Evaluating Visual Analytics Systems for Investigative Analysis: Deriving Design Principles from a Case Study

Youn-ah Kang* Carsten Görg† John Stasko‡

School of Interactive Computing & GVU Center, Georgia Institute of Technology

ABSTRACT

Despite the growing number of systems providing visual analytic support for investigative analysis, few empirical studies of the potential benefits of such systems have been conducted, particularly controlled, comparative evaluations. Determining how such systems foster insight and sensemaking is important for their continued growth and study, however. Furthermore, studies that identify how people use such systems and why they benefit (or not) can help inform the design of new systems in this area. We conducted an evaluation of the visual analytics system Jigsaw employed in a small investigative sensemaking exercise, and we compared its use to three other more traditional methods of analysis. Sixteen participants performed a simulated intelligence analysis task under one of the four conditions. Experimental results suggest that Jigsaw assisted participants to analyze the data and identify an embedded threat. We describe different analysis strategies used by study participants and how computational support (or the lack thereof) influenced the strategies. We then illustrate several characteristics of the sensemaking process identified in the study and provide design implications for investigative analysis tools based thereon. We conclude with recommendations for metrics and techniques for evaluating other visual analytics investigative analysis tools.

1 INTRODUCTION

Recent years have seen a rise in the number of visual analytics systems built to assist investigative analysis. Many stimuli are behind the development of these systems including the availability of example data sets via contests and challenges [9], the increasing importance of this type of work to government and intelligence activities [16], and the emergence of the visual analytics area in general.

Although many new investigative analysis systems are being built, we still do not well understand how to evaluate and assess them. Evaluating interactive visualization systems is challenging in general [8], but investigative analysis scenarios add even more difficulty. Going beyond the typical goals of information visualization such as identifying correlations, outliers, etc., investigative analysts perform sensemaking activities, develop hypotheses about the data, and seek to understand it more thoroughly. One often thinks of analysts "connecting the dots" or "putting the pieces together." Ultimately, analysts seek to develop insight about the data, a challenging activity to identify and measure [11].

One area in particular lacking much research is the controlled, comparative evaluation of investigative analysis systems. A number of systems have been studied in trial usage scenarios by trained analysts [1, 5, 17], but these studies did not compare performance against other systems or more traditional, "low-tech" approaches.

In this study, we examined use of the Jigsaw system [15] in an analysis scenario as compared to three other investigative methods including paper-and-pencil and simple desktop electronic document storage and search. While we were curious if Jigsaw would prove beneficial, the purpose of our study was not to evaluate Jigsaw per se. Instead, our primary goal was to better understand how visualization can assist investigative analysis, if it truly can. We wanted to see how people would approach data analysis using a visual analytics system. What characteristics of the system, if any, lead to the main benefits? We believe that a comparative scenario where one can examine people working on the same problem under different conditions, although limited in certain ways, does provide a valuable context to address these questions.

A second goal of this research was to better understand evaluation methodologies for investigative analysis systems in general. What should evaluators count, measure, and observe in order to determine the utility of systems? Identifying metrics for visual analytics system evaluation is challenging [12] and is important to organizations making decisions about which systems, if any, to use in practice.

This study is one of the first comparative experiments conducted in this area. We evaluated four settings for analysis. One of these used Jigsaw. 16 study participants performed an investigation in one of the settings. Each participant was given the same data collection containing 50 plain text documents each about a paragraph long. The documents simulated intelligence case reports and participants needed to identify an embedded terrorist plot within the allotted 90 minutes. In the sections that follow, we provide more details about the study design and resultant findings.

2 RELATED WORK

Few experiments have investigated the utility of visual analytic tools for investigative analysis. A study by Bier et al. [1] assessed the suitability of their Entity Workspace System in the context of design guidelines for collaborative intelligence analysis. The researchers modified their system based on five design guidelines and evaluated the system in both a laboratory study with intelligence analysts and a field study with an analysis team. Relying on analysts' subjective feedback in conjunction with quantitative logging data, they confirmed the positive effects of the tool on collaboration and the usefulness of the design guidelines for collaborative analysis.

Perer and Shneiderman [6] recognized the limitations of traditional controlled experiments in examining the process of exploratory data analysis and developed an evaluation methodology for studying the effectiveness of their system, SocialAction. Consisting of a long-term case study [14] and in-depth interviews, the evaluation confirmed the core value of SocialAction - integrating statistics with visualization - and further provided guidance for redesign of the tool.

Several studies have captured and characterized the work practices and analytical processes of individual or collaborative analysis through a qualitative approach. Pirolli and Card [7] studied analysts and developed a notional model of the analytic processes they follow. Chin et al. [3] conducted an observational case study with professional analysts in which participants worked on real-world scenarios, either as an individual analyst or as an investigative team. The researchers revealed various characteristics of the analytical processes of intelligence analysts, such as the investigative methodologies they apply, how they collect and triage information, and how they identify patterns and trends.

*e-mail: ykang3@gatech.edu
†e-mail: goerg@cc.gatech.edu
‡e-mail: stasko@cc.gatech.edu

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

Robinson [10] examined how analysts synthesize visual analytic results by studying domain experts conducting a simulated synthesis task using analytical artifacts printed on cards on a large paper-covered workspace. Based on analysis of video coding results, he identified several characteristics in the process of synthesis such as the use of different approaches to collaborative synthesis, a variety of organizational metaphors when structuring information, and the importance of establishing common ground and role assignment.

While these latter three studies did not evaluate specific visual analytic tools or features per se, they provide valuable implications to inform design directions for future support tools. Scholtz [12] emphasizes that the development of metrics and methodologies for evaluation is necessary to help researchers measure the progress of their work and understand the impact on users. She argues that the evaluation of visual analytic environments requires researchers to go beyond performance evaluations and usability evaluations, and proposes five key areas to be considered as metrics and methodologies for evaluation: situation awareness, collaboration, interaction, creativity, and utility.

3 Study Design

We recruited 16 graduate students (8 female) from Georgia Tech to participate in the experiment. We explicitly described the study goals and their simulated actions as an intelligence analyst to find students who would be interested and motivated by such a scenario. Participants received either a $30 or $40 gift card, depending on their setting and experiment duration, as compensation.

3.1 Task and Dataset

We told participants that they would be taking on the role of a government intelligence analyst. We gave them 50 documents, described as intelligence reports, and asked the participants to identify a hidden terrorist plot.

For this task, we adapted documents from an exercise we had learned about from a military intelligence college. Embedded across some of the documents are hints to a fictional terrorist plot with four sub-stories that support the plot. The main plot is an attack on U.S. airports with surface-to-air missiles, and the sub-stories involve the acquisition and movement of the weapons to the pertinent locations. Each document was a few sentences long. 23 of the documents contained information useful to identifying the threat. The other 27 documents described other suspicious activities but were not relevant to the main plot.

We told participants that they needed to identify the plot and ultimately write a short narrative describing the potential threat. In addition, we gave participants task sheets adapted from the VAST Symposium Contest [9], which contained tables for them to list key players, events, and locations relevant to the plot.

3.2 Settings and Procedures

We created four settings in the experiment and assigned each participant to one of the conditions. Each setting had both male and female participants. In setting 1 (Paper), we gave participants the reports as paper documents and asked them to perform the task without any technological aid. In setting 2 (Desktop), we gave participants the documents as separate text files on a computer and made Microsoft Desktop Search available to search for keyword(s) in the documents. In setting 3 (Entity), participants used a limited version of Jigsaw, in which only a modified version of the Document View (tag cloud removed) and text search capability were available. Essentially, this setting was like Desktop except that the Jigsaw Document View highlights identified entities such as people, places, and organizations in the documents. In setting 4 (Jigsaw), participants performed the task using the Jigsaw system. We provided participants in this setting with a short training video of the system three days before the session and gave them an additional 30 minutes of training at the beginning of the session. Neither of these training sessions involved information related to the task used for the evaluation.

In all settings, participants could take notes using pen and paper. For the Desktop, Entity, and Jigsaw settings, participants worked on a four-monitor computer. We gave each participant 90 minutes to work on the problem and conducted a semi-structured interview after each session. We video-taped all the sessions.

3.3 Jigsaw

Jigsaw is a system for helping analysts with the kinds of investigative scenarios encountered in this study. It is a multi-view system, including a number of different visualizations of the documents in the collection and the entities (people, places, organizations, etc.) within those documents. Figure 1 shows some of the visualizations: the Document Views (left) displays documents and highlights identified entities within them, the Graph View (right top) shows connections between documents and entities using a node link diagram, and the List View (right bottom) shows connections between entities that are arranged in lists accordingly to their type. The Jigsaw system is described in detail in [15].

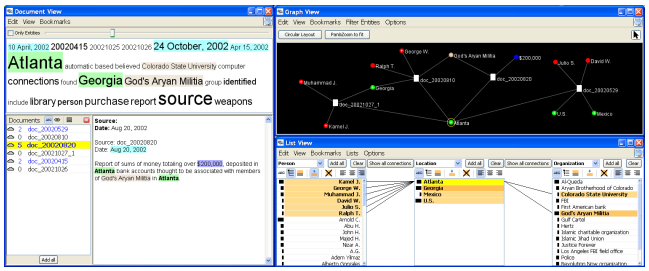

Figure 1: Jigsaw's Document View, Graph View, and List View.

A textual search query interface allows users to find particular entities and the documents in which they occur. In addition, entities and documents can be explored directly by interacting with those objects in the views. For instance, new entities can be displayed and explored by user interface operations in the views that expand the context of entities and documents. In practice these two approaches are often combined: search queries serve to jump-start an exploration and view interaction then yields richer representations and exploration.

3.4 Performance Measures

We created a solution to the exercise and described it in a short text narrative. In addition, we completed the task sheets (relevant people, events, places). Two external raters used this material to grade the anonymized task sheets and debriefings.

For the task sheets the raters awarded each correct item 1 point while misidentified items (false positives) lost 1 point. This grading rule yielded a few negative scores for participants who listed more false positives than correct answers. The maximum reachable score was 29 points. The raters also subjectively graded each narrative debriefing on a scale from 1 to 7, where 7 indicates "Highly accurate; Hits the main plot; Covers all of the supporting evidence and sub-stories" and 1 indicates "Fails to find the main plot; No relevant sub-stories; Points out irrelevant facts and events." We averaged the scores from two raters for final scores.

4 Results and Analysis

The first block of rows in Table 1 summarizes the performance results of the participants by setting. We normalized the ratings from the task sheets and the debriefing (equally weighted) to a 100-point

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Grading Task Sheet	-1.75	17	4	13	13	10.5	10.5	-3.5	5.5	-8.25	4	7.5	14.5	13.5	7	17
Grading Debriefing	2	2.5	1	5.5	3	4	1.5	3	3.5	2.5	1.5	6	6	2.5	5.5	5
Final Score	22.87	65.00	24.26	87.08	62.08	67.13	42.13	29.41	52.23	15.00	29.26	81.19	95.05	58.07	75.20	90.00
Performance	fair	very good	fair	excellent	very good	very good	good	fair	good	poor	fair	excellent	excellent	good	very good	excellent
Avg. Score/Setting	49.80				50.19				44.42				79.59			
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23
Number of Queries					19	18	48	8	23	61	59	91	44	4	26	8
First Query					40:49	19:55	2:47	12:41	1:31	0:29	0:59	3:12	0:18	5:35	25:37	4:18
Amount of Notes	many	none	many	some	many	some	few	some	some	none	none	few	some	few	few	few
First Note Taking	0:07	–	0:05	0:16	1:53	19:57	2:47	8:20	2:37	–	–	3:14	0:48	0:32	5:15	78:48
First Task Sheet	43:20	32:53	70:13	3:25	61:35	20:26	7:33	64:11	28:09	0:52	2:55	7:20	48:26	41:48	43:00	5:33
Strategy Used	OFD	OFD	BFD	OFD	OFD	OFD	FCFT	BFD	BFD	HTK	HTK	FCFT	FCFT	HTK	OFD	FCFT

Table 1: Study results and statistics, grouped by setting. The measures are explained in Section 4 and Section 5.

scale to determine a final score and grouped them into five categories (poor, fair, good, very good, excellent) via quintile rankings.

Our focus here was not on producing statistically significant differences. With such a small subject population, it seems doubtful that such results could even be found. Instead, we view these results as suggestions of overall performance and we relate them to more qualitative findings discussed later.

Within that context observe that participants in the Jigsaw setting earned excellent, excellent, very good and good ratings. If we average the final scores of the four participants in each setting, those using Jigsaw clearly outdistanced those in the other three settings that produced similar average final scores. P4 (Paper setting) and P12 (Entity setting) also performed excellently.

4.1 Activity Patterns

Because of the explorative nature of the task, we were curious about general activity patterns such as how many of the documents were viewed in total, which document was viewed most, and how many times each document was viewed. We also determined how many search queries a participant performed and when the first query was performed. For those participants who took notes on paper, we identified when they first started note-taking, as well as how many and what kind of notes they took. Additionally, we identified when each participant first began completing the task sheets.

Ten of the sixteen participants viewed all the documents at least once (second block of rows in Table 1). Curiously, all of the Paper and Desktop participants read all of the documents, but only one in each of the Entity and Jigsaw settings did so.

The frequency of search queries issued by each participant varies, ranging from 4 times to 91 times (third block of rows in Table 1). (Obviously, participants in the Paper setting could not issue queries.) Overall, those in the Entity setting tended to issue more queries and start doing so relatively early in the session. Large individual differences existed in all settings, depending on how much each person relied on queries in their analysis.

The fourth block of rows in Table 1 summarizes participants' note-taking and task sheet completion behavior. Thirteen out of 16 people took notes on paper, and those in the Paper and Desktop settings took relatively more notes. Participants mostly jotted down important names, places, and events along with the document number. Some drew a simplified map and used arrows to illustrate traces of people. Most participants started taking notes quite early in the process. In particular, those in the Paper setting typically began taking notes as soon as they started reading. The time at which each participant began to complete the task sheets varied; some people worked on them right after figuring out certain pieces of information such as repeated names, locations, or events relevant to the plot. Most read the documents and concurrently worked on the task sheets. Several participants—1, 2, 6, 9, 13, 14, and 15—started the task sheets in the middle of the process, when they had confidence about their hypothesis to some degree. P3 and P8 waited to complete the task sheets almost until the end of the session, and it turned out that they had still not still determined what to write.

4.2 Jigsaw Usage Patterns

To better understand how Jigsaw or a portion of it was used in the Entity and Jigsaw settings, we implemented a logging mechanism and recorded selected user interactions and view operations such as queries and displays of documents in the Document View (Entity setting) and all view actions in the Jigsaw setting.

Since displaying a document does not necessarily mean that the participant actually read it, we decided to impose a criterion on this measure: we consider a document as being read if it was displayed in a Document View for at least five seconds.

Figure 2, at the top, shows an overview of the usage pattern of the different views for the eight participants in the Entity and Jigsaw settings. Each row of pixels in the maps represents one minute and the color encodes the view being used (active window) by the participant at that time. Gray shows periods when participants worked on the task sheets (no active Jigsaw window). The maps for P10, P11, and P12 in the Entity setting are relatively consistent; the map for P9 is slightly different since it has longer note taking periods.

The maps for the participants in the Jigsaw setting reveal quite different usage patterns. P13 worked primarily with the Document and the Graph View (the List, Document Cluster, Calendar View were also used); P14 primarily used the List View (the Document, Timeline, and Graph View were also used); P15 focused on the List and Document View (the Graph View also was used); P16 used all the views and focused on the List and Document Cluster View.

Figure 2, at the bottom, shows a small portion of the detailed usage pattern for P16. Each pixel row represents four minutes and the colors again encode the active views. The rows are annotated with queries issued and documents viewed. This view is synchronized with the recorded video (bottom right) using a slider.

5 Discussion

5.1 Investigative Strategies

After analyzing the video and log data for the 16 sessions, we identified four common investigative strategies participants used throughout their analysis processes.

Strategy 1: Overview, Filter, and Detail (OFD)
The most commonly used strategy was "Overview first, filter and select, and elaborate on details," a strategy quite similar to Shneiderman's InfoVis mantra [13]. Six participants out of 16 performed analysis using this strategy (fifth block of rows in Table 1). They began by quickly scanning all documents and building rough ideas of the plot. While gaining an overview, most people jotted down important keywords with corresponding document numbers, drew circles and lines to indicate connections between keywords and doc-

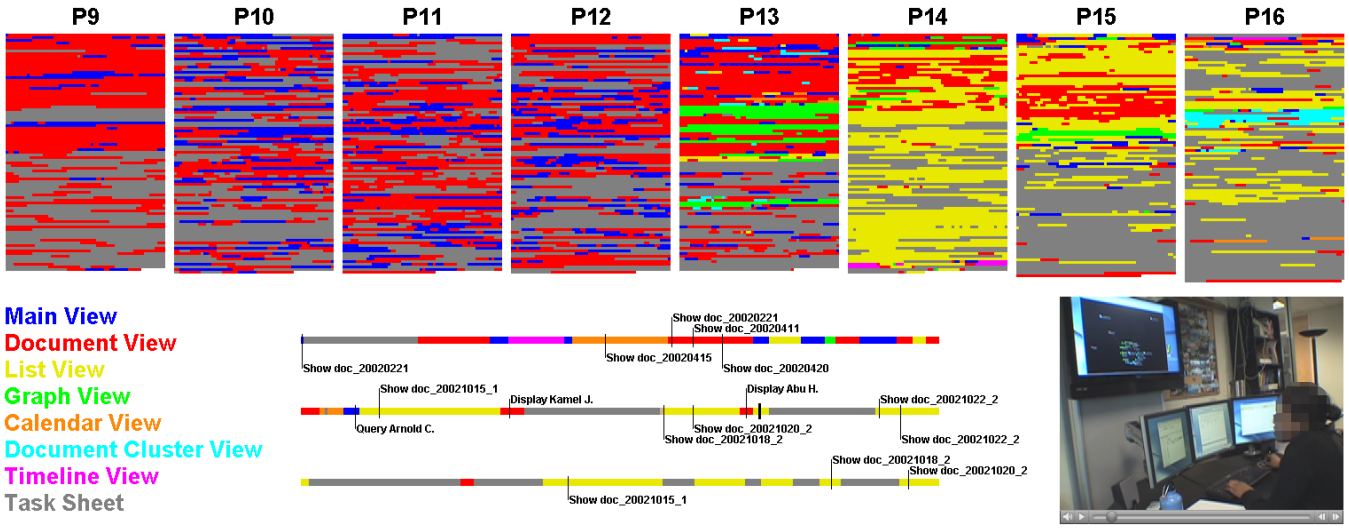

Figure 2: Overview of the Jigsaw usage patterns (at the top) and an extract from the detailed usage pattern with video for P16 (at the bottom). P9-P12 were in the Entity setting, so they accessed only the Main View and the Document View in Jigsaw. P13-P16 used the full system.

uments, and later used these notes as an index for finding relevant documents. After scanning all documents, they revisited relevant documents selectively - either by directly looking up the document or by searching for a keyword that stood out. Then they read each one carefully, extracting key information for the task sheets.

We speculate that this strategy worked well in this task because the dataset was relatively small. Participants were able to gain a rough idea of the important documents or keywords by simply scanning all documents. However, because they made a decision about the importance of each document or keyword based on their own subjective judgment, sometimes they missed important details.

Strategy 2: Build from Detail (BFD)

The strategy, "Build from detail", contrasts the previous one. Three participants used this strategy. They started the analysis from details of each document by carefully reading it. Even though they used the search function when important phrases or words arose (where applicable), it was more of an auxiliary use than a main focus. They issued relatively few queries. Instead, they focused on every sentence of the documents, in the fear of missing any relevant information. Some tried to write down important keywords for every document, which took even more time.

Because they paid attention to every detail, it was difficult for them to see the "big picture" of the plot, and therefore this strategy turned out to be least effective of the different strategies, as mentioned by one participant:.

> P8: *If I had to do it again, I'll scan through all documents several times until I get the big picture. This time, I read the documents much too carefully one by one and it took so long. I still haven't figured out what the story is about.*

Strategy 3: Hit the Keyword (HTK)

Some participants used an unexpected strategy - an intensive keyword-based exploration. They did not begin the analysis by reading a specific document, but directly looked for a few specific keywords such as "terrorist" or "Al-Qaeda". They read only the related documents and then searched for other terms that emerged during that time. This did not cover all of the documents, and these participants ignored the rest of documents that might not have been brought up.

Since the effectiveness of this strategy depended on the appropriateness of the terms chosen in the initial stage, performance varied across participants using this strategy. While P10 and P11 showed poor performance, P14 performed quite well using this strategy. He was in the Jigsaw setting, and he started using the List View even before he read any document or used the search control panel. He first added all documents in the first list, all people in the second list, and all places in the third list. Then he sorted the second list by frequency of appearance, which resulted in the most frequently appeared people moving to the top. Selecting a person's name highlighted documents that contained the name in the first column, and he read those documents in the Document View. After reading a few documents relevant to those people who were at the top of the list, he moved to the third column, places, and repeated the same process. In this way, he was able to read most of the documents relevant to important people and places. This is a similar result to those who searched for particular names and places, but it was much more efficient in that he did not have to spend time in deciding which keywords to search and which documents to read. In fact, he made only 4 search queries total. In contrast, P10 and P11 made about 60 queries but only a few of them retrieved the most vital documents, which resulted in poor performance.

Strategy 4: Find a Clue, Follow the Trail (FCFT)

The "Find a clue, follow the trail" strategy is a hybrid approach of the previous strategies, and four participants followed it. They invested some time in reading the first few documents to understand the context and find a clue, then followed the trail rigorously using search or other functionalities provided by the tool.

In theory, this may be a good strategy because the analyst's attention is focused on relevant documents only. The initial investment in reading a few documents pays off because it increases the possibility of finding the right clue. The performance of participants who used this strategy is notably good.

When we more closely examined this strategy, we found two sub-strategies. While following the trail, P7 and P12 tried to read every document in the dataset at least once and made sure they did not miss any connections. This may work for a relatively small set of documents as was present here, but as the size of a dataset increases, an issue of scarcity of attention likely will arise because the analyst must keep track of what has been read and what has not.

Jigsaw participants P13 and P16, however, did not skim the rest of documents that were not in the trail. They read only 31 and 23 out of 50 documents, respectively. Since they gained the highest scores among participants, it seems clear that they focused only on important parts of the dataset, along the trail. From the log data we identified that both read all 23 important documents and that most

of the documents irrelevant to the plot were not viewed. P16 identified one of the main players of the plot in the beginning of the analysis, and effectively explored the document collection following the person's trail.

P16: *I like this people-first approach. Once I identify key people, then things that are potentially important come up, too. I'm an impatient person and don't want to read all documents chronologically.*

This may be a fruitful strategy when there are a large number of documents. However, there still is a possibility of a dead-end if the analyst follows a wrong trail. In that case, the ability to quickly turn to another track is crucial.

P13: *I started from a name and followed it. I had this one direction for a while and realized that it wasn't a good way. I was kind of running into a dead end. Then I saw other names coming out over and over again, other areas coming out, then I got a story of what's going on.*

5.2 Jigsaw's Influence

Among the four study conditions, the group using Jigsaw generally outperformed the other groups on the whole. The worst performance of a participant in this group was "good", whereas the performance of participants in the other settings varied more. Based on observations, interviews, videos, and log analyses, we identified several benefits Jigsaw seemingly provided to users.

5.2.1 Supporting Different Strategies

Examining each participant's analysis process, we note that the four Jigsaw setting individuals used three different strategies. This suggests that Jigsaw supported different analysis strategies well. For example, as discussed in the previous section, P14 was able to do keyword-based exploration effectively using the "sort by frequency" function of the List View. P15, who used the "overview, filter, and details" strategy, used the List View to grasp the main idea about important people, their organizations and locations after quickly going through all documents in the Document View. He opened an additional List View, put all the people and all the documents in two lists, and used it as an index when he needed to revisit documents in his second iteration. P13 and P16 both used the "find a clue, follow the trail" strategy, which was effective across settings. However, we found that these two individuals performed even better and more efficiently than those who used the same strategy in the other settings.

5.2.2 Showing Connections between Entities

Showing connections between entities such as people, organizations, and places is the key functionality of Jigsaw. We had a belief that showing connections would help the analysis process, and the study clearly revealed evidence to support this. Participants using Jigsaw performed well even though they did not fully take advantage of many system capabilities, partly due to limited training and unfamiliarity with Jigsaw. Mostly, they used the List View to explore connections. Multiple participants in the non-Jigsaw settings wanted to see comprehensive connections between entities. Many of the generated notes contained important names, dates, and places. They were linked by lines and were used to assess the centrality of certain items, to understand what is important, and to decide what to examine further. The connections participants drew on paper and the functionalities they desired are similar to capabilities provided by Jigsaw. Figure 3 shows some examples. When asked about what were the most challenging aspects of the analysis, 6 out of 12 participants who did not use Jigsaw mentioned the difficulty in making connections:

P9: *Making connections was the most difficult part. I started from one person but there were so many connections around it and it was impossible to trace all the connections.*

P8: *Connecting names and documents was hard. Sometimes when two documents are related, there's no way to look it up if I hadn't marked [the connection] on each document.*

P3: *It was really hard to connect current information to what I read previously. Just too many names and places.*

Some participants also stated that they would change strategy and make connections more visible if they had to do the task again:

P3: *I'd write down important names, places, and events and put them in different shapes by type and then show connections between them by lines and arrows.*

In contrast, none of the participants in the Jigsaw setting identified connections as an issue. Rather, they focused on the challenges in organizing and keeping track of relevant information.

5.2.3 Helping Users Find the Right Clue

Finding an appropriate clue early in the analysis is crucial and sometimes even determines the entire performance. Participants often seemed to take on a kind of "tunnel vision" about what was important, which may be problematic with large document collections. Even though the dataset used in this study was relatively small, participants still benefited from Jigsaw in finding the right starting point. Tag clouds, entity highlighting, and the connections in the List View helped to find the right clue:

P9: *Entity highlighting helped a lot. I didn't have to read all the documents and still could recognize the pattern.*

P15: *I think the tag cloud is really interesting. It was helpful to see some important terms when I did not know what it is about.*

P15: *I scanned all the documents first and got some rough ideas. But still I wasn't sure where to start. Then I opened the List View to explore connections between people and locations, and I started from there.*

5.2.4 Helping Users Focus on Essential Information

Even though analysts may find appropriate initial clues, it is still important to follow the trails in an efficient manner. If relatively unimportant information diverts their attention, the investigative process may suffer no matter how quickly a good clue was discovered. We found that Jigsaw helped participants to follow the right trail and ignore irrelevant documents, thereby saving the participant's attention for important information. As we described earlier, two participants in the Jigsaw setting read only about half of the documents while the majority of other participants read all 50 documents at least once. These two Jigsaw setting participants (P13, P16) earned the two highest final scores. The other two participants (P7, P12) who used the same strategy and performed relatively well, in the Desktop and Entity settings respectively, both tried to read all the documents and keep track of other names or events while following the trail. This diverted their attention and hindered them from totally focusing on the main story. P12 (Entity setting) stated:

P12: *Because I searched for key phrases, read relevant stories, and went back to another document, I tended to lose track of all the dates that were going on.*

5.2.5 Reviewing Hypotheses

During analysis, the participants generated hypotheses about the hidden plot and gathered evidence that could support their hypotheses. Two of the Jigsaw setting participants found the Graph View to be useful as a confirmatory aid. P15 explored the dataset primarily using the Document View and the List View, and narrowed down to the most three important persons surrounding the plot. Then he used the Graph View to review his hypothesis and to check whether they were really key people in the plot, by quickly reviewing related documents and their connections to other entities.

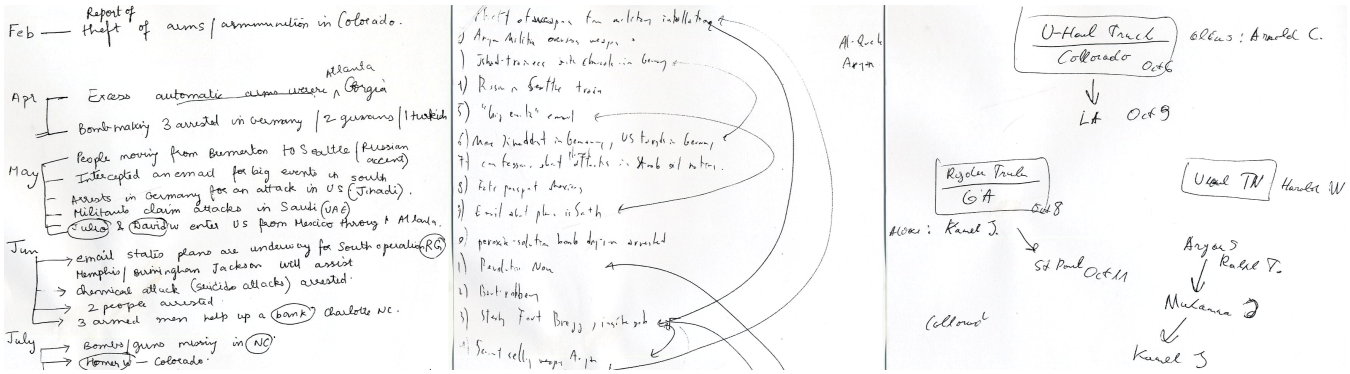

Figure 3: Notes made by participants not using Jigsaw.

P13: *Once I got some names and locations that I wrote down on paper, about one and half pages, I used the Graph View to get an idea of what's related. This is more confirmation rather than fact finding in that case. Everything in the middle was basically what I already knew about ... so ... I used it to validate what was going on. It was helpful but in a different sense. It's not about finding new facts but just asking like, was I right? What were things the graph is showing?*

5.3 Observations on Sensemaking

Pirolli and Card have proposed a Think Loop Model of Sensemaking [7] consisting of two major loops, a foraging loop and a sensemaking loop, and several intermediate stages including "shoebox", "evidence file", and " schema". We observed study participants and how their actions related to this model.

5.3.1 Diversity in Sensemaking Processes

While the model is not linear and can proceed top-down or bottom up with many loops, we found that the sequence of analysis significantly differed across individuals even in the same task with the same dataset. Some participants followed the sequence linearly with iterations; they extracted and jotted down important information while reading the documents, then organized the information according to a certain scheme such as time and location, eventually leading to a hypothesis. Some participants started organizing information as soon as they read documents, either by filling out the task sheet or drawing timelines/maps on paper, thus skipping the process of building an evidence file. Once they created a hypothesis, they took out snippets of information from the schema that did not support the hypothesis. On the other hand, some participants immediately started from a hypothesis without the schema stage, and then worked on organizing to confirm the hypothesis. In this case, the schematizing stage took place at the end of the analysis process.

Individual differences also existed in each stage of the model. For example, the "read & extract" stage, in which evidence files are collected from the shoebox, exhibited individual differences. When encountering much unfamiliar information, it is not easy to extract nuggets of evidence simply by reading documents; the analyst usually needs some criteria to decide what to pull out. In our study, some participants started from a specific set of people and extracted information related to those people. Those who used location as a criterion gathered all information related to specific cities or countries. Participants also extracted evidence files based on specific events such as arms thefts or truck rentals. Although participants used different approaches in this stage, it did not make a significant difference in the overall analysis process because the evidence files gathered tended to be similar regardless of the extraction criteria, as long as the analyst carried out the process thoroughly.

5.3.2 Power of Schematizing

It was the schematize stage that showed the most significant variance between individuals. During this stage, it seemed that each person had his/her own preferred organizational scheme such as a timeline, map, or diagram. For example, while most people wanted a timeline, the representations they envisioned were all different. Some people wanted a timeline organized by person and event; some wanted a timeline by location; others wanted a timeline categorized by story. Clustering was another organizational method employed by participants, but the classification scheme varied - by organization, by location, and by connectedness. The variances in this stage seemed to affect the entire analysis performance.

The time at which a participant first reached the schematize stage and how much effort the participant invested in this stage significantly affected the performance. When we further examined those who performed well independent of the setting, we found a commonality that all of these people spent considerable time and effort in organizing information. Most people used the task sheet as a tool for gathering their thoughts since the task sheet was structured by certain schemes (e.g., people, events, and locations). During the interviews, many participants explicitly described how completing the task sheet helped their sensemaking process.

P12: *There were a couple of themes that kept popping up. And so I think I was more mentally taking notes about those and then once I started feeling there were too many references and things got intertwined in my head, I started using these task sheets to drop them down and organizing.*

P9: *Filling out the task sheet - all the events by date - was really helpful. At first, I started from people's names, but at some point I jumped to the events instead of names, and found that much more helpful to make sense of the story. Jotting down didn't help that much.*

As the quotes indicate, participants did not expect the task sheets to help their investigation at first, but they noted the sheets' usefulness at the end. Note, however, that the participants were simply marking down entities from the documents on the task sheets, not new or meta information. The key difference was that the entities were being organized under a particular point of view.

Those participants who did not build schema or undertake some organizational process performed poorly on both task sheets and debriefing. Some of them did take a fair amount of notes, but no process of organizing the notes followed. Simply rewriting information without imposing an organizational scheme did not appear to help the sensemaking process.

5.3.3 Insight Acquisition

It is still difficult for us to identify exactly when people gained a key insight during the investigative process. When we asked the participants how they knew they were progressing towards the goals,

the common answer was "when the pieces of a puzzle started being connected and put together." Rather than a spontaneous insight occurring (the "light bulb going on"), insight seemed to form continuously throughout the investigation, not unlike that described by Chang at al. [2]. Participants had a difficult time identifying when they "got" the plot. P13 who gained the highest score, when asked about this, stated:

P13: *Well, that's interesting. I don't know. Names coming up a lot, there's all these relationships like, for example, there seems to be Colorado and Georgia, and there were organizations there. You have this idea that just validates itself.*

5.4 Design Implications for Investigative Analysis Tools

The study and its results suggest several design implications for visual analytics systems for investigative analysis. Investigative analysis tools need to support analysts in finding appropriate starting points or clues and then following the trail of these clues efficiently. The study showed that the "find a clue, follow the trail" analysis strategy generally led to a positive result. Further, the performance of those participants who were able to focus only on relevant documents was outstanding. Investigative analysis tools need to direct the analyst's attention to the most critical information.

The study demonstrated that people do frequently move between stages of the Think Loop Model, particularly in the middle parts of the model. Investigative analysis tools should allow smooth transitions between the "shoebox", "evidence file", and "schema" stages so that different sequences of the sensemaking process can be supported. Currently, the focus of Jigsaw is on the "shoebox" and the "evidence file" stages, but it lacks powerful support for the "schematize" stage. While Jigsaw does appear to help analysts finding nuggets of information effectively, it does not really support putting those pieces of evidence together. In other words, analysts may easily discover the pieces to be put in a puzzle and have a sense of which piece goes where, but they should also receive help in putting the pieces together. The ability to work on extracting evidence files and organizing them into a schema will significantly help the sensemaking process.

For Jigsaw to be a comprehensive investigative analysis tool, it is crucial for the system to include a workspace in which the analyst can simply drop/paste entities, draw connections between them, and add annotations, capabilities found in systems such as Analyst's Notebook [4], the Sandbox [17], and Entity Workspace [1]. Several participants pointed out this issue as well, including:

P16: *Remembering what I had already found was hard. Keeping track of names was really hard, too. When I was reading a document about truck rentals in different cities, I remembered I read a similar document before. Oh yeah, there was somebody who rented a truck from Chicago to Minneapolis but then I forgot his name and it was really frustrating.*

P12: *I'd probably do something like this [the task sheets] but either spread them out or do it on notepad to give me more room so that I can just cut and paste things and move things around.*

When supporting the "schematize" stage, developers of investigative analysis tools should consider that individuals will choose different organizational metaphors or schemes. For example, even for a timeline, individuals imagined many different types of timelines and they were quite insistent about this approach. Rather than providing one fixed schema, allowing flexibility and room for customization will be beneficial. One participant wanted to have the ability to organize a timeline by "story", which also requires flexibility in organizational schemes.

P7: *It would be good to have categorized keywords or events with relevant people/activities sorted by time. For example, I can have multiple stories such as passport, truck rental, Al-Qaeda, things like that, and under each keyword, all related people/activities are listed in a sequential order.*

Tool developers may consider having a system suggest a few organizational schemes when the analyst has created a significant evidence file but still does not have a schema, particularly for novice analysts. Staying too long at the "evidence file" stage appears to impede the analysis process so suggestions of organizational schemes may be beneficial.

It is not uncommon for an analyst to confront a dead-end or find evidence that refutes an existing hypothesis. Investigative analysis tools need to support the analyst to find appropriate next steps or alternatives by making the milestones of the investigative process explicit. In this way, the analyst can come back to the point where she/he was earlier and start over from that point. This also ensures that the analyst can proceed further without being too concerned about keeping track of past states.

P16: *I was managing too much information. While in the analysis, I was really afraid of [getting] out of track, so I didn't want to go further at some point. I always kept coming back to the previous stage because I wanted to keep the main story line.*

5.5 Evaluation Implications for Investigative Analysis Tools

The study also suggested a number of ways to help evaluate investigative analysis systems. By comparing system usage to more traditional methods but otherwise giving participants freedom to perform as they wished, we feel that the findings are both realistic and provide ample grounds for contextual analysis and comparison.

We also suggest that the evaluation of investigative analysis tools focus on collecting more qualitative data. While quantitative data is useful when a solution is well-defined and measurable, the nature of investigative analysis is exploratory and flexible. It may be too limiting to assess the value of a system solely based on statistical results. Identifying best practices supported, particular pain points, and future design requirements can be better achieved through interviews and observations. When possible, we suggest using quantitative data such as usage log files and analysis scores to help understand qualitative results.

Findings from the study suggest potential questions to be answered in the evaluation of investigative analysis tools:

- Does the tool help to provide information scent appropriately, thus helping to find initial clues?
- Does it guide the analyst to follow the right trail, without distraction?
- Does it support different strategies (sequences) for the sensemaking process? That is, does it support smooth transitions between different stages of the model?
- Does it allow flexibility in organization?
- Does it help to find appropriate next steps when encountering a dead-end?
- Does it facilitate further exploration?

In this study, we identified and used several metrics, which are broadly applicable to evaluation of investigative analysis tools:

- The number of important documents viewed, relative to the entire collection
- When the analyst first started creating representations such as notes and drawings
- The quantity of representations created

We also suggest two possible metrics for evaluating investigative analysis tools:

- Amount of time and effort in organizing
- Amount of time the analyst spent in reading/processing essential information

5.6 Study Limitations

The study had several limitations that likely affected our findings. First, our participants were graduate students, not professional analysts. None of the students had formal training in investigative analysis, so it is unclear if and how the results would change by using a professional analyst participant population. (Note that it would likely be extremely difficult to gain access to enough professional analysts to conduct a comparative study such as this one.) All of the student participants were familiar with research, however, so we believe that they were at least somewhat representative of the behavior one might expect from professionals.

Though it would have been interesting to see if correlation between an investigative strategy and a setting exists, the small sample size (16) did not allow us to examine the relationship. For example, we could take the "Find a clue, follow the trail" strategy and see if a particular setting better supported that strategy compared to other settings. Examining that correlation yields 16 (4 settings x 4 strategies) experimental conditions and thus requires many more participants.

We compared Jigsaw to other traditional tools, but not to other visual analytics systems. Comparing the usage of our tool to other existing systems developed for investigate analysis would have generated more insightful findings and implications.

For the study, we used a relatively small document collection, which likely would not be the case in reality. The collection size was chosen to make the experiment feasible in a reasonable amount of time. We speculate that some of the findings would only be amplified when working with larger document collections. Throughout the discussion we identified numerous situations where larger datasets would place even more importance on highlighting connections, following evidence trails, and organizing data and evidence.

The analytic scenario used in the study was a *targeting* scenario, one in which analysts seek to "put the pieces together" and identify a hidden plot. Many investigative scenarios have no clear, specific solution, however. Instead, they involve general knowledge acquisition over a long time period. Developing evaluation strategies and measures for these scenarios appears to be particularly challenging.

It was clear to us that even with the two-phased training for Jigsaw, participants in that condition still overlooked many useful capabilities of the system. With further experience and training, we would hope that the system would be even more beneficial. In particular, the experience of performing one investigation like this appeared to place participants in a position where they could better understand system capabilities if given further training.

6 Conclusion

While many researchers in the visual analytics community firmly believe that new visual analytics technologies can benefit analysts, showing that is the case is still a challenging proposition. Clearly, one necessary step is to compare the use of new technologies to existing, more traditional methods. We conducted an experiment comparing students performing an investigative analysis exercise under one of four conditions. While lacking the size and depth to identify statistically significant differences, the study nonetheless suggested that visual analytics systems such as Jigsaw can benefit investigative analysis. Two aspects of Jigsaw turned out to be particularly helpful: showing connections between entities and narrowing down the focus.

Beyond that, this research makes several contributions to the visual analytics community:

- It provides an experimental design and methodology that others can emulate and apply.
- It describes how participants used visualization for analytic benefit and how its absence amplified challenges and difficulties.
- It provides a description of four analytic strategies employed by participants in seeking to identify a hidden plot.
- It identifies a number of design suggestions and capabilities to make visual analytics investigative analysis systems more effective.
- It suggests new evaluation metrics and qualitative factors for conducting experiments on these types of systems.

Evaluation of visual analytics systems must progress in step with new technical development for continued progress. Understanding how and why systems aid analysts will help to inform future designs and research. Our study provides initial evidence and insight in this area, and sheds light on many challenging open questions.

Acknowledgements

This research was supported in part by the National Science Foundation via Award IIS-0915788 and the National Visualization and Analytics Center (NVAC^{TM}), a U.S. Department of Homeland Security Program, under the auspices of the Southeast Regional Visualization and Analytics Center.

References

[1] E. Bier, S. Card, and J. Bodnar. Entity-based collaboration tools for intelligence analysis. In *IEEE VAST*, pages 99–106, Oct. 2008.

[2] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining insight for visual analytics. *IEEE CGA.*, 29(2):14–17, 2009.

[3] G. Chin, O. A. Kuchar, and K. E. Wolf. Exploring the analytical processes of intelligence analysts. In *ACM CHI*, pages 11–20, April 2009.

[4] i2 - Analyst's Notebook. `http://www.i2inc.com/`.

[5] D. H. Jeong, W. Dou, H. Lipford, F. Stukes, R. Chang, and W. Ribarsky. Evaluating the relationship between user interaction and financial visual analysis. In *IEEE VAST*, pages 83–90, Oct. 2008.

[6] A. Perer and B. Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *ACM CHI*, pages 265–274, April 2008.

[7] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *International Conference on Intelligence Analysis*, May 2005.

[8] C. Plaisant. The challenge of information visualization evaluation. In *AVI*, pages 109–116, May 2004.

[9] C. Plaisant, J.-D. Fekete, and G. Grinstein. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE TVCG*, 14(1):120–134, 2008.

[10] A. Robinson. Collaborative synthesis of visual analytic results. In *IEEE VAST*, pages 67–74, Oct. 2008.

[11] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE TVCG*, 11(4):443–456, 2005.

[12] J. Scholtz. Beyond usability: Evaluation aspects of visual analytic environments. *IEEE VAST*, pages 145–150, Oct. 2006.

[13] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, pages 336–343, Sept. 1996.

[14] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *BELIV*, pages 1–7, May 2006.

[15] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[16] J. J. Thomas and K. A. Cook. *Illuminating the Path*. IEEE Computer Society, 2005.

[17] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The sandbox for analysis: Concepts and methods. In *ACM CHI*, pages 801–810, April 2006.