

Wright State University

CORE Scholar

[Browse all Theses and Dissertations](#)

[Theses and Dissertations](#)

2018

Comparison of Cyber Network Defense Visual Displays

Christen Elizabeth Lopez Sushereba
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all



Part of the [Industrial and Organizational Psychology Commons](#)

Repository Citation

Sushereba, Christen Elizabeth Lopez, "Comparison of Cyber Network Defense Visual Displays" (2018).
Browse all Theses and Dissertations. 1953.
https://corescholar.libraries.wright.edu/etd_all/1953

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

COMPARISON OF CYBER NETWORK
DEFENSE VISUAL DISPLAYS

A thesis submitted in partial fulfillment of the
Requirements for the degree of
Master of Science

By

CHRISTEN ELIZABETH LOPEZ SUSHEREBA
B.A., University of Dayton, 2010

2018
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

April 24, 2018

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY
SUPERVISION BY Christen Elizabeth Lopez Sushereba ENTITLED Comparison of
Cyber Network Defense Visual Displays BE ACCEPTED IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of
Science..

Kevin Bennett, PhD
Thesis Director

Debra Steele-Johnson, PhD
Chair, Psychology

Committee on Final Examination:

John Flach, PhD

Adam Bryant, PhD

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Sushereba, Christen Elizabeth Lopez. M.S. Department of Psychology, Wright State University, 2018. Comparison of Cyber Network Defense Visual Displays.

This work describes an Ecological Interface Design (EID) comparison of five displays (Alphanumeric, 2D and 3D Aggregate, Radial, and Treemap) on accuracy and latency performance for simple cyber network data analysis tasks. Twenty students from the Computer Science and Engineering Department at Wright State University participated for compensation. Questions ($n = 12$) ranged from global to specific aspects of the data and required two types of responses: numerical estimates and binary visual judgments. EID principles of attunement and specificity (Bennett & Flach, 2011) guided the interpretation of results. Participants answered faster when the display's visual forms (vertical extent, area, or angle) aligned with Cleveland's (1985) principles of graphical perception (i.e., attunement), and when the displays reflected the task structure of the question (i.e., specificity). Performance was best using the vertical extent displays. This research emphasizes the importance of using EID to create graphical displays to support cyber network defense analysts.

TABLE OF CONTENTS

I.	INTRODUCTION	1
	The Cyber Domain.....	3
	Cyber Network Monitoring.....	4
	Selected Cyber Data Visualizations	7
	Questions.....	15
	Ecological Interface Design.....	16
	Hypotheses	19
II.	METHOD	21
	Participants.....	21
	Apparatus	21
	Network Data	22
	Displays.....	22
	Procedure	23
III.	RESULTS.....	26
	Latency.....	26
	Accuracy	29
IV.	DISCUSSION	30
	Accuracy	30
	Latency.....	32
	Hypotheses Revisited.....	41

Limitations	45
V. CONCLUSION	49
VI. APPENDICES	53
Appendix A: Background Questionnaire	53
Appendix B: Counterbalanced Presentation Sequence	54
VII. REFERENCES	55

LIST OF FIGURES

Figure	Page
1. Alphanumeric display, based on Wireshark interface	9
2. Treemap display	10
3. 3D aggregate data and transmissions display	12
4. 2D aggregate data display and transmissions display	13
5. Radial display	14
6. Screenshot of the experimental software, showing the display, question, and answer-input number pad	25
7. Mean latency scores for each pair of numerical and binary questions	27
8. Mean latency for numerical questions	28
9. Summary of the significant favorable and unfavorable comparisons between displays and across questions	34
10. Number of positive and negative comparisons for each display	42
11. Treemap showing the hierarchical organization of a computer's hard drive	46
12. Hierarchically nested levels of evaluation	50

LIST OF TABLES

Table	Page
1. Experimental questions and variants	16
2. Performance of graphical displays by question	43
3. Comparison of Alphanumeric display to graphical displays	44

I. INTRODUCTION

Society is moving towards increased connectivity, with increased reliance on cyber networks. High profile attacks on the cyber networks of commercial companies (e.g., the Sony Pictures hack in 2014), nation states (e.g., the cyber attacks against Estonian government and public service websites in 2007), and financial institutions (e.g., Equifax hack in 2017; Indian Banks data breach in 2016; JPMorgan and Chase data breach in 2014) highlight the vulnerability of cyberspace and that there are many individuals interested in exploiting those vulnerabilities. To protect proprietary information, financial assets, and physical systems controlled with technology, organizations need to invest in cyber network defense (CND). CND analysts have a variety of tasks, one of which is to monitor data communications (i.e., “traffic”) for a given cyber network and determine whether there are any anomalies. This task is difficult because network traffic data is multivariate, and analysts need to correlate information from a variety of sources (e.g., intrusion detection system alerts, external websites that give information about specific signatures, “Hot IP” lists at the organization, analysts’ own memories and mental models) to determine whether activity appears to be anomalous (D’Amico, Tesone, Whitley, O’Brien, & Roth, 2008). Because of these challenges and the sheer quantity of traffic generated over cyber networks, information overload is a constant threat for CND analysts (Aschenbrenner, 2008).

A promising solution to aid CND analysts in these difficult tasks is to use information visualization to show cyber network data. Graphical displays allow for

parallel perceptual processing, which is faster than the serial processing of tabular displays of data. Parallel processing increases the efficiency of working memory, thus amplifying cognition (Goodall, 2008). There are many types of displays to assist CND analysts in the task of network traffic monitoring, but research indicates that analysts do not use them regularly. D'Amico and colleagues (2007) conducted a cognitive task analysis in which they observed CND analysts and interviewed them about their use of visualization tools. The results of the cognitive task analysis indicated that while CND analysts use some visualization tools, they are only useful for certain tasks like threat analysis and correlating activity. The analysts often used visualizations from non-CND applications, modifying their data to fit the constraints of the other visualization tool. The researchers also noted that analysts needed many different kinds of visualizations, because a single graph could not provide all the information the analysts needed to achieve their goals. Finally, the researchers noted that existing CND visualizations did not translate well from the laboratory to real-world use because they did not fit into existing operational workflow. The visualizations seemed to fail in keeping up with the volume of data, interfacing with other systems the analysts used, and were too complicated to learn (D'Amico, Goodall, Tesone, & Kopylec, 2007).

Many attempts to graph multivariate data fail because the designers do not account for the needs of the analyst and the constraints of the work domain (i.e., CND). Ecological interface design (EID; Rasmussen & Vicente, 1989) is an approach that requires careful analysis and consideration of the work domain and the agents who interact with the domain to create semantically meaningful representations of the domain in the interface. Bennett and Flach (2011) noted how many designers employ a dyadic

approach to interface design, in which the system constraints are based on the human's informational processing abilities and the display's ability to not exceed the limitations of the human. However, EID is a triadic approach, in which constraints are introduced in the human, the interface, and the work domain. The human as information processing abilities that are limited in certain ways (e.g., working memory limits) but are close to unlimited in other ways (e.g., creative problem solving). The work domain has its own set of constraints that the human operator must understand in order to engage appropriately with the domain. The goal of the interface is to couple these two elements (the human and the work domain) in a way that the meaning of the domain is represented in a way that the human can interpret.

CND is a complicated system, incorporating a deeply complex and dynamic work domain and analysts who have to interpret meaning from multivariate data coming from disparate sources. Dyadic approaches to visualization interface design seem to have failed, because CND analysts are not using them as originally intended. The purpose of this study was to compare five types of CND visualizations from an EID approach. Specifically, I evaluated how well each interface represented domain information and whether participants could decode the information accurately. In the following sections, I describe the complexities of the cyber domain and CND in more detail, introduce the displays and questions used in the experiment, and describe the principles of EID and how they apply to this experiment.

THE CYBER DOMAIN

Cyber crime is a major threat in terms of the number of attacks and the financial consequences of dealing with the results of attacks. According to the International

Business Machines Corporation's (IBM) X-Force Threat Intelligence Index for 2018, there were close to 100,000 cyber attacks in 2017 that indicated intentional malicious activity. The industries that attackers targeted the most in 2017 were financial services, information and communications, manufacturing, retail, and professional services (IBM, 2018).

According to the X-Force report (IBM, 2018), in 2017 cyber attackers used many different types of malware (i.e., malicious software designed to inflict harm in some way) like Trojans to infiltrate the networks of companies and organizations. The results of these attacks were large financial costs and loss of personal data. Another popular type of attack used ransomware to lock users' data until they paid the attackers money. Newer versions of ransomware attacks focused more on destroying data than giving it back, adding loss of important data to the growing list of the results of cyber crime.

Organization insiders also inadvertently helped many cyber attackers in 2017 in a variety of ways. For example, insiders who responded to phishing messages by clicking links or downloading attachments created openings for attackers to infiltrate their networks.

Cyber attackers also exploited the use of weak passwords, unsecured personal electronic devices, and confidential log-in credentials that were stored on open repositories to gain access to networks. The results of these crimes included large financial losses, loss of data, the compromise of confidential information, and even physical damage to network components.

CYBER NETWORK MONITORING

Analysts who work in the field of CND (also known as information assurance and information security, or InfoSec) are responsible for protecting networks against cyber

attacks. Analysts' tasks, workflows, and tools vary depending on where they work and the type of network they supervise. Some analysts focus on one type of activity like threat identification (e.g., some of the analysts studied in D'Amico et al., 2008), while other analysts work with an incident "from cradle to grave" (Gutzwiller, Hunt, & Lange, 2016).

One common task analysts engage in is monitoring network traffic (D'Amico & Whitley, 2008). The goal of this task is to identify anomalous activity on the network. Network monitoring has several subtasks, each with its own set of challenges. For analysts to identify anomalous activity, they have to determine the norms of the particular network they are monitoring (D'Amico et al., 2005). Each network is different and normal traffic patterns vary over time, so analysts must update their knowledge of the normal state for each of the networks they monitor. What is normal for one network may be completely abnormal for a different network; there are no *a priori* indications of what a normal network looks like.

Analysts use tools like automated intrusion detection systems (IDS), network sensors, router logs, etc. to help identify potentially suspicious activity (D'Amico et al., 2008), but these tools generate many false alarms. False alarms are alerts that do not actually indicate malicious activity. These alerts draw analysts' attention away from real alerts because they must determine whether an alert is likely to be true or false. For example, if an analyst classifies an alert as true, the analyst must spend time analyzing the threat and devising a strategy for threat mitigation. Depending on the nature of the perceived threat, these actions could disrupt the network. If it is a sensitive network and the threat is perceived to be severe, the analyst may have to take part of the network

offline to prevent the perceived threat from compromising more of the system. Therefore, it is very important for analysts to be accurate in classifying alerts as either true or false alarms. False alarm rates are often high, leading to information overload (D'Amico et al., 2005; Aschenbrenner, 2008). According to an analysis of cyber threats in medium to large organizations in 2012-2013, the average number of security events was 1,574,882 events per week (81,893,882 events annually). Less than 1% of those events (about 90 events per year) were true security events that required mitigating actions; the rest were false alarms (IBM, 2013). One potential use of effective CND visualizations is to help analysts determine whether an alert represents a true threat or a false alarm more quickly and accurately, limiting the negative outcomes of unwarranted mitigating actions.

After analysts identify suspicious activity that warrants additional investigation, they start fusing data from a variety of sources (e.g., different types of system logs) to find correlations and trends that could indicate whether the suspicious activity indicates a true threat (D'Amico et al., 2005). Existing tools lack the capability to show data from different sources in a single display (D'Amico et al., 2005) and link the data in a meaningful way (Best, Endert, & Kidwell, 2014). If an analyst identifies a threat, s/he must investigate its extent and attempt to resolve it (Best et al., 2014; D'Amico et al., 2008).

Many researchers and designers have attempted to develop visualization tools that would help CND analysts monitor network traffic, but analysts do not adopt these tools (Best et al., 2014) and published evaluations are varied in terms of the tools' evaluations (in terms of methods and metrics used; Staheli et al., 2014). Staheli et al. (2014)

conducted a review of visualization evaluations and suggested that evaluations using non-expert participants could fill an important gap in the literature related to cyber security visualizations. The authors suggested that future studies could break down CND tasks into component perceptual, cognitive, and motor elements to be tested on non-expert users. By breaking down these tasks into their basic components, researchers can more readily compare different displays at the basic perceptual, cognitive, or motor level; findings of such studies could then inform higher level design efforts. In this experiment, I worked to break down complex tasks related to cyber network monitoring into simple questions to test the following displays on how well they supported those perceptual and cognitive tasks.

SELECTED CYBER DATA VISUALIZATIONS

In this study, I compared five displays that have been applied to the CND domain: Alphanumeric, Treemap, 3D Aggregate data display, 2D Aggregate data display, and a Radial display. I modified each display used in this study so I could show comparable information in each display. Each display shows source and destination Internet Protocol (IP) addresses, the type of data protocol associated with each transmission (TCP, UDP, or ICMP), and the size of the data transmissions (in bytes). IP addresses represent individual machines (or “nodes”) on a network (e.g., laptops, desktops, routers, printers, etc.). IPs (or hosts) transmit data between each other (from a “source” IP to a “destination” IP). The kind of data that one host sends to another determines the protocol type. For example, the Internet Control Message Protocol (ICMP), transmits error and operational data. The Transmission Control Protocol (TCP) pieces fragmented data together, checks that the order is correct, the destination is correct, and there are no errors

in the data content. The User Datagram Protocol (UDP) is useful for real-time data transmission (e.g., live streaming) because it does not employ error-checking functions like the previous types of protocols and therefore avoids any time lags associated with error-checking functions.

I also modified the displays in terms of interactivity. Each display is most powerful when used interactively, but the purpose of this experiment is to test how well participants are able to interpret the data encoded in each display, not to determine how participants used the displays. To determine how well each display holds up to graphical perception-type tasks, I made them into static screenshots.

Alphanumeric display. The alphanumeric display is based on the format of the Wireshark interface (Wireshark Foundation, n.d.; Figure 1), pared down to present the selected categories used in the present experiment: source and destination IP addresses, the type of protocol each data packet uses (TCP, UDP, or ICMP), and the size of transmissions between two IP addresses (in bytes). Each data entry is color coded according to the protocol (note that while color is used in the Wireshark interface, the colors applied in this experiment are not the same as those employed by Wireshark; rather, these colors were chosen for aesthetic reasons.). This is a tabular display in which cyber network data is listed in individual rows and columns.

SOURCE	DESTINATION	PROTOCOL	BYTES
192.168.2.87	192.168.1.2	ICMP	26,656
192.168.1.2	192.168.2.87	ICMP	18,072
192.168.2.73	192.168.1.2	UDP	15,364
192.168.2.159	192.168.1.2	ICMP	15,076
192.168.2.58	192.168.1.2	ICMP	15,076
192.168.2.30	192.168.1.2	ICMP	15,076
192.168.2.32	192.168.1.14	ICMP	15,060
192.168.1.2	192.168.2.73	UDP	6,724
192.168.1.2	192.168.2.58	ICMP	6,668
192.168.1.2	192.168.2.159	ICMP	6,372
192.168.1.2	192.168.2.30	ICMP	6,372
192.168.1.14	192.168.2.32	ICMP	6,372
192.168.2.81	192.168.1.6	TCP	6,216
192.168.1.6	192.168.2.81	TCP	5,416
192.168.2.81	192.168.1.14	TCP	2,728
192.168.1.14	192.168.2.81	TCP	2,616
192.168.2.87	192.168.1.6	TCP	1,048
192.168.1.6	192.168.2.87	TCP	616

Figure 1. Alphanumeric display, based on Wireshark interface.

Treemap display. The Treemap display (Figure 2) was originally proposed by Shneiderman (1992) as a way of showing hierarchical relationships. For this study, the area of each square in the display represents the total amount of traffic sent from the source IP (the top IP address listed) to the destination IP (the bottom IP address listed). In this experiment, I applied the squarified algorithm (Bruls, Juizing, & van Wijk, 2000) to produce rectangles that approximate squares. The squarified algorithm is an alternative to the original “slice and dice” algorithm (Shneiderman, 1992), which produces long, thin rectangles. Kong, Heer, and Agrawala (2010) found that participants experienced more difficulty in judging the area when the rectangles in a Treemap had extreme aspect ratios, like those generated by the slice and dice algorithm. In his original proposal of the Treemap technique, Shneiderman (1992) recommended the use of color to identify additional dimensions of the represented data, such as file type or owner. In this experiment, I applied color to represent the type of protocol that is associated with each transmission.

Because it is a space-filling visualization, the Treemap takes up the same amount of space regardless of the total amount of data that it shows. Therefore, a scale was added to show how much data each square represented. The vertical scale on the left and right sides of the Treemap represent the total number of bytes transmitted. The horizontal scale along the top and bottom of the Treemap represent the proportion of the total horizontal axis that each rectangle occupies. To calculate the area of a single box, one looks at how many bytes the box's containing row represented on the vertical axis (height), then multiplies that number by the proportion of the horizontal axis the box takes up (width).

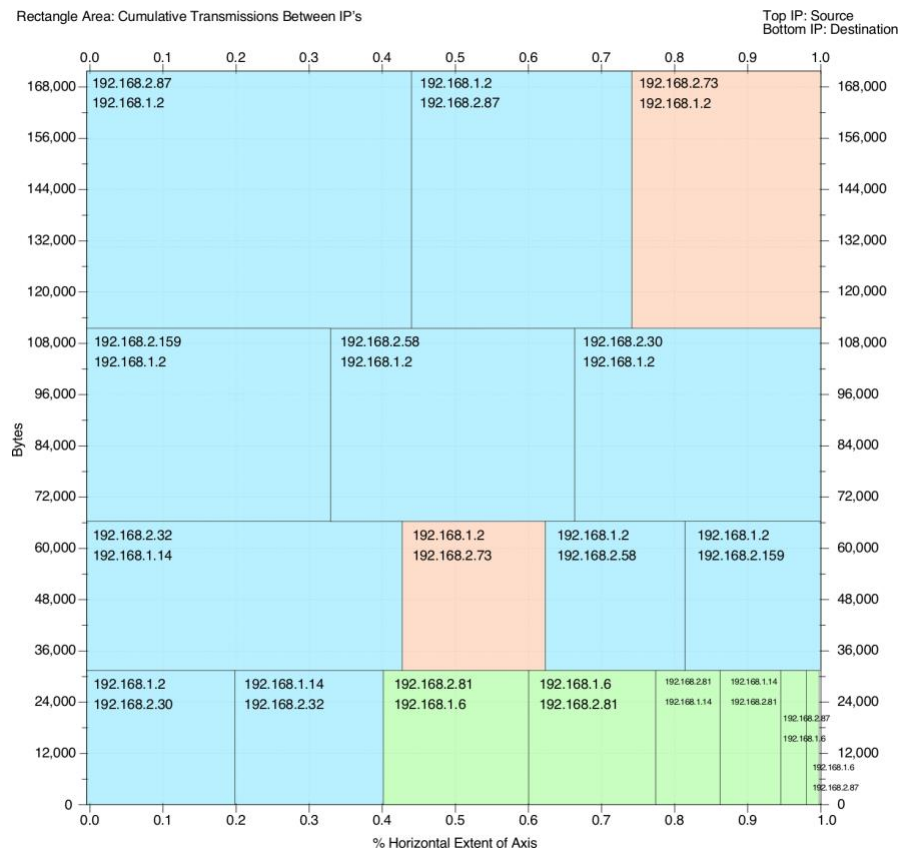


Figure 2. Treemap display

3D Aggregate display. The 3D Aggregate display was originally proposed by Bennett (2014). The display can be envisioned as a three-dimensional cube that has had its front left, front right, and top faces removed (see Figure 3). What remains are the back two faces (or data walls, left and right) and the base of the cube (front and center). The two front axes of the base are used to represent the source (left) and destination (right) IP addresses. The IP addresses are ordered by size of transmissions, with larger data transmissions appearing closer to the data walls and smaller transmissions appearing closer to the center of the display. This ordering was to prevent large columns appearing near the front of the cube and occluding smaller columns behind.

The two axes are projected inside the cube's base to form a matrix. Each cell of this matrix contains a three-dimensional column graph which represents the total amount of information that has been transmitted between a particular set of source and destination IP addresses. In contrast, the two-dimensional bar graphs located on the data walls are used to represent aggregated contributions that are specific to an individual IP. For example, all data transmitted by an individual source IP (left cube axis) is represented by a bar graph on the right data wall (located in the appropriate spot on the bottom wall axis). Furthermore, the stacked and color coded segments in both the bar and column graphs are used to represent aggregated transmission levels for each data protocol type.

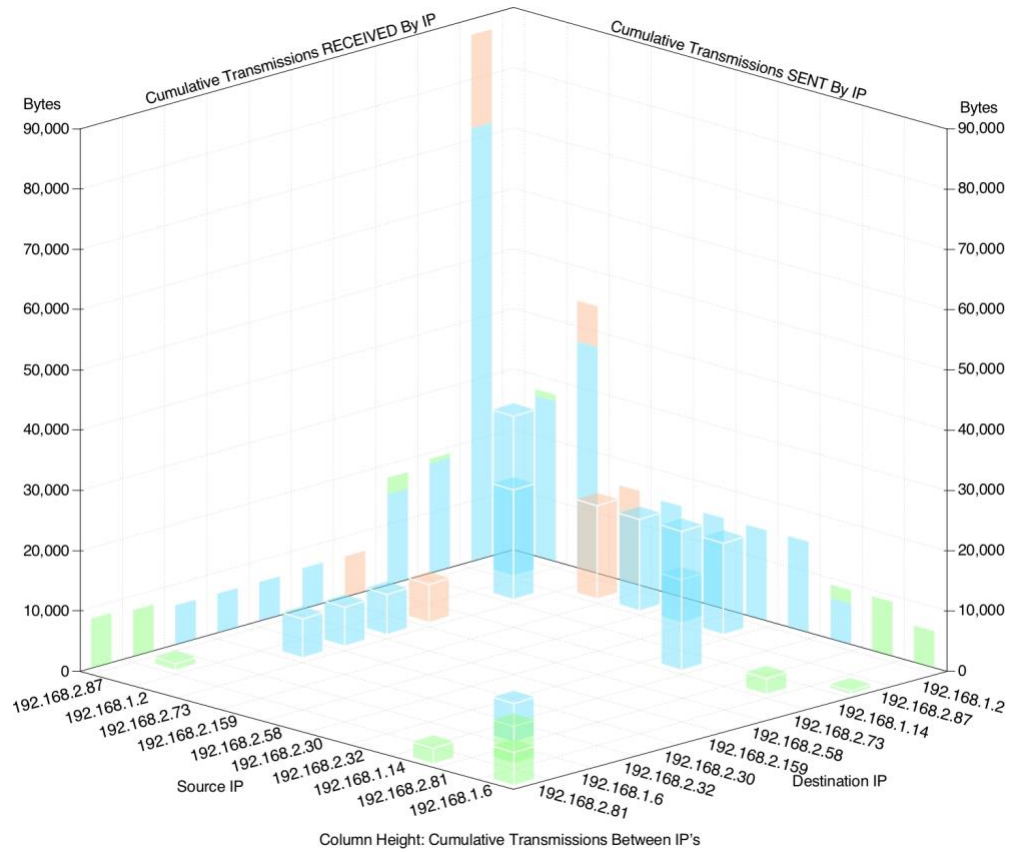


Figure 3. 3D Aggregate data and transmissions display.

2D Aggregate display. There are potential issues with 3D displays that can affect interpretability. One issue is occlusion, in which visual forms might obstruct the view of other visual forms. Another issue is the use of linear perspective, which gives the illusion of 3D space, but can inhibit accurate determinations of scale. Because of these potential issues with the 3D Aggregate display, I created the 2D Aggregate display, which is a two-dimensional version of the 3D Aggregate display (Figure 4). The primary matrix (left) represents aggregated data transmissions between source (right axis) and destination (bottom axis) IP addresses using segmented, color coded, two-dimensional bar graphs located in corresponding cells. Two additional matrices, located to the right of the primary matrix shows aggregated transmissions according to source IP (top) and

destination IP (bottom). These aggregate data matrices contain segmented, color coded, two-dimensional bar graphs that represent data transmissions that are specific to individual IP addresses. All segmenting, color coding, and ordering conventions used in the 3D aggregate display are also applied to this display.

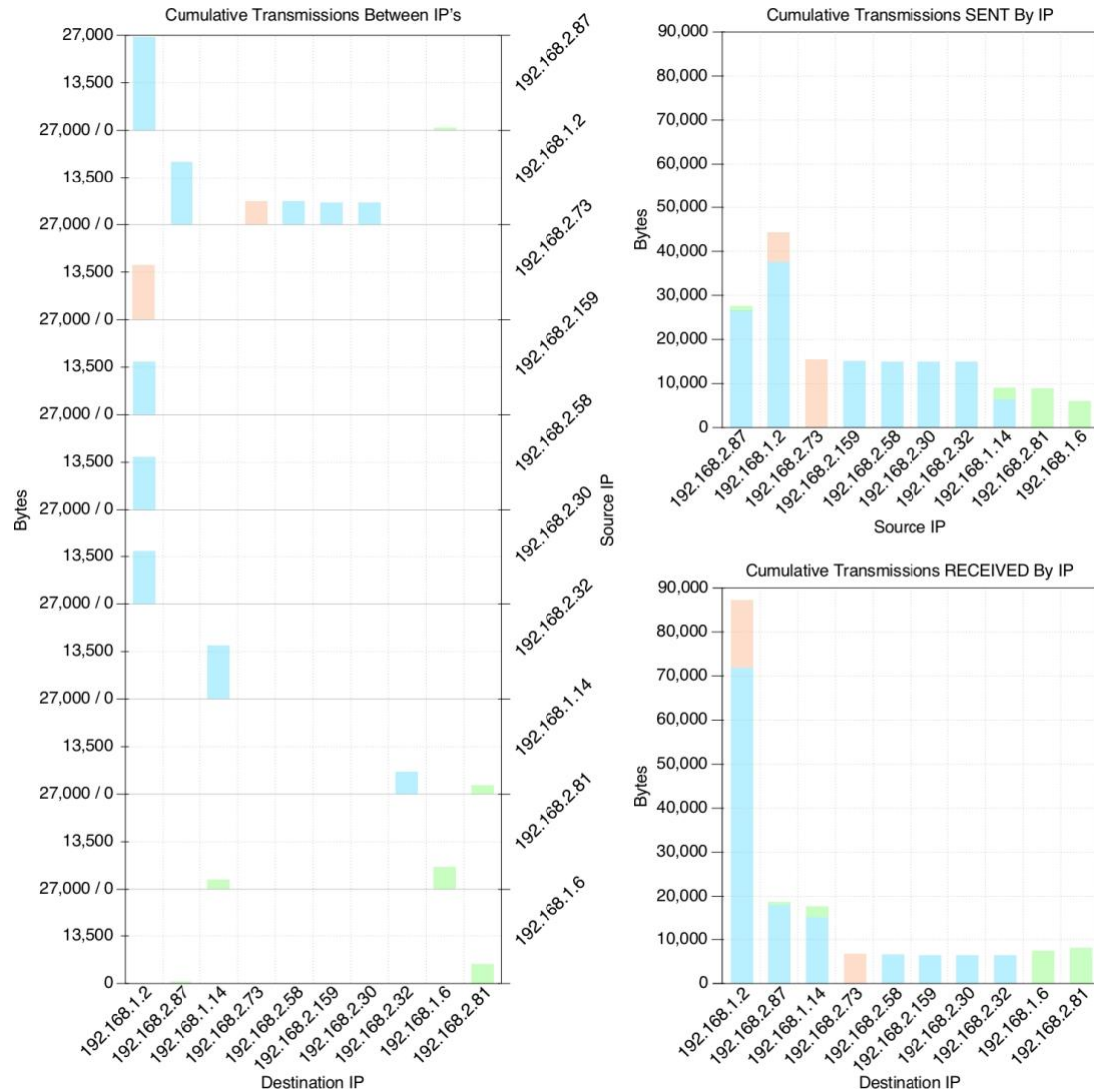


Figure 4. 2D Aggregate data display and transmissions display.

Radial display. The Radial display is based on the Radial Traffic Analyzer described in Keim, Mansmann, Schnedewind, and Schreck (2006). The innermost ring corresponds to the source IP address and the outer ring corresponds to the destination IP

address. The size of each sector corresponds to the total amount of data transmitted between two IP addresses, and color coding within the rings show the network protocols used. When possible, sectors that represent the same IP address are placed next to each other. This allows the sectors to be combined in cases when the sectors are too narrow to show IP address labels. When narrow sectors are combined, faint lines indicate the actual divides, but the IP address label is superimposed over all of the sectors that correspond to that IP address. Like the Treemap display, a scale is included that represents the total amount of bytes transmitted.

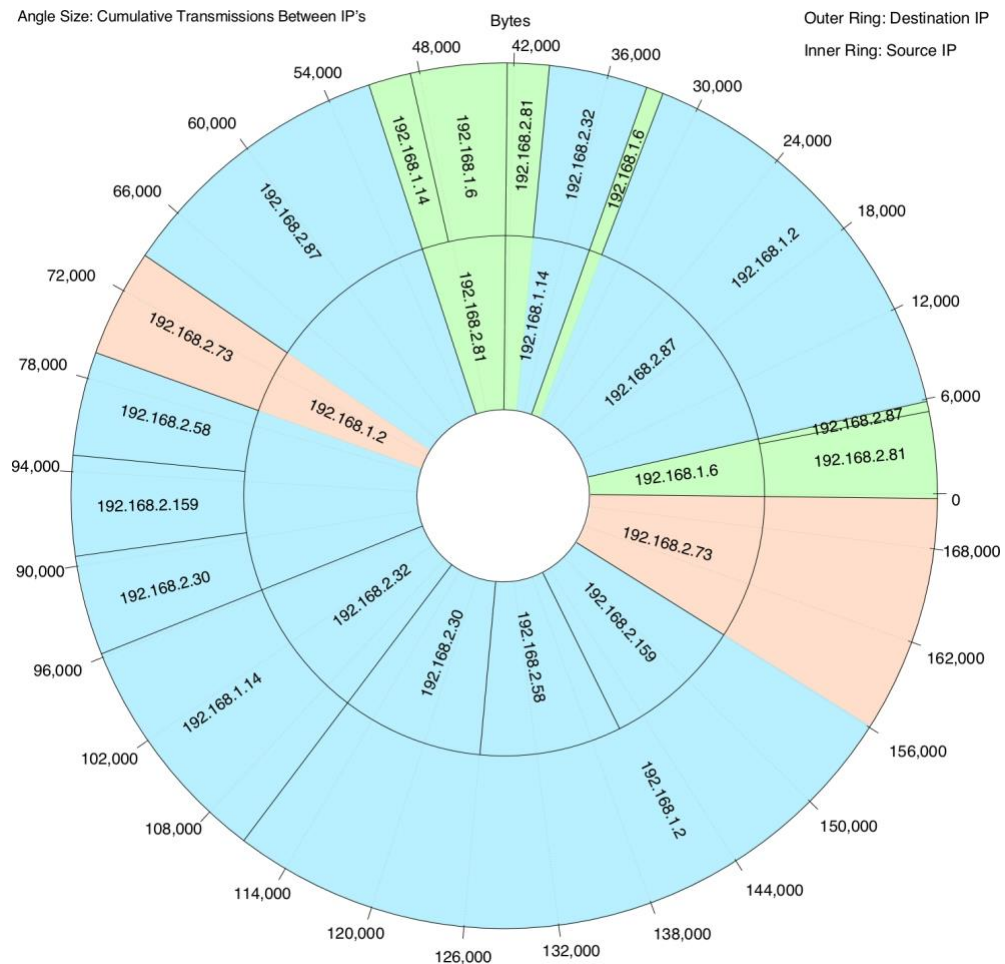


Figure 5. Radial display.

QUESTIONS

I developed twelve questions to approximate different ways in which CND analysts might be required to consider networking data. These questions were structured using a factorial combination of two dimensions (host and protocol, see Table 1). The host dimension varied the number and type of relationships between hosts (or IP addresses) that needed to be considered: 1) total transmissions across all hosts, 2) total transmissions for a single host, and 3) total transmissions exchanged between two hosts (i.e., a dyad). The protocol dimension varied the type of data that needed to be considered: 1) transmissions across all three protocol types or 2) transmissions within a single protocol.

Two different types of responses were also required. A primary purpose of analogical displays is to represent quantitative values; Questions 1-6 required the participant to provide exact numerical responses. The quality of performance on these questions will assess the effectiveness of each display in fulfilling this primary purpose (i.e., the degree to which participants can obtain the information that has been represented in an effective fashion). A second purpose of analogical displays is to support activities that require more global estimates of the information that is being represented (e.g., “spot checking” variables to make sure they are within an acceptable or anticipated range). Questions 7-12 retain the host/protocol structure, but require visual judgments and comparisons that do not need a high degree of precision and can be answered with a binary response (e.g., “yes” or “no”).

Table 1.

Experimental questions and variants. Note that text in parentheses represents variables.

Exact Values (i.e., Numerical Responses):		
<u>Host:</u>	<u>Protocol:</u>	<u>Question:</u>
All hosts	Across	1. What are the total number of bytes being (sent/received)?
	Within	2. What are the total number of (TCP/UDP/ICMP) bytes being (sent/received)?
Single host	Across	3. What are the total number of bytes being (sent/received) by (IPaddress)?
	Within	4. What are the total number of (TCP/UDP/ICMP) bytes being (sent/received) by (IPaddress)?
Host dyad	Across	5. What are the total number of bytes being sent from (IPaddress1) to (IPaddress2)?
	Within	6. What are the total number of (TCP/UDP/ICMP) bytes being sent from (IPaddress1) to (IPaddress2)?
Visual Judgments (i.e., Binary Responses):		
<u>Host:</u>	<u>Protocol:</u>	<u>Question:</u>
All hosts	Across	7. Are the total number of bytes being sent more than <value>?
	Within	8. Are there more (TCP/UDP/ICMP) bytes being sent than (TCP/UDP/ICMP)?
Single host	Across	9. Is (IPaddress) sending more bytes or receiving more bytes?
	Within	10. Is (IPaddress) sending more (TCP/UDP/ICMP) bytes or receiving more (TCP/UDP/ICMP) bytes?
Host dyad	Across	11. Is (IPaddress1) sending more bytes to, or receiving more bytes from (IPaddress2)?
	Within	12. Is (IPaddress1) sending more (TCP/UDP/ICMP) bytes to, or receiving more (TCP/UDP/ICMP) from (IPaddress2)?

ECOLOGICAL INTERFACE DESIGN

Ecological interface design (EID; Rasmussen & Vicente, 1989) is an approach that is closely related to the cognitive systems engineering (CSE; Rasmussen, 1986; Rasmussen, Pejtersen, & Goodstein, 1994) framework. While both CSE and EID focus on work in complex sociotechnical systems, EID is specifically tailored to leveraging interface elements to aid users with cognitive requirements, such as decision making, pattern recognition, and problem solving related to the work domain. According to the

EID framework, there are three elements of a system, each with their own constraints: the work domain, the agent controlling and interacting with the domain, and the interface that connects the two (Bennett & Flach, 2011). The work domain needs to be analyzed and well understood so an interface can accurately represent key constraints and limits of the domain to the user. The user needs to be understood in terms of limitations, skills, knowledge, expertise about the domain, etc. The interface then must represent the work domain in a way that falls within the limits of the perceptual and cognitive skills and abilities of the user. An ideal interface amplifies human strengths while supporting limitations to allow for proper interaction between the human and the work domain.

Because the three components (work domain, user, and interface) are tightly linked to each other, the connecting mechanisms are important to understand as well. Bennett and Flach (1992) define several of the mappings that link each of the key components. For this experiment, I am most concerned with specificity and attunement. *Specificity* links the work domain and the interface and is the extent to which the visual representations in the display accurately represent the work domain. This link represents the meaning of the domain. The link between the display and the human operator is *attunement*, which is the extent to which the visual elements in the display (e.g., emergent features) can be perceived by the user, and whether the user has the knowledge to decode the meanings of the visual representations. This link represents a user's interpretation of the domain. The goal of an interface is to represent the meaning of the work domain in a way that guides the user for how to interpret the domain. The user's actions will depend on how well the user's interpretation of the domain and the actual meaning of the domain match.

From a visual perception standpoint, Cleveland and McGill (1985) found that humans are more sensitive to certain types of emergent features over others. Emergent features are visual properties that emerge from the arrangement of graphical forms that become more meaningful than the individual graphical forms (Pomerantz & Pristach, 1989, cited in Bennett & Flach, 2011). Emergent features affect the attunement of a display. Cleveland (1985) and Cleveland and McGill (1985) did a series of studies to test basic graphical perception abilities, or how well people were able to decode graphed information using different graphical elements. They found that humans are relatively good at interpreting visual elements that show vertical extent from a common baseline (a straight line from a base, e.g., a bar graph). Humans are not as good at interpreting visual elements that rely on angles (e.g., a pie chart) or area. According to Cleveland et al.'s research, performance on each of the graphical displays in this study should vary according to the primary graphical elements used to encode the data. In other words, the graphical elements (i.e., vertical extent, angle, or area) will affect the mapping between the display elements and the participants' perceptual skills (attunement).

Specificity is also likely to vary between the displays, because each display represents the same data differently, affecting the semantic mapping between the domain and the interface. The alphanumeric display shows the data as a series of individual transmissions in a tabular format without any graphical elements to show relative magnitude. The 2D and 3D displays use graphical elements (columns) to show individual transmissions, and also aggregates the data by host (both source and destination hosts). The Treemap and Radial displays use graphical forms to show individual transmissions, which can show relative magnitude of different transmissions.

The Treemap does not aggregate any of the data. The Radial display is organized by source destination, so most transmissions associated with a single source are grouped together, providing some aggregation. Each of the types of displays represent the domain differently, which affects how the user interprets the underlying meaning that the display is trying to communicate.

The alphanumeric display is a fundamentally different type of representation than the other analogical graphical displays. Unlike the four graphical displays, there is an arbitrary relationship between visual form and underlying meaning. The differences in performance between alphanumeric and graphical displays has been investigated thoroughly (Boles & Wickens, 1987; Hanson, Payne, Shively, & Kantowitz, 1981). Alphanumeric displays are very precise and effective when the response requires the exact value of a variable or property (e.g., Bennett & Walters, 2001; Hansen, 1995). However, graphical displays are better suited to quickly determining relative magnitudes between values which is more useful for quick estimations and parallel perceptual processing (Goodall, 2008).

HYPOTHESES

I hypothesized that participants would perform differently across the five displays and across the two question types (binary and numeric response questions). In terms of differences between the displays, first I predicted that performance would follow the pattern of graphical principles that Cleveland and colleagues identified. In other words, I predicted that participants would perform better when the attunement mapping was based on vertical extent (i.e., the 2D and 3D aggregate displays) than when the attunement mapping was based on angle (i.e., Radial display) or area (i.e., Treemap).

Second, I predicted that performance would be better for displays that represented data in a manner that was coherent with the type of questions asked (i.e., work domain constraints). For example, the 2D and 3D displays show aggregate transmissions between dyads of hosts, along with total transmissions for each individual IP; the Treemap, and to a lesser degree, the Radial display, do not. Therefore, the 2D and 3D displays should yield faster and more accurate responses to questions related to information transmissions between two hosts (i.e., questions 5 and 6) and summarized transmissions for a single IP (questions 3 and 4). In contrast, the Radial and Treemap displays have numerical scales that show aggregate data transmissions between all hosts, while the 2D and 3D displays do not. Therefore, I predicted that the Radial and Treemap displays would yield better performance for Questions 1 and 2.

Third, I predicted that performance with the Alphanumeric display would differ from the graphical displays in that participants would give more precise responses (because they had access to exact numerical values). I also predicted that responses would be quicker when a small number of entries needed to be considered (e.g., only one entry will be needed for Question 6), but slower when many entries had to be added together (e.g., for Question 1 or 7 that require consideration of the entire dataset).

In terms of differences between question types (numerical estimates versus binary visual judgments), I predicted that performance would be faster for the binary questions (Questions 7-12) than for the numeric response questions (Questions 1-6). Because Questions 7-12 rely on making visual judgements instead of estimating values, they should be easier for participants to answer.

II. METHOD

PARTICIPANTS

Twenty students from the Computer Science and Engineering Department at Wright State University (3 female, 17 male) between the ages of 18 and 37 participated in this study (mean age $M = 24.60$ years, $SD = 4.99$ years). Average years of education beyond high school was $M = 3.95$, $SD = 1.36$. Participants were computer science majors ($N = 13$), computer engineering majors ($N = 4$), or enrolled in the cyber security graduate program ($N = 3$). Participants were recruited through emailed flyers distributed by the Computer Science and Engineering Department's office assistant. Participants had knowledge of basic networking concepts (protocols, packets, etc.). I recruited specifically for students who had taken a networking course or were graduate students in the cyber security program. Participants were compensated \$25 for participation. All participants had normal or corrected to normal color vision.

APPARATUS

I conducted the experiment on a general purpose laboratory computer (Apple Mac Pro, Model A1186, 3.0 GHz dual core Xenon processor, 5 GB memory, ATI Radeon HD 5770 graphics card) with a color video monitor (Apple Cinema HD Display, Model A1083, 30", 2560 by 1600 resolution, 60 Hz refresh rate) and a standard keyboard located in an enclosed experimental room. I used Adobe Director 11.5 (Adobe Systems, Inc.) software to control experimental events.

NETWORK DATA

For this experiment, I used simulated network data from the IEEE Visual Analytics Science and Technology (VAST) Mini-Challenge 2 from 2011 (the IEEE VAST Challenge, Mini-Challenge 2, 2011). This network data included a series of suspicious events including a denial of service attack, various port scans, a social engineering attack, and the addition of an undocumented machine to the internal network. For this experiment, I used the packet capture (PCAP) data file associated with the port scan that occurred on the second day of the simulated dataset. The PCAP data file contained transmitted packets of data that simulated a port scan attack. A port scan is when an attacker sends data to a host on a network using different port numbers. The open ports yield information about the types of services on the machine and the type of operating system. The PCAP dataset I used organized data transmissions by time. There were 300 time segments; I captured twelve snapshots of data spaced out across the entire dataset. I extracted data for source and destination IP addresses, size of transmissions, and protocol type. Additionally, I only used data that could be portrayed equally across all displays (e.g., not all available IP addresses were included). This ensured that each display portrayed exactly the same network data.

DISPLAYS

I evaluated the five displays that were described in Chapter 1 (i.e., Alphanumeric, Treemap, 3D Aggregate, 2D Aggregate, and Radial). I generated sixty images (12 data snapshots for each of the five displays). Each image was initially captured as a screenshot and then manually transformed into a high-resolution graphic (approximately 1470 x 1470 pixels) using Canvas Draw 3 (ACD Systems).

PROCEDURE

Participants signed up for individual two-hour experimental sessions. Upon arrival, participants read and signed an informed consent form. They completed a brief demographic questionnaire (Appendix A). I then gave a training presentation using PowerPoint. There were three sections in the training presentation. The first section was an introduction to each of the displays, in which I showed the participant the displays and described how to read each of them. During the second section of the training presentation, the participant individually walked through each of the displays again, with PowerPoint animations demonstrating how to answer the same question for each display. The purpose of the final section of the training presentation was to determine whether the participant understood how to answer different types of questions using each display. I asked the participant how to answer a different sample question for each display, correcting the participant when necessary. By the end of the training presentation, participants had seen each display three times.

I used a Latin square to determine the presentation order of the displays (Appendix B). The first 10 participants randomly drew a number that corresponded to one of the 10 orders listed in the Latin square in Appendix B (random assignment without replacement). The next 10 participants followed the same procedure. I entered the drawn number into the software to cue the correct presentation order of displays.

After the training presentation, I opened the experimental software and gave directions to the participant about what s/he would see, when s/he could take breaks if required, and how to input answers. There was a calculator (Texas Instruments TI-30X

IIS), pencil, and paper available to them if they needed. I stayed in the room to assist with any questions or software malfunctions.

Each participant then completed an experimental session lasting approximately 60 minutes with five experimental blocks. Each block contained a single display; the order of display presentation was counterbalanced between participants (see Appendix B). Participants answered each of the 12 questions exactly once within a block. Each question was randomly paired, without replacement, to one of the 12 snapshots of data across time. Thus, participants saw each of the 12 images developed for a display exactly once within a block. The software presented the questions in a random order.

Ten of the 12 questions contained variables that the software filled in randomly to determine the specific question to be asked. For example, consider Question 3 in Table 1: “What are the total number of bytes being received by (IPaddress)?” The software replaced the variable “(IP address)” with a randomly determined destination address (from the pool of destination addresses that received information in that particular data snapshot) to instantiate the question.

I instructed participants to answer each question as quickly and as accurately as possible. Participants initiated a trial by clicking on a “Begin” button on the screen. The display appeared in the left portion of the screen, the question appeared in the upper right, and an on-screen number pad appeared in the middle-right (see Figure 6). Participants clicked the buttons of the number pad on the screen to record their response in a text box. Participants could clear this text box by clicking on the clear button or complete their response by clicking on the “Record Answer” button. The screen cleared, participants received feedback (their response, the correct answer, and their response time), and then

Rectangle Area: Cumulative Transmissions Between IP's

Top IP: Source
Bottom IP: Destination

Bytes

ICMP
TCP
UDP

What are the total number of ICMP bytes being Received by 192.168.2.87 ?

Answer:

Record Answer

7 8 9
4 5 6
1 2 3
0
Clear

Top IP (Source)	Bottom IP (Destination)	Transmission Type	Bytes
192.168.2.87	192.168.1.2	ICMP	168,000
192.168.1.2	192.168.2.87	ICMP	168,000
192.168.2.73	192.168.1.2	UDP	168,000
192.168.2.159	192.168.1.2	ICMP	108,000
192.168.2.58	192.168.1.2	ICMP	108,000
192.168.2.30	192.168.1.2	ICMP	108,000
192.168.2.32	192.168.1.14	ICMP	60,000
192.168.1.2	192.168.2.73	UDP	60,000
192.168.1.2	192.168.2.58	ICMP	60,000
192.168.1.2	192.168.2.159	ICMP	60,000
192.168.1.2	192.168.2.30	ICMP	24,000
192.168.1.14	192.168.2.32	ICMP	24,000
192.168.2.81	192.168.1.6	TCP	24,000
192.168.1.6	192.168.2.81	TCP	24,000
192.168.2.81	192.168.1.14	TCP	24,000
192.168.1.14	192.168.2.81	TCP	24,000
192.168.2.87	192.168.1.6	TCP	12,000
192.168.1.6	192.168.2.87	TCP	12,000
192.168.2.87	192.168.1.2	ICMP	0

At the end of the experiment, I thanked the subjects compensated them \$25 for their participation. Some participants expressed curiosity about the study and offered reactions to the different displays, but this was not prompted.

III. RESULTS

LATENCY

Latency was measured (1/20th sec accuracy) from the time the question appeared on the screen to the time the participant completed their response. I conducted a two-way within subjects repeated measures ANOVA with five display and 12 question levels. The main effects of display, $F(4,76) = 13.86$, $p < .000001$, and question, $F(11,209) = 29.72$, $p < .000001$, were significant, as was the display by question interaction, $F(4,76) = 11.51$, $p < .000001$.

Question. I conducted a contrast to test for any overall differences between question type (i.e., numerical vs. binary); it was significant, $F(1,19) = 94.28$, $p < .000001$. Contrasts between pairs of numerical and binary questions were significant for questions 1 versus 7, $F(1,19) = 30.66$, $p < .00003$, 2 versus 8, $F(1,19) = 121.19$, $p < .000001$, 3 versus 9, $F(1,19) = 43.11$, $p < .000003$ and 4 versus 10, $F(1,19) = 35.21$, $p < .00002$. Responses to numerical questions were significantly slower than those for binary responses (see Figure 7). The contrasts for 5 versus 11 and 6 versus 12 were not significant.

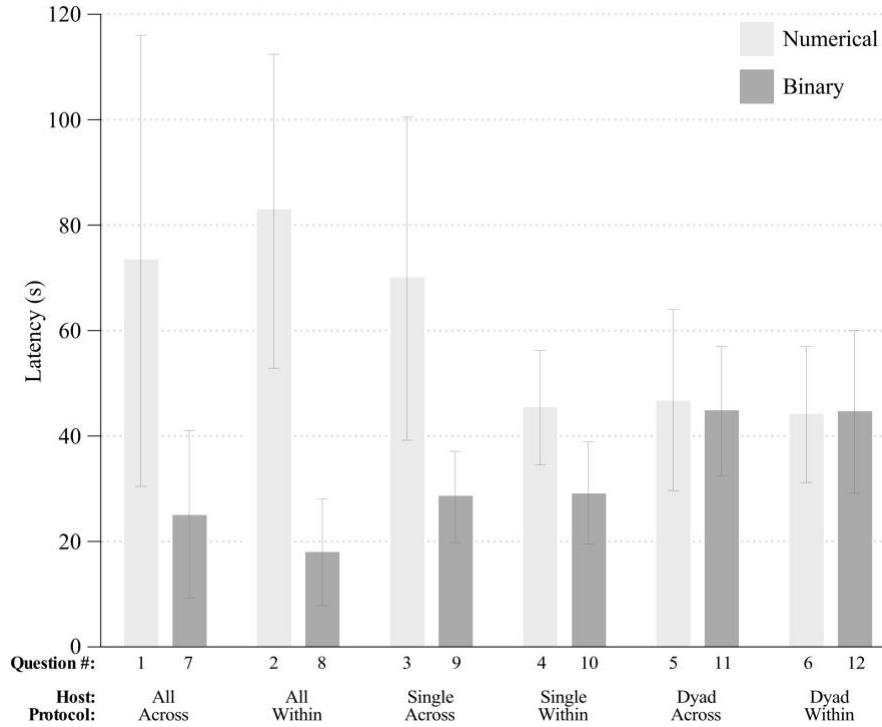


Figure 7. Mean latency scores (in seconds) for each pair of numerical and binary questions. Error bars show 95% confidence intervals.

Display by Question. I conducted contrasts testing the simple main effect of display at each question. The results were significant for Questions 1 [$F(4,76) = 6.84, p < .0001$], 2 [$F(4,76) = 8.39, p < .00002$], 3 [$F(4,76) = 20.49, p < .000001$], 4 [$F(4,76) = 25.62, p < .000001$], 5 [$F(4,76) = 15.56, p < .000001$], 6 [$F(4,76) = 14.22, p < .000001$], 7 [$F(4,76) = 2.82, p < .04$], and 8 [$F(4,76) = 3.06, p < .03$]; they were not significant for Questions 9-12. I conducted contrasts between displays when there was a simple main effect for display. Figure 8 represents the average latency performance for each display and each question (for Questions 1-6). All significant ($p < .05$) contrasts between displays are represented by the underscoring of two icons (i.e., there was a significant difference between two icons connected by an underscore).

The results for the binary Question 7 revealed that the Treemap display ($M = 21.48$ s) produced significantly lower latencies than the 3D display (32.15 s) and the Radial display (17.52 s) produced significantly lower latencies than the Alphanumeric (30.62 s) and the 3D display (32.15 s). The contrasts for the binary Question 8 revealed that the 2D display (10.31 s) produced significantly lower latencies than the Alphanumeric (24.95 s) and the 3D display (26.82 s).

Questions Requiring Exact Values (i.e., Numerical Responses):

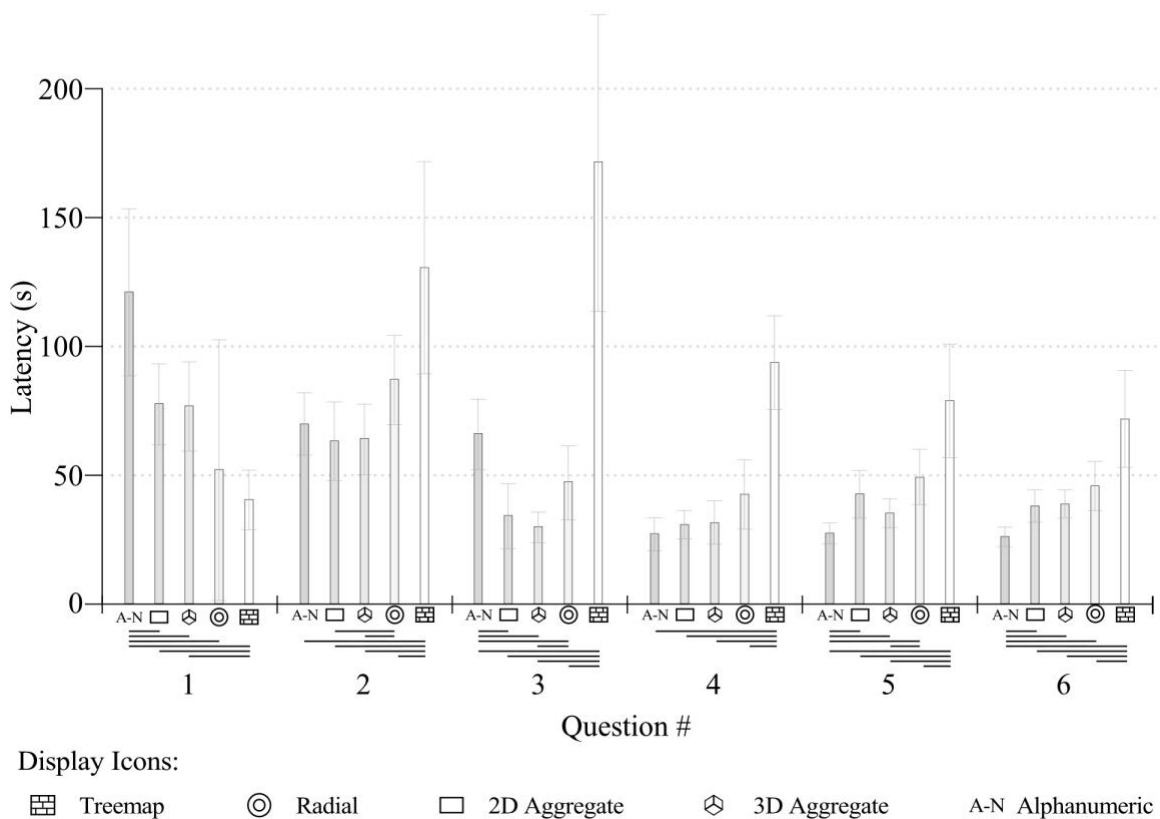


Figure 8. Mean latency for numerical questions (error bars show 95% confidence intervals.). Significant comparisons between two displays are represented by horizontal lines appearing under the x-axis.

ACCURACY

I calculated accuracy scores for Questions 1-6 by subtracting the participant's estimate from the actual number of bytes transmitted and taking the absolute value of the difference. The 5 (display) by 6 (question) repeated measures ANOVA revealed no significant effects for display, question, or the display by question interaction. I scored the accuracy for Questions 7-12 as correct or incorrect; the Cochran Q Tests revealed no significant differences between displays.

IV. DISCUSSION

I found a number of significant differences in latency performance across displays and across question types. However, there were no significant effects for accuracy. I begin the discussion section by providing different interpretations of this outcome for the two question types (numerical and binary). I then describe the pattern of results for each display, and finish with revisiting my original hypotheses.

ACCURACY

Binary Questions (7-12). There are two potential explanations for the lack of significant accuracy effects for the binary questions. The first is that there may have been a ceiling effect. These questions were not intended to be particularly difficult to answer. They dealt with global aspects of the information that was represented in the display. Participants did not have to conduct a detailed examination of the data to provide the correct answer. For example, to answer Question 8 (Are more bytes of one protocol type being sent than a different protocol type?), participants had to simply scan the display and compare the relative amount of one color to another. They did not have to calculate the amount of the first protocol type, calculate the amount of the second protocol type, then find the difference. The high overall accuracy of the responses (approximately 92% accuracy) in combination with the quick completion times (see Figure 7) provide some support for a ceiling effect.

The second explanation for the lack of significant results for the binary questions is a simple one: the statistical power was low. The binary responses required the use of

non-parametric statistics to determine significance; these tests are inherently less powerful, and therefore less capable of picking up differences in performance.

Numerical Questions (1-6). The lack of significant effects for the accuracy of questions that required numerical estimates (i.e., questions 1-6) was a more surprising outcome. This is particularly true for the alphanumeric display, because there is the capability to produce a completely accurate response on every trial (a capability that does not hold true for the graphical displays). One contributing factor is that the fundamental nature of the task introduced a great deal of variability. Participants were required to estimate transmission levels that could vary across a wide range of values: the smallest and largest correct responses in the dataset were 616 bytes and 338,528 bytes. I allowed (but did not require) participants to use pencil, paper, and/or a calculator with the express purpose of minimizing errors. Despite this, variability in performance was substantial: the average error was 27,435 bytes and the standard deviation was 182,278 bytes.

As these numbers suggest, the combination of working with these inherently large numbers, and perhaps doing so without use of pencil, paper, or calculator, produced the potential for very large errors. Examination of the numeric responses revealed a common strategy adopted by participants, one that does not appear to have involved these memory aids: to generate a “ball park” estimate using a small number of lead digits followed by an appropriate number of zeros (for example, “250,000” instead of “248,588”). Providing the correct number of zeroes on a consistent basis with this strategy would be difficult; adding one too many, or one too few zeros would be an easy mistake and would produce large errors. Thus, it is likely that this common strategy contributed to the large errors and variability of responses.

Similarly, an unintended digit in the response (e.g., an entry error) would have the same effect. For example, consider the least accurate score in the experiment: an error score of 3,009,448 bytes. The correct answer was 224,552 bytes and the participant's response was 3,234,000 bytes. It is possible that the leading digit "3" was a simple entry error. If this digit is removed (i.e., assuming that the intended response was 234,000), then the result would have been a very reasonable error score of 9,448 bytes. Devising an alternative method to measure accuracy with this type of data is a goal of future research.

LATENCY

The majority of the latency results for Question types (i.e., numerical versus binary response questions, see Figure 7) support the prediction that the binary questions produced quicker responses compared to the numerical questions. Performance for all questions addressing total transmissions across all hosts (i.e., Questions 1 and 2) and total transmissions for a single host (i.e., Questions 3 and 4) was significantly different. Performance for total transmissions between a pair of hosts (i.e., Questions 5 and 6) was not significantly different. As Figure 7 shows, the latency for numerical questions generally decreased from Question 1 to Question 6, but the latency for the binary questions generally increased from Question 7 to Question 12. For the numerical questions, this trend is likely explained by fewer mathematical calculations being required as the questions became more specific (i.e., asking about a dyad of hosts, usually referencing single transmissions versus asking about all data sent across all hosts). For the binary questions, the upward trend in latency implies that it took participants longer to make the finer discriminations required to answer the later questions. Instead of looking at perceptually large graphical elements to answer the questions, participants

were required to look at smaller elements and make finer distinctions to answer the questions.

There were many significant results between displays for the numerical latency responses of Questions 1-6 (see Figure 8). Figure 9 reorganizes and summarizes these findings in terms of the overall pattern of significant contrasts both between and within displays. The graph on the left (Figure 9a) summarizes the number of significant comparisons between displays across all questions. The numbers in each cell in the matrix summarize performance between displays. Consider the top cell in the matrix (2 | 2). The number to the left (2) indicates the number of comparisons that favor the display that appears on the y axis (3D); the number to the right (2) indicates the number of comparisons that favor the display on the x axis (Alphanumeric).

The graph on the right (Figure 9b) summarizes these numbers within each individual display. The bar labeled “+” indicates the number of significant comparisons favoring that display; the “-“ bar indicates the number of significant comparisons favoring all other displays. The displays are arranged along the x axis in terms of overall performance with the best display on the left and progressively poorer performance to the right. In the following sections, I discuss the results for each display and interpret them in terms of the EID principles of constraint matching between task demands, visual structure in displays, and visual attention/form perception that I outlined in the introduction.

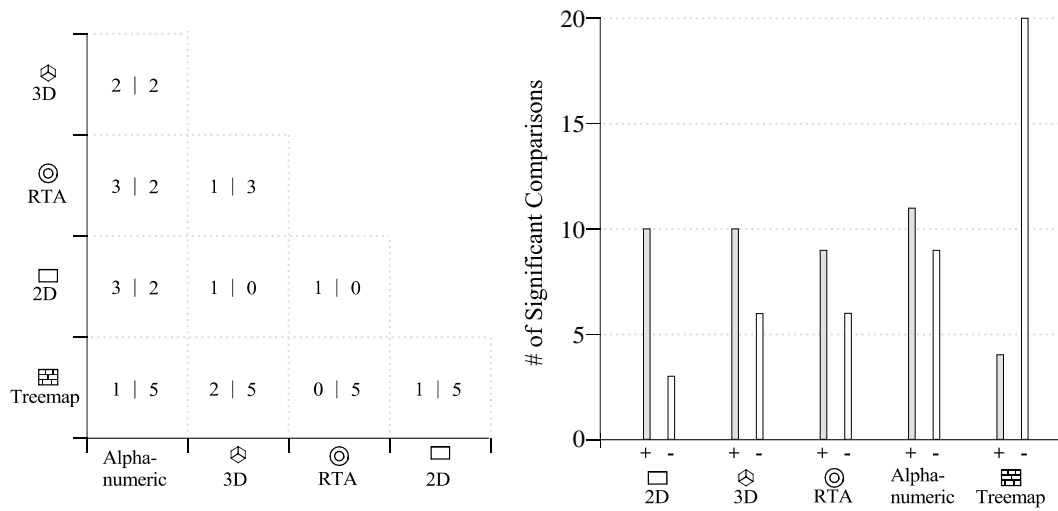


Figure 9a (left). Summary of the significant favorable (left number in each cell) and unfavorable (right number) comparisons between displays and across questions. Figure 9b (right). Summary of significant favorable (+ column) and unfavorable (- column) comparisons for each display.

Alphanumeric Display. I begin with a consideration of results for the alphanumeric display. Question 1 required consideration of global properties of network traffic; all alphanumeric entries in the display needed to be added together. This task constraint produced significantly longer response latencies with the alphanumeric display than all of the other graphical displays (see Figure 8). In contrast, Question 6 required consideration of very specific aspects of network traffic: transmissions between two specific hosts that involved a single protocol type. The alphanumeric display contained only one entry that specified the correct answer; the participant had only to find it and enter the response. As a result, the alphanumeric display produced response latencies that were significantly faster than all other graphical displays on Question 6. On average, Questions 2 through 5 required progressively fewer transmission entries to be considered;

the time required to complete responses varied in a reasonably systematic fashion between the two extremes outlined above.

These results are consistent with previous research investigating tradeoffs between alphanumeric and graphical displays (e.g., Bennett & Walters, 2001; Hansen, 1995). Alphanumeric displays provide detailed, precise information that can be extremely beneficial when exact values are required. However, the lack of analogical graphical properties severely limits their utility as the primary representation in an interface (e.g., Bennett & Flach, 2011). Domain semantics (e.g., relationships, properties, goals, constraints) are not visible directly; unlimited perceptual resources are not leveraged. Meaning must be derived mentally using limited capacity cognitive resources.

The four graphical displays did not have exact values and participants needed to employ a visual estimation process to produce a response. The principles of specificity and attunement mentioned in the introduction are particularly relevant to the interpretation of the results from the graphical displays. In the interpretations that follow, I refer to performance differences between the four graphical displays (i.e., not the alphanumeric display), unless otherwise noted.

Attunement. The EID principle of attunement refers, in part, to perceptual perspicuity: how visually salient, in a psychophysical sense, are the emergent features produced by a display? One attempt to provide an empirical answer to this question is provided by Cleveland and his colleagues (e.g., Cleveland, 1985) who evaluated a number of “elementary graphical perception tasks.” This includes the primary emergent features produced by each of the four graphical displays: position along a common scale (2D and 3D displays), angle (Radial), and area (Treemap).

The ranking of relative effectiveness obtained by Cleveland et al. for these emergent features is an exact match to the overall pattern of significant differences in this experiment (i.e., the left to right ordering of the displays along the axis of Figure 9b). Thus, one interpretation is that the fundamental representational choices contributed to the ease with which observers could pick up the information encoded into the various graphical displays.

Specificity. The EID principle of specificity refers to the extent to which the constraints of the work domain have been faithfully represented in the geometrical constraints of the display (i.e., does the visual evidence provided by the display map directly onto the significant possibilities or affordances of the work domain?). The differences between how the displays structured information affected how easily participants could decode the information about the underlying domain. In the next few sections, I discuss each of the displays in terms of how their specificity and attunement affected participant performance (in terms of latency).

Treemap. The Treemap display produced the poorest performance of all displays that were evaluated (see Figure 9). It produced significantly slower response latencies for **all** comparisons with **all** other displays in Questions 2-6. As mentioned previously, part of this poor performance is likely to be due to attunement and the fundamental representational choice (area). More specifically, the visual estimation process for the Treemap display was more complicated than the other graphical displays. After locating a relevant rectangle, two emergent features (height and width) needed to be estimated (in conjunction with grid lines, scale markers, and labels on the y axis) to generate two exact numerical values. An estimate of height (corresponding to the number of bytes

transmitted by all hosts in the row) needed to be multiplied by an estimate of width (corresponding to the proportion of the row occupied by that particular host) to obtain the final value. All other graphical displays required visual estimation of only one emergent feature (i.e., angle or linear extent) to generate a value.

Aspects of the Treemap display that are related to specificity are also likely to have contributed to the poor performance for Questions 3-6. The Treemap was originally designed as a way to represent hierarchical data structures (Johnson & Shneiderman, 1991; Shneiderman, 1992; Johnson, 1993). I used the squarified algorithm (Bruls, Juizing, & van Wijk, 2000) with the goal of producing rectangles that are more square-like (i.e., not long and thin), and therefore easier to work with (Kong, Heer, & Agrawala, 2010). The algorithm divides the available space into rectangles with the goal of achieving aspect ratios as close to 1 as possible. Because there were no parent-child hierarchical relationships in the data, the algorithm divided the entire display space and created rectangles accordingly. All five displays were organized to order transmissions in terms of size; in the Treemap, this meant that the largest squares appeared in the upper left and the squares got progressively smaller to the lower right. Thus, the current version of the Treemap display provided a poor mapping in terms of specificity: the hierarchical structure that was a key component of the original Treemap was missing. This hierarchical structure was a fundamental component required to answer Questions 3-6. Instead of having all data transmissions relative to a particular IP address in one physical location, the participant needed to search the entire display by reading the labels for each rectangle.

In contrast to Questions 2-6, performance on Question 1 was significantly better

for the Treemap display relative to both the 2D and the 3D displays. Performance with the Treemap display was also significantly better than performance with the 3D display for Question 7. The quality of specificity mapping is likely to have contributed to this result. Recall that both of these questions assess global levels of network traffic (i.e., the total amount of data for all transmissions). Unlike the 2D and 3D displays, the Treemap display (and the RTA display) had graphical representations and a scale (located on the y axis) that directly specified these global levels. Forming an estimate was therefore a fairly simple process: reading the label of the highest value gridline and adding an estimate for the graphical portion above the last scale value. In contrast, the representations and scales for the 2D and 3D displays were not constructed to reflect this property of the work domain. As a result, forming a response was a far more complicated process: participants needed to estimate the values of multiple graphical elements (up to 10) and combine them to arrive at a global estimate.

Radial Display. The Radial display produced intermediate levels of performance relative to the other graphical displays. Like the Treemap, the Radial display contained a scale that directly specified global levels of network traffic; latency for Question 1 was faster than the 2D and 3D displays but slower than the Treemap display (although not significantly different in any instance). For Questions 2-6 the average latency with this display was always faster than the Treemap display and always slower than both the 2D and 3D displays. These performance differences were significantly better than the Treemap display for all 5 questions, significantly worse than the 3D display for Questions 2, 3, and 5 and significantly worse than the 2D display for Question 2.

As mentioned previously, this intermediate pattern of performance is consistent with attunement and the general discriminability of the primary emergent feature (angle). The effects of this representational choice may have been exacerbated when transmission levels were high, but not specified by the overall scale. The psychophysical literature provides some evidence that, in general, sensitivity decreases as angle sizes become larger (e.g., Maclean & Stacey, 1971). Thus, the large angle sizes associated with the transmission levels required in the responses to Questions 2 and 3 may have contributed to the significantly poorer performance with the Radial display.

Specificity may have contributed to the intermediate pattern of results as well. Recall the earlier discussion of the hierarchical structure that was critical in answering Questions 3-6: with the Treemap display, this structure was completely removed. This was also true for the Radial display when the question concerned received transmissions. All of the data transmissions received by a particular IP address were spread out across the outer ring of the display, because transmissions were organized by source IP in the inner ring. As a result, the participant was required to search the outer ring for all relevant transmissions by reading the labels for each wedge. In contrast, this hierarchical structure was always maintained for data transmissions sent by a particular IP address: all relevant wedges were located in a contiguous physical location in the inner ring. Note that the hierarchical structure was always present in both the 2D and the 3D displays regardless of whether the concern was sending or receiving (see ensuing discussion). Thus, the requirement to search for, estimate, and combine numerical values from multiple representations was always required for the Treemap display, sometimes required for the Radial display, and never required for the 2D and 3D displays.

2D and 3D Displays. The 2D and 3D displays produced the best overall performance of the four graphical displays. Overall, there were 18 significant contrasts between these two displays and the other two displays (i.e., Radial and Treemap). Fourteen contrasts favored the 2D and 3D displays. As mentioned previously, attunement is likely to have played a role: Cleveland et al. (e.g., Cleveland, 1985) found that the emergent feature produced by these two graphical displays (vertical extent, or position along a common baseline) was the most visually salient of those that were tested (and those that were used in the present experiment).

Specificity is also likely to have played a role in the positive results for the 2D and 3D displays. The majority (10 out of 14) of the significant contrasts favoring these displays were obtained for Questions 3 through 6. The 2D and 3D displays provided visual structure that corresponded directly to the information that was required to answer these questions. Questions 3 and 4 required that network traffic be considered in terms of aggregated transmissions that occurred for a single host (see Table 1). Both the 2D and the 3D displays provided a single graphical form that specified these summarized values directly (see Figures 3 and 4): either a contribution bar graph (Question 3) or a segment within a contribution bar graph (Question 4). Questions 5 and 6 required the participant to consider network traffic in terms of aggregated transmissions that occurred between a dyad of hosts. Both the 2D and the 3D displays provided a single graphical form that specified these values directly as either a contribution bar graph (or column, Question 5) or as a segment within a contribution bar graph (or a contribution column, Question 6).

In contrast, the visual information required to answer these questions with the Treemap and Radial displays was often scattered around numerous spatial locations in the

display. Thus, the increased time and effort required to search the display and integrate the information produced increases in the latency of responses.

As mentioned previously, specificity is likely to have played a major role in interpreting the four significant contrasts that did not favor the 2D and 3D displays. All four of these findings were obtained for the two questions (1 and 7) that required assessments of global levels of network traffic. Both the Treemap and the RTA display had visual forms and numerical scales that corresponded directly to these global levels, whereas the 2D and 3D displays did not (see the discussion in the Treemap section).

The 2D and 3D displays were very similar. They used the same fundamental emergent feature (vertical extent from a common baseline) to specify information and the same conceptual structure (aggregated transmissions for a specific IP and aggregated transmissions between two IPs) to specify various aspects of network traffic. The only real differences between them arise from the application of a 3-dimensional perspective in the 3D display. In terms of direct statistical comparisons between them, the average latency for the numerical questions was virtually identical with no significant differences; for the binary questions a single contrast (Question 8) was found to favor the 2D display.

HYPOTHESES REVISITED

I predicted that there would be performance differences between different displays and the different question types. Because I did not find any significant results with the accuracy data, I discuss the hypotheses in terms of latency.

I predicted that performance on the graphical displays would follow the pattern that Cleveland et al. identified: displays using vertical extent as a graphical form would be better than displays using angle as a graphical form; displays using area as a graphical

form would yield the worst performance of the three. Figure 10 shows the number of significant comparisons in for each display. The left bar represents the number of times a particular display outperformed the other displays, and the right bar represents the number of times a particular display was outperformed by the other displays. Looking at the dark gray segment of each display (representing the comparisons to the other graphical displays), Cleveland's pattern of graphical perception appears. The two vertical extent displays (2D and 3D) had 6 (2D) and 8 (3D) positive comparisons and 1 negative comparison each. The area display (Treemap) had only 2 positive comparisons and 15 negative comparisons. The area display (Radial) fell in the middle of the other two, with 5 positive and 4 negative comparisons. This pattern of results supports my first hypothesis.

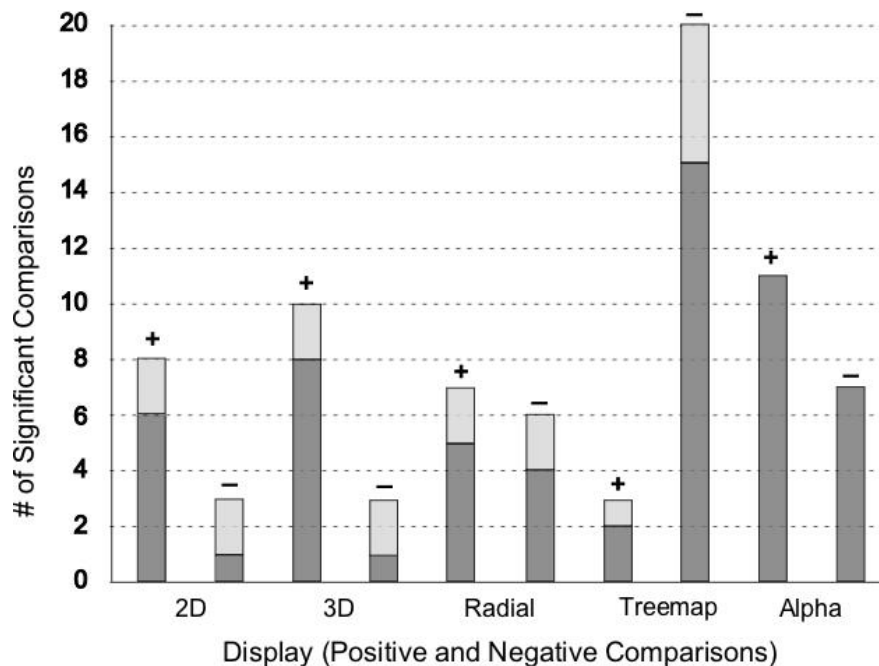






































Figure 10. Number of positive (left column) and negative (right column) comparisons for each display. Dark gray rectangles are comparisons to other graphical displays, light gray rectangles are comparisons to Alphanumeric display.

Next, I predicted that performance with the graphical displays would vary based on how well the graphical forms supported the task constraints associated with answering each of the questions (i.e., specificity). I expected the displays that represented data in a global manner (i.e., the Treemap and Radial display) to yield better performance on the questions that asked participants to consider the global properties of the data (i.e., Questions 1, 2, 7, 8). I predicted that the displays that showed transmissions between dyads and aggregated across hosts would do better on the single host (Questions 3, 4, 9, 10) and host dyad (Questions 5, 6, 11, 12) questions. Table 2 below shows the predicted and actual outcomes for each of the questions.

Table 2.

Performance of graphical displays by question: Predicted versus actual outcomes.





Exact Values (i.e., Numerical Responses):			
<u>Host:</u>	<u>Protocol:</u>	<u>Predicted:</u>	<u>Actual:</u>
All hosts	Across	1.  	 
	Within	2.  	 
Single host	Across	3.  	
	Within	4.  	 
Host dyad	Across	5.  	
	Within	6.  	 
Visual Judgments (i.e., Binary Responses):			
<u>Host:</u>	<u>Protocol:</u>	<u>Predicted:</u>	<u>Actual:</u>
All hosts	Across	7.  	
	Within	8.  	
Single host	Across	9.  	n.s.
	Within	10.  	n.s.
Host dyad	Across	11.  	n.s.
	Within	12.  	n.s.

The 2D and 3D displays did better on the within protocol questions in the “All hosts” category (Questions 2 and 8). When asked about specific types of protocols, participants had to look at individual graphical elements instead of the global picture. The scale on the Radial and Treemap displays did not support participants when answering Questions 2 and 8 (poorer specificity). The Radial display did better than expected on Questions 4 and 6; there were significant differences between it and the Treemap, but not when compared to the 2D and 3D displays. The results partially supported my second hypothesis.

I hypothesized that the Alphanumeric display would be faster for host dyad questions and slower for the all hosts questions when compared to the graphical displays. Because the host dyad questions usually referred to a single line on the display, the participant only needed to find the appropriate line and enter the response, yielding fast response times. Table 3 shows the results for each question.

Table 3.

Comparison of Alphanumeric display to graphical displays: Predicted versus actual outcomes.

Exact Values (i.e., Numerical Responses):			
<u>Host:</u>	<u>Protocol:</u>	<u>Predicted:</u>	<u>Actual:</u>
All hosts	Across	1. Graph > Alpha	
	Within	2. Graph < Alpha	
Single host	Across	3. Alpha = Graph	
	Within	4. Alpha = Graph	Alpha
Host dyad	Across	5. Alpha < Graph	Alpha
	Within	6. Alpha < Graph	Alpha
Visual Judgments (i.e., Binary Responses):			
<u>Host:</u>	<u>Protocol:</u>	<u>Predicted:</u>	<u>Actual:</u>
All hosts	Across	7. Alpha > Graph	


	Within	8. Alpha > Graph	
Single host	Across	9. Alpha = Graph	n.s.
	Within	10. Alpha = Graph	n.s.
Host dyad	Across	11. Alpha < Graph	n.s.
	Within	12. Alpha < Graph	n.s.

Table 3 shows that the Alphanumeric display was slower than the graphical displays for the all hosts questions (Questions 1, 2, 7, 8) and faster than the graphical displays for the host dyad questions (Questions 5, 6). These results supported my third hypothesis.

Finally, I predicted that participants would answer the binary questions faster than the numerical questions. Figure 7 (in the Results section) shows the average latency for each question, organized by related pair (i.e., Questions 1 and 7, Questions 2 and 8, etc.). There were significant differences between the first four pairs (all hosts and single host questions) in which participants answered faster on the binary versions of the questions. There were no significant differences between the last two pairs of questions (host dyad questions). As discussed earlier, the latency for the numerical questions quickened from Question 1 to Question 6, and the latency for the binary questions slowed from Question 7 to Question 12. These results partially support my fourth hypothesis.

LIMITATIONS

Participants performed poorly using the Treemap display. A few participants even commented after the experiment was over that the Treemap was the worst display. However, treemaps are used in many different contexts and seem to be an accepted form of visualizing information. Its poor performance in this study could be due to a couple factors.

First, the Treemap was originally designed as a way of showing hierarchical data (Shneiderman, 1992). For example, it has been applied to visualizing the distribution of file types on a computer's hard drive (see Figure 10). For this experiment, I did not incorporate a hierarchical organization, thus separating it from one of its principal elements. I may have unintentionally created a strawman, but the data did not readily lend itself to a hierarchical organization. In a future iteration of this study, I will test a redesigned Treemap display that retains a hierarchical organization.

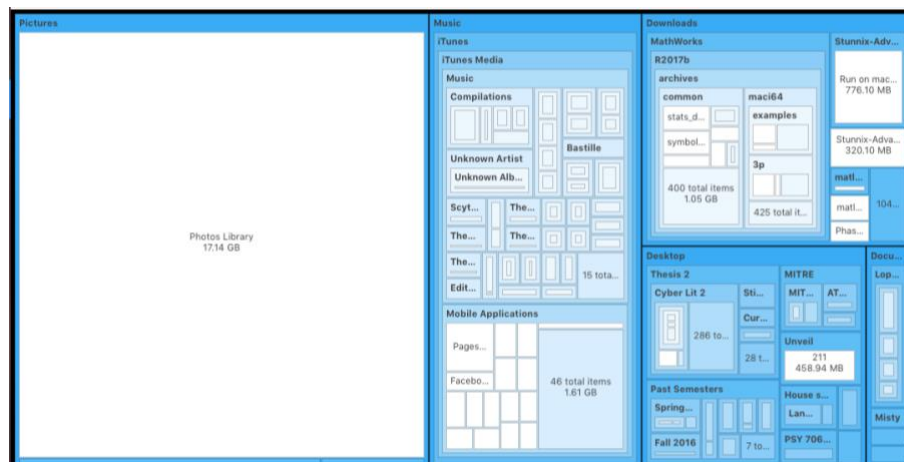


Figure 11. Treemap showing the hierarchical organization of a computer's hard drive

Second, performance using the Treemap could have suffered because none of the questions that I asked in this experiment played to its strengths as a display. If the Treemap is best used as a way to visualize hierarchically organized data, none of the questions in this experiment would have matched it in terms of specificity. Question 1 was the only question in which participants answered faster using the Treemap; however, that was most likely due to the inclusion of a scale, rather than an inherent strength in the Treemap display for that particular question. In future studies, I should attempt to find questions that are more specific to the Treemap's original design and organization.

A third factor that may have hindered performance on the Treemap is that participants had to perform an extra step when estimating values. With the other graphical displays, they could estimate values based on a single emergent feature – vertical extent for the 2D and 3D displays, and angle for the Radial display. With the Treemap, participants had to estimate two linear extents (height and width), then multiply them together to arrive at the estimated answer.

The final factor that may have hindered performance on the Treemap is also relevant to the other displays. Each of these displays were originally designed to be interactive. However, I had to make them static to fit the constraints of this experiment. When a Treemap is used in the real world, users are usually able to hover over a box to get a digital read-out of its value, or there are labels available. The 3D display was originally designed so the user could move the cube around to better see occluded columns or visually compare values. Even Wireshark, the program on which the Alphanumeric display was based, is interactive, allowing users to sort data on a range of criteria. While this lack of interactivity may have hindered performance, I argue that the logic behind making the displays static was valid. It is likely that users of these displays would look at the graphical elements first, then drill down to find exact values when needed. The goal of this experiment was to compare displays with different graphical perception elements using common measures, not to evaluate how participants interacted with dynamic displays.

Another limitation of this study was the poor quality of the accuracy data. For Questions 1-6, the variance for both latency and accuracy was extremely high. I suspect that the extreme variances seen in the accuracy data for each question occluded any

possible trends that may have emerged. As discussed earlier, part of this variance could be due to entry errors (e.g., adding a random digit before the intended answer) or the variance could be due to the magnitude of errors growing with larger numbers. For Questions 7-12, there was a ceiling effect for both accuracy and latency. The binary judgement questions were too easy for participants. Participants answered significantly faster on the binary questions for each display than the quantitative questions (Questions 1-6). To address these limitations, future iterations of this study should implement a different input mechanism that may reduce the number of entry errors. One possibility is a slider that a participant can use to calibrate to the correct answer, instead of entering it in using a number pad. In this study, I attempted to identify obvious input errors and eliminate them from the data set, but they did not affect the variance. In future, I will do logarithmic transforms of the data to see if that addresses the high variance in a statistical manner.

V. CONCLUSION

Evaluating a display for a complex system like CND should not be a single activity. Rasmussen et al. (1994) and Bennet and Flach (2011) described the importance of hierarchically nested levels of evaluation for a system (Figure 12). In terms of display design, the initial evaluations should focus on the coherence of the display, control, or navigation with human capabilities and limitations (Boundary Level 1 in Figure 12). Intermediate evaluations should focus on how well the display supports higher level cognitive functions, like decision making, pattern recognition, and problem solving (Boundary Levels 3 and 4). The highest level of evaluation should focus on testing the interface in increasingly realistic scenarios to assess how well it matches to the constraints of the work domain (Boundary Level 5).

Based on Staheli et al.'s findings about the current published literature about CND display evaluations (2014), most evaluations to date have focused on Boundary Level 3 or above – realistic work scenarios in simulated task environments or field studies. However, there is a lack of studies focusing on Boundaries 1 and 2. Staheli et al. called for more studies at these levels to fill the existing gap in the literature. They suggested that non-experts could be used for these types of experiments, capitalizing on larger subject pools. The current study falls within this category. These types of studies will help create a solid foundation of how to design interfaces for the complex domain of CND.

Although I used non-expert participants, I found support for my hypotheses that will likely generalize to CND analysts because I broke down a complex task into simple perceptual and cognitive subtasks. The results validated the use of the triadic approach found in EID; namely that the mappings between the human user and the interface (attunement) and the work domain and the interface (specificity) affect performance in specific ways. When the display presented information in a manner that was consistent with the task demands (specificity), the participants answered the questions quicker. Additionally, when human constraints related to graphical perception were accounted for in the design of the display, participants also answered more quickly.

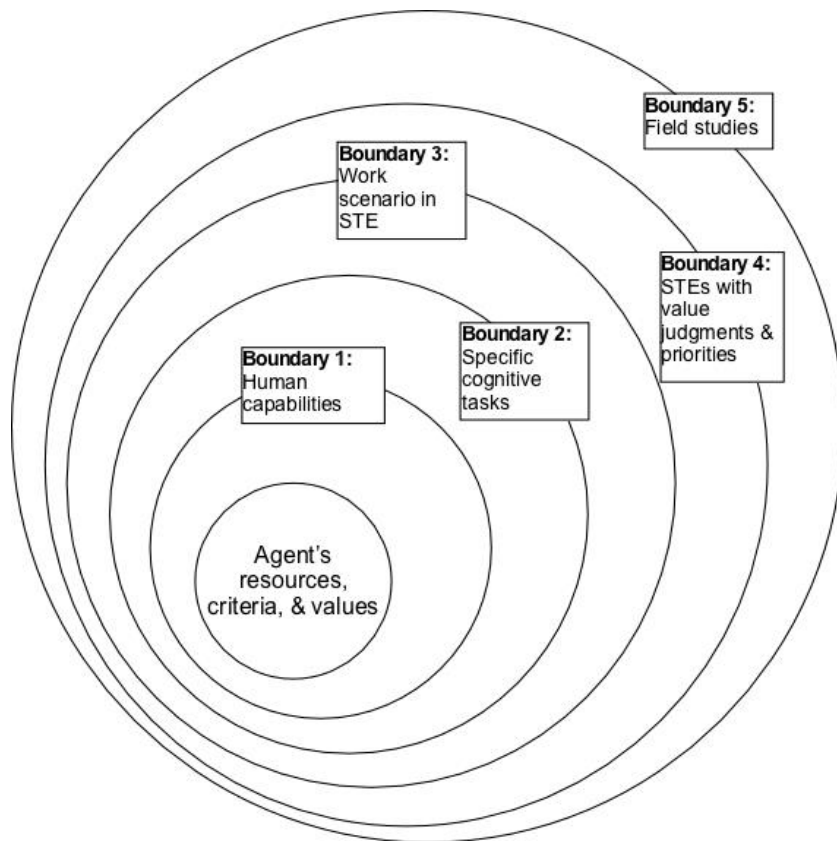


Figure 12. Hierarchically nested levels of evaluation (Rasmussen, Pejtersen, & Goodstein, 1994; Bennett & Flach, 2011)

The findings from this experiment can be used to design candidate visualizations that are better tied to CND tasks. For example, if there is a network monitoring task that requires comparing values across dyads of a hosts, a display using vertical extent like the 2D and 3D displays in this experiment might be better suited than a display using area or angle to represent the important information. Performance on the displays varied based on how well the display represented the meaning of the domain and how that cohered with the goals of the participant in answering specific questions. With a complex domain like CND, it is vital that the interface represents the domain so analysts can interpret and find the underlying meaning that is important to their current goals and tasks.

The study described in this paper was the first of a series of studies comparing the four graphical displays (3D, 2D, Treemap, and Radial). The next study will focus on the effects of scales on visual comparisons. Participants in this study will make visual comparisons without the use of a scale. The inclusion of a scale in the current study most likely skewed the responses to Question 1. It will be interesting to see if the pattern of results changes when removing the numerical scales. I will also address the problem with the design of the Treemap in another iteration of the study by using a version of the display that retains its original hierarchical organization.

Future research should continue to break CND tasks down into perceptual and cognitive components so graphical display elements can be tested in their most basic forms within the context of the CND domain. There are challenges unique to the CND domain that are not being supported by existing visualizations (Best et al., 2014). It is important to analyze these challenges at different levels and design solutions that support each level – the users’ perceptual skills, cognitive and macrocognitive requirements,

along with the constraints of the work domain. Only then will we be better positioned to create cyber data visualizations that will be likely to support CND analysts in the complex, ever-changing environment in which they operate.

VI. APPENDICES

APPENDIX A: BACKGROUND QUESTIONNAIRE

1. How old are you?

2. Are you male or female (circle one)?

Male Female Other Prefer not to answer

3. How many years of education beyond high school do you have (circle one; if you are a freshman, please circle 1, sophomore, circle 2, etc.)?

1 2 3 4 5+

4. What is your major or field of study?

APPENDIX B: COUNTERBALANCED PRESENTATION SEQUENCE

	Presentation Order				
Group 1	1	2	5	3	4
Group 2	2	3	1	4	5
Group 3	3	4	2	5	1
Group 4	4	5	3	1	2
Group 5	5	1	4	2	3
Group 6	4	3	5	2	1
Group 7	5	4	1	3	2
Group 8	1	5	2	4	3
Group 9	2	1	3	5	4
Group 10	3	2	4	1	5

Latin Square counterbalancing technique, taken from Shaughnessy, Zechmeister, & Zechmeister, 2012.

VII. REFERENCES

1. Adobe Director (Version 11.5) [Computer software]. Adobe Systems, Incorporated.
2. Aschenbrenner, B. (2008). Identification of command and control information requirements for the cyberspace domain. (AFIT/GIR/ENG/08-01). Wright Patterson Air Force Base, OH: Air Force Institute of Technology.
3. Bennett, K. B. (2014). VEILS: An ecological interface for computer network defense. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1233-1237. Los Angeles: SAGE Publications.
4. Bennett, K. B. & Flach, J. M. (1992). Graphical displays: Implications for divided attention, focused attention, and problem solving. *Human Factors*, 34(5), 513-533.
5. Bennett, K. B. & Flach, J. M. (2011). *Display interface design: Subtle science, exact art*. CRC Press.
6. Bennett, K. B. & Walters, B. (2001). Configural display design techniques considered at multiple levels of evaluation. *Human Factors*, 43(3), 415-434.
7. Best, D. M., Endert, A., & Kidwell, D. (2014). 7 key challenges for visualization in cyber network defense. *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*, 33-40. ACM.
8. Boles, D. B. & Wickens, C. D. (1987). Display formatting in information integration and nonintegration tasks. *Human Factors*, 29(4), 395-406

9. Bruls, M., Huizing, K., & Van Wijk, J. J. (2000). Squarified treemaps. In *Data Visualization 2000* (pp. 33-42). Springer: Vienna.
10. Canvas Draw (Version 3) [Computer software]. ACD Systems International.
11. Cleveland, W. S. (1985). *The elements of graphing data*. Belmont, CA: Wadsworth.
12. Cleveland, W. S. & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716), 828-833.
13. D'Amico, A., & Whitley, K. (2008). The real work of computer network defense analysts. In *VizSEC 2007* (pp. 19-37). Springer: Berlin Heidelberg.
14. D'Amico, A., Whitley, K., Tesone, D., O'Brien, B., & Roth, E. (2005). Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3), 229-233. SAGE Publications. D'Amico, A. D., Goodall, J. R., Tesone, D. R., & Kopylec, J. K. (2007). Visual discovery in computer network defense. *IEEE Computer Graphics and Applications*, 27(5), 20-27.
15. D'Amico, A., & Whitley, K. (2008). The real work of computer network defense analysts. In *VizSEC 2007* (pp. 19-37). Springer Berlin Heidelberg.
16. Goodall, J. R. (2008). Introduction to visualization for computer security. In *VizSEC 2007* (pp. 1-17). Springer Berlin Heidelberg.
17. Gutzwiller, R. S., Hunt, S. M., & Lange, D. S. (2016). A task analysis toward characterizing cyber-cognitive situation awareness (CCSA) in cyber defense analysts. *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2016 IEEE International Multi-Disciplinary Conference, pp. 14-20.

IEEE.

18. Hansen, J. P. (1995). An experimental investigation of configural, digital, and temporal information on process displays. *Human Factors*, 37: 539-552.
19. Hanson, R. H., Payne, D. G., Shively, R. J., & Kantowitz, B. H. (1981). Process control simulation research in monitoring analog and digital displays. *Proceedings of the Human Factors Society 25th Annual Meeting*, pp. 154-158.
20. IEEE. (2011). *IEEE VAST Challenge 2011, Mini-Challenge 2: Cybersecurity - situational awareness in computer networks*.
21. International Business Machines Corporation (2013). IBM security services cyber security intelligence index. Somers, NY: IBM Global Technology Services.
22. International Business Machines Corporation (2018). *IBM X-force threat intelligence index 2018*. Armonk, NY: IBM Security.
23. Johnson, B. S. (1993). Treemaps: Visualizing hierarchical and categorical data (Unpublished doctoral dissertation). The University of Maryland, College Park, MD.
24. Johnson, B. & Shneiderman, B. (1991). Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the 2nd conference on Visualization '91*, 284-291. IEEE Computer Society Press.
25. Keim, D. A., Mansmann, F., Schneidewind, J., & Schreck, T. (2006). Monitoring network traffic with Radial traffic analyzer. In *2006 IEEE Symposium on Visual Analytics Science and Technology* (pp. 123-128). IEEE.
26. Kong, N., Heer, J., & Agrawala, M. (2010). Perceptual guidelines for creating rectangular treemaps. *IEEE transactions on visualization and computer*

- graphics*, 16(6), 990-998.
27. Lovie, P. (1986). Identifying outliers. In A.D. Lovie (Ed.) *New developments in statistics for psychology and the social sciences* (pp. 44-69). London: British Psychological Society and Methuen.
 28. Maclean, I. E., & Stacey, B. G. (1971). Judgment of angle size: An experimental appraisal. *Perception & Psychophysics*, 9(6), 499-504.
 29. Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. North-Holland.
 30. Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive Systems Engineering*. Wiley.
 31. Rasmussen, J. & Vicente, K. J. (1989). Coping with human errors through system design: Implications for ecological interface design. *International Journal of Man-Machine Studies*, 31(5), 517-534.
 32. Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2012). *Research methods in psychology* (7th ed.). Boston, MA: McGraw Hill.
 33. Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1), 92-99.
 34. Staheli, D., Yu, T., Crouser, R. J., Damodaran, S., Nam, K., O'Gwynn, D., ..., & Harrison, L. (2014). Visualization evaluation for cyber security: Trends and future directions. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*, 49-56. ACM.
 35. Wireshark (Version 2.2.5) [Computer software]. Retrieved from <https://www.wireshark.org>.