certainly have experience with Ethereal. This could be somewhat mitigated by encouraging them to use TNV before the testing to ensure that they have at least some exposure to both tools.

This article has presented a comparative evaluation between TNV, an information visualization tool, and Ethereal, the de facto standard for network packet analysis. Participants performed more accurately using TNV than Ethereal in the well-defined tasks. 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. Finally, 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, contextual knowledge of a computer network needed to accomplish intrusion detection analysis,[22] 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 proportions of data. Those who already have expertise in Ethereal may also benefit from the added context that TNV provides, but this requires further study.

## References


1 Conti, G. and Abdullah, K. (2004) *Passive Visual Fingerprinting of Network Attack Tools*. VizSEC/DMSEC: Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security. New York: ACM, pp. 45–54.

2 Erbacher, R.F., Walker, K.L. and Frincke, D.A. (2002) *Intrusion and misuse detection in large-scale systems*. IEEE Computer Graphics and Applications 22(1): 38–48.

3 Lakkaraju, K. *et al* (2005) *Closing-the-Loop: Discovery and Search in Security Visualizations*. Proceedings of the IEEE Workshop on Information Assurance and Security (IAW), pp. 58–63.

4 Lee, C.P. and Copeland, J.A. (2006) *Flowtag: A Collaborative Attack-analysis, Reporting, and Sharing Tool for Security Researchers*. VizSEC: Proceedings of the 3rd International Workshop on Visualization for Computer Security. New York: ACM, pp. 103–108.

5 McPherson, J. *et al* (2004) *PortVis: A Tool for Port-based Detection of Security Events*. VizSEC/DMSEC: Proceedings of the 2004

ACM Workshop on Visualization and Data Mining for Computer Security. New York: ACM, pp. 73–81.
6 Phan, D. *et al* (2008) Visual analysis of network flow data with timelines and event plots. In: J.R. Goodall, G. Conti, and K.L. Ma (eds.) *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*. Springer, pp. 85–99.
7 Pike, W.A., Scherrer, C. and Zabriskie, S. (2008) Putting security in context: visual correlation of network activity with real-world information. In: J.R. Goodall, G. Conti, and K.L. Ma (eds.) *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*. Springer, pp. 203–220.
8 Goodall, J.R. *et al* (2006) Focusing on context in network traffic analysis. *IEEE Computer Graphics and Applications* 26(2): 72–80.
9 Komlodi, A., Sears, A. and Stanziola, E. (2003) *Information Visualization Evaluation Review*.
10 Sebrechts, M.M. *et al* (1999) *Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces*. SIGIR. ACM Press, pp. 3–10.
11 Stasko, J. *et al* (2000) An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human Computer Studies* 53(5): 663–694.
12 Plaisant, C., Grosjean, J. and Bederson, B.B. (2002) *Space Tree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation*. Proceedings of the IEEE Symposium on Information Visualization, Boston, MA. IEEE Computer Society, pp. 57–64.
13 Risden, K. *et al* (2000) An initial examination of ease of use for 2D and 3D information visualizations of web content. *International Journal of Human Computer Studies* 53(5): 695–714.
14 Goodall, J.R. (2007) *Defending the Network: Visualizing Network Traffic for Intrusion Detection Analysis*. University of Maryland-Baltimore County, Department of Information Systems.
15 Ball, R., Fink, G.A. and North, C. (2004) *Home-centric Visualization of Network Traffic for Security Administration*. VizSEC/DMSEC: Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security. New York: ACM, pp. 55–64.
16 Wehrend, S. and Lewis, C. (1990) *A Problem-oriented Classification of Visualization Techniques*. Proceedings of the IEEE Conference on Visualization Vis, pp. 139–143.
17 Plaisant, C. (2004) *The Challenge of Information Visualization Evaluation*. Proceedings of the Conference on Advanced Visual Interfaces AVI, pp. 109–116.
18 Mark, G., Kobsa, A. and Gonzalez, V. (2003) *Do Four Eyes See Better than Two? Collaborative Versus Individual Discovery in Data Visualization Systems*. Proceedings of the International Conference of Information Visualisation, pp. 249–255.
19 Juarez, O., Hendrickson, C. and Garrett Jr, J.H. (2000) *Evaluating Visualizations Based on the Performed Task*. Proceedings of the International Conference of Information Visualisation, pp. 135–142.
20 Norman, D.A. (1988) *The Design of Everyday Things*. New York: Doubleday.
21 Nielsen, J. (1993) *Usability Engineering*. San Francisco, CA: Morgan Kaufmann.
22 Goodall, J.R., Lutters, W.G. and Komlodi, A. (2009) Developing expertise for network intrusion detection. *Information Technology People* 22(2): 92–108.