Visualization is Better! A Comparative Evaluation

John Goodall

johng@securedecisions.avi.com Secure Decisions division of Applied Visions, Inc.

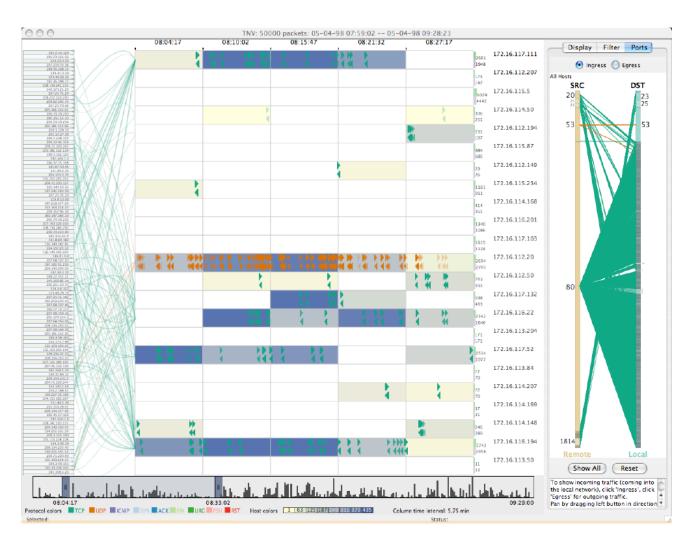




Context

- This work was part of larger research study
- Field study, interviews with security analysts, and survey to understand intrusion detection work practice
- Development of vis tool for analysis
 - Iterative heuristic reviews and usability testing
- Summative comparative evaluation

tnv



User Testing

- Controlled experiments comparing design elements: a comparison of specific widgets
- 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

User Testing

- Controlled experiments comparing design elements: a comparison of specific widgets
- 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

Study Design

- Goal: Compare tnv and the standard tool for network packet analysis
- Design: Repeated measure within subject
- Participants: 8 IS undergrad/grad students
- Tools: tnv & Ethereal
- Data: small (200 packets) & large (750 packets)
- Tasks: well-defined & exploratory

Why Novice Users?

- Learning: research showed that novices 'play' with tools to learn; tnv was designed to facilitate learning
- Background: domain experts would have lots of experience with Ethereal, which could skew the results
- Accessibility: domain experts are hard to come by

Tools

Wireshark tnv

The Wireshark Network Analyzer

60 53.550000

60 53 550000

54 53.550000

240 54,210000

189 54.250000

60 54.650000

66 54.710000

Source port: 22587 (22587)

Destination port: 110 (110)

[Next sequence number: 35

Acknowledgement number: 134

0030 7d 78 3d cc 00 00 53 54 41 54 0d 0a

Header length: 20 bytes

▶ Flags: 0x0018 (PSH, ACK)

Sequence number (tcp.seq), 4 bytes

Len Time

Elle Edit View Go Capture Analyze Statistics Help

204.252.103.16

207.183.142.87

00:20:af:92:d4:5f

08:00:4e:08:5d:56

204.252.102.2

Frame 122 (60 bytes on wire, 60 bytes captured)

□ X ⊗ □ □ → → ¬ 7 ₽

00000000.00609739b071 00000000.ffffffffffff

Internet Protocol, Src: 207.183.142.87 (207.183.142.87), Dst: 204.252.102.2 (204.252.102.2)

(relative sequence number)]

(relative ack number

Destination

207.183.142.87

207, 183, 142, 87

204.252.103.16

03:00:00:00:00:01

01:80:c2:00:00:00

207.183.142.87

Ethernet II, Src: 00:c0:4f:c7:eb:c0 (00:c0:4f:c7:eb:c0), Dst: 00:00:0c:36:00:19 (00:00:0c:36:00:19)

▼ Transmission Control Protocol, Src Port: 22587 (22587), Dst Port: 110 (110), Seq: 29, Ack: 134, Len: 6

1013 > 22 [FIN, ACK] Seq=3084 Ack=644 Win=

22 > 1013 [ACK] Seg=644 Ack=3085 Win=16384

22 > 1013 [FIN, ACK] Seq=644 Ack=3085 Win=

1013 > 22 [ACK] Seq=3085 Ack=645 Win=32256

Conf. Root = 65535/08:00:4e:08:5d:56 Cost

[Packet size limited during capture]

[Packet size limited during capture]

[Packet size limited during capture]

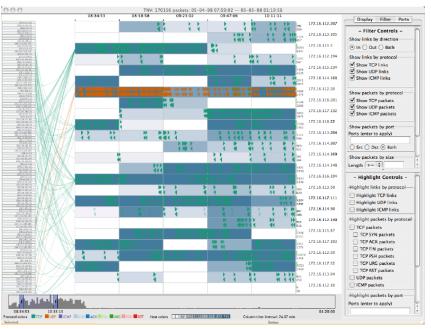
Response: +OK 2 3467

▼ 🖶 Expression... 🗞 Clear 🗸 Apply

TCP

STP

Protocol Info



0000 00 00 0c 36 00 19 00 c0 4f c7 eb c0 08 00 45 00

0010 00 2e 75 02 40 00 40 06 34 ba cf b7 8e 57 cc fc

0020 66 02 58 3b 00 6e 6a 0f a9 ba a6 bd ae 90 50 18

De facto standard for packet analysis: 88% of survey respondents used Ethereal at least occasionally (62% frequently)

...6.... O....E.

..u.@.@. 4....W...

f.X;.ni.P.

P: 3632 D: 3632 M: 0

Designed to facilitate high-level and detailed understanding of network traffic

Tasks

- Well-defined
 - Representative of 'typical' tasks; 1 correct answer
 - Task categories: comparison & identification
 - 16 tasks for each tool
- Exploratory
 - Asked participants to draw open ended conclusions from the data; no correct answer
 - Predefined time limit
 - 1 exploratory task for each tool

Tasks

- Well-defined
 - Representative of 'typical' tasks; 1 correct answer
 - Task categories: comparison & identification
 - 16 tasks for each tool
- Exploratory
 - Asked participants to draw open ended conclusions from the data; no correct answer
 - Predefined time limit
 - 1 exploratory task for each tool

Procedure

- Introduction to the study and each of the tools
- Training using either tnv or Ethereal
- Timed tasks using that tool
- Exploratory task using that tool
- Training using the second tool
- Timed tasks using the second tool
- Exploratory task using the second tool
- A satisfaction questionnaire on both tools

Variables

- Independent Variables
 - Tool: tnv, Ethereal
 - Task Type: Comparison, Identification
- Dependent Variables
 - Accuracy
 - Completion Time
 - User Perceptions

Expected Results

Expect users to perform better with tnv...

...Especially for comparison tasks, since the shows much more data at once

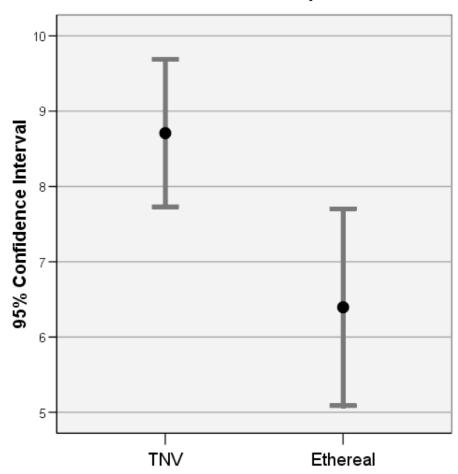
...But identification tasks will be closer, since Ethereal has easy to use search capability

Analysis

- A repeated measures analysis of variance (RMANOVA) with repeated measures for tool (tnv, Ethereal) and task type (Comparison, Identification)
- 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

Accuracy

Number of Accurate Responses



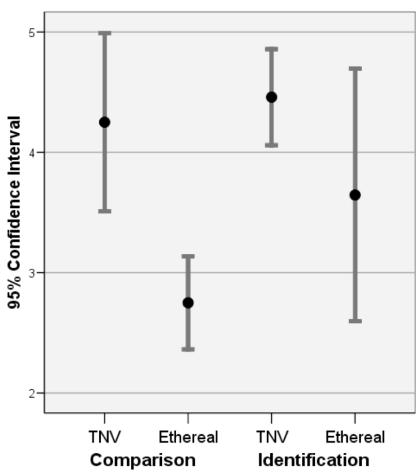
Mean and 95% confidence interval of accurate responses by tool. (maximum = 10)

Interaction effect of tool: F(1,6) = 14.72, p = 0.009

Participants had significantly fewer errors using tnv than using Ethereal

Accuracy

Number of Accurate Responses



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

Interaction effect between tool and task type:

$$F(1,6) = 2.139, p = 0.194$$

But, looking at comparison tasks for each tool, there is an effect

$$t = 5.612, p = 0.001$$

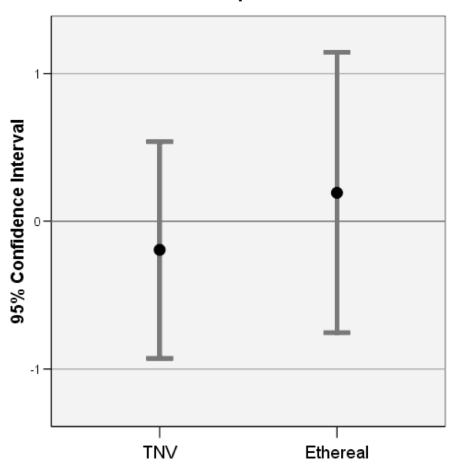
Time

- Time to completion for *successful* tasks
 - Not partially successful tasks or timed out tasks
 - Incorrect responses could have been guesses
- Standardized time
 - Tasks were of varying levels of difficulty
 - Average time for each task varied greatly
 - Negative number means faster than average

StandardizedTime = (ParticipantTime – TaskMeanTime) / TaskStandardDeviation

Time

Standardized Time to Complete Successful Tasks



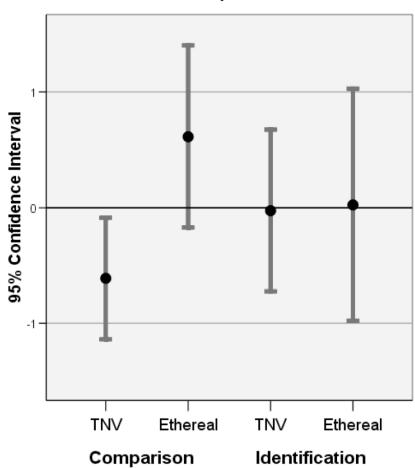
Mean and 95% confidence interval of standardized time to successful tasks by tool

Interaction effect of tool: F(1,6) = 5.581, p = 0.056

Trend suggests faster performance, but not significant

Time

Standardized Time to Complete Successful Tasks



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

Interaction effect between tool and task type

$$F(1,6) = 2.558, p = 0.161$$

But, looking at comparison tasks for each tool, there is an effect

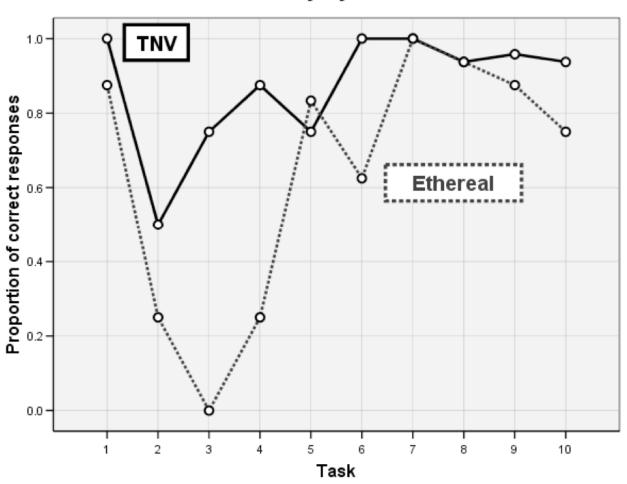
$$t = -4.615$$
, $p = 0.002$

Discussion: Task Type

- Larger difference in comparison tasks
 - Ethereal: Statistics were underused; comparisons were done by sorting and mental addition
 - tnv: Comparisons could be seen at a glance
- Less of a difference in identification tasks
 - Ethereal: Search on small data sets removed all but the relevant information
 - tnv: Search highlighted relevant information, but kept all data on the screen, so participants didn't always see where it was

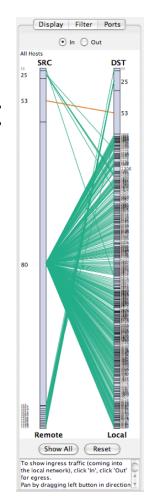
Discussion: Tasks

Accuracy by Task



Port Related Tasks

- Tasks 2, 3: compare port activity
- tnv port visualization is hidden by default
- Participants couldn't answer by looking at main display
- Participants learned in task 2, so task 3 was much faster (81 s -> 22 s)

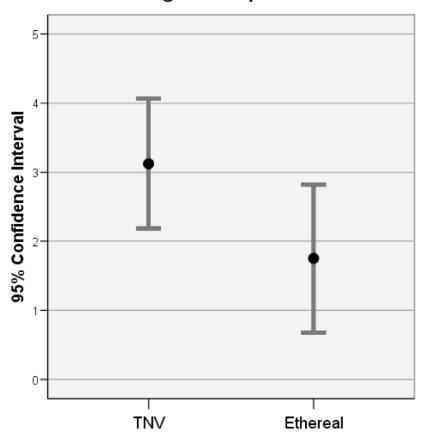


Exploratory Tasks

- Measured number of 'insights' that were not mentioned in timed tasks and not incorrect
- Results: participants often started out talking about the tools, not the data
- Several simply gave up (especially for Ethereal)

Results: Exploration

Number of Insights in Exploration Tasks

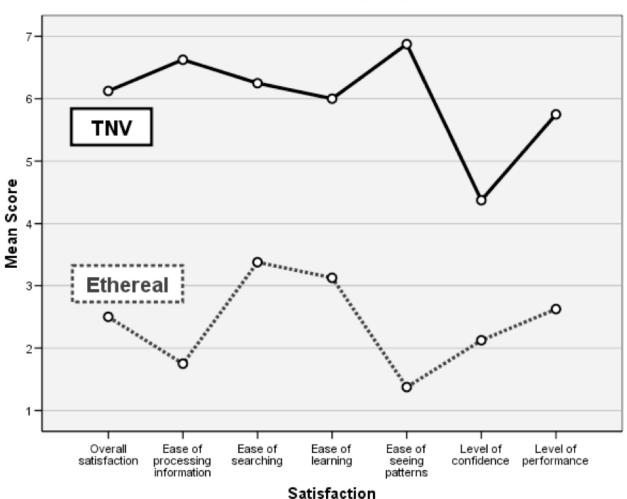


Mean and 95% confidence interval of the number of insights discovered

- tnv: higher-level
 - Gap in activity
- Ethereal: packetlevel details
 - Unencrypted passwords

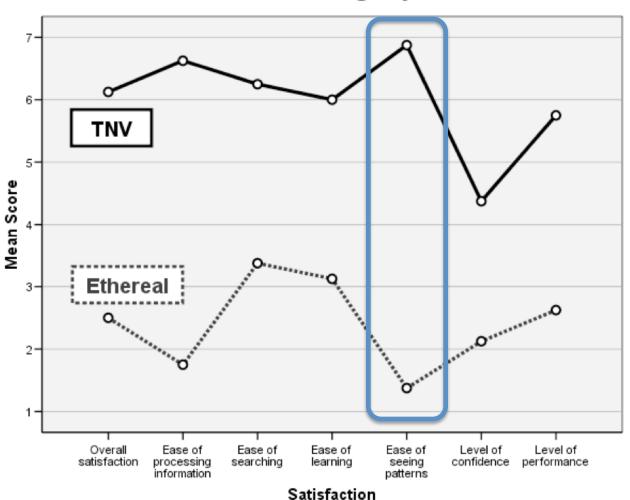
User Perceptions

Satisfaction Ratings by Tool



Ease of Seeing Patterns

Satisfaction Ratings by Tool



Lessons

- Domain experts are difficult to recruit
 - Include them in the design process
- Training can take a lot of test time
 - Self-directed training matches how analysts learn
- Data sets are problematic and unlabeled
 - http://vizsec.org/datasets/
- 'Realistic' tasks that can be answered quickly with both tools are hard to define
 - **— ???**

Questions?

John Goodall

johng@securedecisions.avi.com Secure Decisions division of Applied Visions, Inc.



