UNDERSTANDING THE BI-DIRECTIONAL RELATIONSHIP BETWEEN ANALYTICAL PROCESSES AND INTERACTIVE VISUALIZATION SYSTEMS

by

Wenwen Dou

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2012

Approved by:

____

Dr. William Ribarsky

____

Dr. Jing Yang

____

Dr. Robert Kosara

____

Dr. Heather Richter Lipford

____

Dr. Mark Faust

ABSTRACT

WENWEN DOU. Understanding the bi-directional relationship between analytical processes and interactive visualization systems. (Under the direction of DR. WILLIAM RIBARSKY)

Interactive visualizations leverage the human visual and reasoning systems to increase the scale of information with which we can effectively work, therefore improving our ability to explore and analyze large amounts of data. Interactive visualizations are often designed with target domains in mind, such as analyzing unstructured textual information, which is a main thrust in this dissertation.

Since each domain has its own existing procedures of analyzing data, a good start to a well-designed interactive visualization system is to understand the domain experts' workflow and analysis processes. This dissertation recasts the importance of understanding domain users' analysis processes and incorporating such understanding into the design of interactive visualization systems.

To meet this aim, I first introduce considerations guiding the gathering of general and domain-specific analysis processes in text analytics. Two interactive visualization systems are designed by following the considerations. The first system is ParallelTopics, a visual analytics system supporting analysis of large collections of documents by extracting semantically meaningful topics. Based on lessons learned from ParallelTopics, this dissertation further presents a general visual text analysis framework, I-Si, to present meaningful topical summaries and temporal patterns, with the capability to handle large-scale textual information. Both systems have been evaluated by expert users and deemed successful in addressing domain analysis needs.

The second contribution lies in preserving domain users' analysis process while using interactive visualizations. Our research suggests the preservation could serve multiple purposes. On the one hand, it could further improve the current system. On the other hand, users often need help in recalling and revisiting their complex and sometimes iterative analysis process with an interactive visualization system. This dissertation introduces multiple types of evidences available for capturing a user's analysis process within an interactive visualization and analyzes cost/benefit ratios of the capturing methods. It concludes that tracking interaction sequences is the most un-intrusive and feasible way to capture part of a user's analysis process. To validate this claim, a user study is presented to theoretically analyze the relationship between interactions and problem-solving processes. The results indicate that constraining the way a user interacts with a mathematical puzzle does have an effect on the problem-solving process. As later evidenced in an evaluative study, a fair amount of high-level analysis can be recovered through merely analyzing interaction logs.

ACKNOWLEDGEMENTS

Though a thesis is an encapsulation of the work throughout my Ph.D. life, in truth this dissertation would never have been possible without the support of my mentors, family, and friends.

As a graduate student, I was very fortunate to benefit from mentors and colleagues with a wide range of research perspectives. At UNC Charlotte, my advisor William Ribarsky has been both a teacher and a supporter. He shaped my thinking and helped me develop my vision for future research. Dr. Jing Yang and Robert Kosara have sparked my interest in the field of visual text analytics. Dr Heather Richter Lipford introduced me methodologies and perspectives from the HCI community. Dr. Mark Faust has introduced me to a fascinating field - Cognitive Science, which in turn provides theoretical grounds for my research. Russ Burtner at the Pacific Northwest National Laboratory demonstrated the value of creativity and aesthetics in the research process and contributed to my thinking. Antonio Sanfilippo have sparked my interest in coupling machine learning methods with Visual Analytics and he also made my internship at PNNL joyful. Last but not least, Michelle Zhou from IBM research, has given me valuable feedback and inspirations on my research in this thesis.

My dissertation research was generously supported by the National Science Foundation and the Department of Homeland Security. I'm also indebted to the participants of the studies I have conducted over the years.

I was lucky to be surrounded by supporting fellow graduate colleagues. A particular acknowledgment to Remco Chang, who provided my guidance and support especially in the earlier stage of my graduate study. Other collaborators, including Caroline

Ziemkiewicz, Dong Hyun Jeong, Tom Butkiewicz, Alex Godwin, and Lane Harrison have also contributed to my fulfilling research experience at UNC Charlotte.

Finally, I dedicate this dissertation to the family and friends who make life worth living. To Yinghua Zhang, Cindy Zhao, and Yong Ye, for their support. To my husband, Xiaoyu Wang, who provides me both joy in my life and valuable insights on my academic research. And to my parents Xiulan Chu and Hongxi Dou, who taught me to be a better person.

Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Visual analytics is the science of analytical reasoning facilitated by visual interactive interfaces [115]. Visual analytics integrates new computational and theory-based tools with innovative interactive techniques and visual representations to enable human-information discourse. There exists a tight relationship between a user's analysis process and the interactive visualization system designed to facilitate such a process. On the one hand, the visualization system is designed by incorporating as much general and domain-specific analysis processes as possible. On the other hand, when the visualization system is in use, the individual analysis processes are carried out through user interacting with the system and therefore reflected by the interactions between the user and the system.

Understanding the intimate relationship between analysis processes and interactive visualization systems brings profound impacts. First, the understanding could inform future visualization system design on how to better incorporate general and domain-specific analytical processes so that the system fits more naturally to the user's analysis flow. Second, the understanding could in turn advise how to extract an individual's analysis processes when using interactive visualization systems. The extracted analysis process could potentially be used for multiple purposes such as self-recall, reporting, or training. Last, the collected individual's analysis processes could be analyzed and further utilized by the interactive visualization systems to better support user's analysis processes.

1.1 Dissertation Problem and Approach

This dissertation focuses on understanding the relationship between analytical processes and interactive visualization systems. In particular, we examine how analytical processes are gathered and then incorporated when designing interactive visualization systems. We also study whether and how an individual's analytical processes can be recovered from examining evidences collected (such as interaction logs and annotation) during the use of interactive visualization systems.

Interactive visualization systems advances analytical reasoning, which is central to analysts' task of applying human judgments to reach conclusions from a combination of evidences and assumptions [98]. Therefore, from a design perspective, it is essential for interactive visualization systems to support analytical processes in general and within target application domains. In order to provide such support, interviewing potential users and even observing users working in their natural work environment is common practice before designing an interactive visualization system. Information gathered from the interviews are then translated into design guidelines and embedded in both visual representations and interaction techniques. The resulting interactive visualization system can then support the analysts' analytical flow by providing insightful patterns and desired information on demand through user interaction.

It is through the interactive manipulation of a visual interface, the dialogue between a user and a visualization could be established and maintained; It is also through the interactive manipulation, a user is able to explore and analyze the underlying data within a visual interface. Therefore, a fair amount of the analysis/reasoning is reflected in the process of interacting with the visualization system. With visual analytics tools becoming more sophisticated and prevalent in the analysis communities, it is now apparent that not only is the final analysis product important, how the analyst arrived at the conclusions is also essential. Since the analysis process is often complex and iterative, there is a real need to help the analysts keep track of their thoughts and procedures [77]. In addition, the process of reaching final conclusions through interacting with visualization systems bears tremendous experience and expertise. If such expertise could be partially recovered after the analysis, the results could potentially be used for versatile purposes, such as generating a report of the analysis process, training novices, further improving the interactive visualization system, etc. In fact, research has shown that merely studying the interaction logs can provide system usage information [52] and even a peek into user's high-level reasoning processes [37].

To illustrate how we incorporate general and domain specific analysis processes into an interactive visualization system, we provide several application examples, specifically in the domain of text analytics. To showcase how we extract the analysis process from evidences recorded when using such system, we first present a study to evaluate the possibility of doing so on a theoretical note. We then present a combination of applications of evaluations to demonstrate our findings.

1.2 Dissertation contribution

This dissertation raises awareness of the bi-directional relationship between analysis processes and interactive visualization systems. The contributions can be categorized into two areas:

1. Design considerations for incorporating general and domain-specific analysis processes into interactive visualization systems.

- We analyze various domains in the context of text analytics and present several common and domain-related tasks. The resulting tasks and sub-tasks well reflect users' analysis processes when making sense of large text corpora. Support for text analysis processes are later incorporated into interactive text visualization systems.

- The design and evaluation of novel text analytics environments - *ParallelTopics* and *I-Si*, based on a thorough understanding of target users' analysis processes. ParallelTopics is a visual analytics system supporting analysis of large collections of documents by extracting semantically meaningful topics. Based on lessons learned from ParallelTopics, this dissertation further presents a general visual text analysis framework, *I-Si*, to present meaningful topical summaries and temporal patterns, with the capability to handle large-scale textual information. Both systems have been evaluated by expert users and deemed successful in addressing domain analysis needs.

2. Theoretical and practical demonstration of extracting analysis process from evidences recorded during analysis with interactive visualization systems. In particular,

we focus on analyzing the interactions between user and computer since it is through the interactive manipulation, the analysis process is supported within visualization systems. More specifically, this contribution can be divided into two sub-categories:

a. Studying the outcome of the analysis process under different interaction constraints validates the importance of interaction to a user's analysis process. Experimental results have shown that constraining the way a user could interact with a problem significantly affects the outcome of the problem solving process.

b. Examining how much of a high-level analysis process can be extracted from merely analyzing interaction logs recorded during the analysis. According to our study, a large amount of high-level reasoning process when using an interactive visual interface could be recovered from analyzing sequences of user interactions. Our findings have sparked a lot research interest on capturing a user's analysis process within visualization through recording user interactions and other means.

1.3 Dissertation Outline

Chapter 2 begins by covering related work and background. The remainder of the dissertation is organized into two areas: an investigation of how to incorporate analysis processes into interactive visualization systems during the design stage, followed by analyzing how to in turn extract individual analysis processes from the use of interactive visualization systems.

1.3.1 Design Considerations and Systems for Incorporating Analysis Processes

Chapter 3 introduces design considerations for incorporating analysis processes into interactive visualization systems during the design stage. We consult the Sensemaking model [94], ethnography methods well-studied in the HCI community, as well as several visualization task taxonomies to derive general analysis processes. We combine results from our interviews with target users and findings summarized by other researchers to present a list of domain-specific tasks in the field of text analytics.

Chapters 4 and 5 introduce examples of incorporating general and domain-specific analytical processes into interactive visualization systems in the area of text analytics. Chapter 4 presents ParallelTopics, a visual analytics system designed to help users to analyze longer text documents such as scientific publications and research proposals. In particular, we focus on the gathering of the general and domain-related analysis processes and how we designed ParallelTopics based on the collected information. Chapter 5 introduces a framework called *I-Si*, which is a pipeline for analyzing large-scale text corpora. In addition to the interactive visualization component, the I-Si framework also leverages parallel computing to process a large amount of textual data. The choice of incorporating a parallel computing component into the pipeline is to address the need for analyzing large scale textual information such as data harvested from social and conventional media, thus to support the analysis process of identifying consistent or bursty trends from a large amount of textual data.

1.3.2 Extracting Analysis Process from Recorded Usage of Interactive Visualization Systems

Chapter 6 through 9 describe means to extract high-level analysis process from evidences collected during the use of visual analytics systems.

Chapter 6 introduces an overview of the types of evidences available for recording during a user's analysis process within an interactive visualization system. This chapter further analyzes the benefits and drawbacks of capturing each type of evidence. Based on our analysis, studying interaction logs has the most benefit-to-cost ratio with respect to both end users and researchers, we then focus on the capturing of interaction in subsequent chapters.

Chapter 7 presents a study that theoretically analyzes the relationship between interactions and problem-solving process. In particular, we study how constraining users interactions could affect the outcome and strategies developed while solving a mathematical puzzle. The results indicate that constraining the way a user interacts with the puzzle does have an effect on the problem-solving process.

With the theoretical support from chapter 7, chapter 8 introduces a study that examines how much of the high-level analysis process could be derived through merely analyzing interaction logs. The chapter describes in detail our approach of capturing interaction sequences and analyzing such sequences. The chapter further discusses the implications for improving capturing and analysis of interaction logs.

Finally, chapter 9 summarizes the contributions of this dissertation, describes recent developments, and outlines remaining challenges for understanding the intimate relationship between analysis processes and interactive visualization systems.

CHAPTER 2: BACKGROUND AND RELATED WORK

2.1 Models of Visualization and Analysis Process

As the field of information visualization and visual analytics mature, many visualization models have been proposed. Some models focus on providing guide to the creation and analysis of visualization systems [20, 19, 25]. For example, Card presented a Visualization Reference Model [20] that emphasizes , among other things, the specific mappings of data tables into visual structures, and the interactive effects of human interactions with these mappings. Chi [24] divides existing visualization techniques into several data categories (scientific visualization, geographic InfoVis, 2D, multi-dimensional, information landscapes, trees, networks, web, and text). He extends Card's reference model into a Data State Reference Model [25] in order to isolate common operational steps within each visualization type.

More recently, Munzner [83] proposed a nested model for both the visualization design and validation. The four layers of the nested model provide not only guidelines for visualization design, but also prescriptive guidance for determining appropriate evaluation approaches. In the nested model, the most fundamental layer that all other layers are built upon is "Domain Problem and Data Characterization", at which a visualization designer *must* learn about the tasks and data of target users in the target domains. This step is analogous to our focus of understanding the users analysis process in their own domains. Each domain usually has its own vocabulary for describing its data and problems, and there is usually some existing workflow or process of how to solve the problems. Munzner further provided guidelines and methods to ensure that the problems of the target audience are clearly understood by visualization designers. Other work has also noticed the importance of understanding the analysis process before designing a visualization system. Wang et al. has proposed a two-stage framework for designing visual analytics systems in organizational environments [119]. In particular, understanding the potential users' analysis processes is the most essential step in the first stage, namely the "observation and design stage". To understand the analysis processes, visualization designers must communicate well with domain users. Some of the challenges inherent in bridging the gaps between designers and target users are discussed by Van Wijk [125].

2.2 Designing Visual Encoding and Interaction Techniques

Among aforementioned visualization models, the design of visual encodings has received a great deal of attention in the foundation of information visualization literature, starting with the influential work by Mackinlay [81] and Card et al. [20] (Chapter 1). The literature provides comprehensive guidelines on the choices of visual encodings given data types, and further presents criteria to measure the expressiveness and effectiveness of graphical presentations.

In contrast, the theory of interaction design for visualization is less developed. Early work focused on the use of interactivity, such as brushing techniques to select and highlight visualized data points [12]. Based on years of visualization design experience, Shneiderman [105] proposed a task by data type taxonomy for information visualization, providing a list of tasks that visualizations should support and a classification of the different data types subject to visualization. In particular, he identified the interactive tasks of obtaining an overview of data, zooming in on items of interest, filtering out uninteresting items, getting more details on demand, highlighting relationships between items, providing an interactive history, and extracting and exporting collections of data. The interactive tasks provide an important categorization of interaction techniques.

More recently, Yi et al. [127] presented a survey that covers seven general categories of interaction techniques widely used in the InfoVis community. The survey is a step forward in developing a framework to design and evaluate interaction techniques. Lam [74] extended Yi's taxonomies and presented a descriptive framework of interaction costs that provide understanding of how interaction can contribute to visualization use and how designs can fall short in supporting these roles. Lam presented extensive evidences from visualization literatures on interaction costs and further divided the costs into seven categories. The framework suggests a need to diversify from the traditional focus on visual encoding to cover user interactions as well.

Based on the number of models and guidelines provided regarding visual encodings and interaction techniques, there seems to be a lack of models on designing and evaluating interaction techniques in the literature. Our work contributes to this area in that we present means to capture interactions between a user and a visualization system and analyze the captured user interactions in informative manners.

2.3 Provenance

As noted by Kindlmann [64] and Silva et al [108], the lack of reproducibility of visualization research has the potential of hindering the advancement of visualization as a science. They argue that in order to recreate and extend specific visualization results, knowing the complete process of how the results are generated is just as important as the techniques used and the final outcome. This process of recording how a user interacts with a visualization is sometimes referred to as provenance tracking, which is defined by Anderson et al as "the logging of information about how data came into being and how it was processed."

2.3.1 Data provenance

Data provenance refers to the logging of low-level interactions. It is the most prevalent type of provenance tracking, and is closely related to the undo/redo functions in nearly all applications today. The primary goal of these systems are often to capture and archive a user's interactions for the purpose of replaying the user's session at a later time. In interactive visualization, one of the most notable systems in data provenance is the GlassBox system by Greitzer [45, 29] which records low-level user's interactions in an analysis environment (such as copy, paste, mouse clicks, window activation, etc). In scientific visualization, VisTrails is an open-source provenance management system that provides infrastructure for data exploration and visualization through workflows [11]. The stored provenance allows the users to query, interact, and understand another session histories with the visualization tool. Jankun-Kelly et al. further generalized how to capture the visualization states as a set of parameters and actions into recordable form [61] that they referred to as the P-Set Model. This model is complete in that every user interaction with a visualization can be described within it.

2.3.2 Information provenance

Groth and Streefkerk [47] recently coined the term “information provenance” to distinguish systems that capture low-level user interactions from systems that record the information discovery process in using a visualization. In their model, they focus on recording the user’s interaction independently from the data in a way that the same set of logged interactions can be applied to a different dataset. Additionally, a user can attach annotations to the user’s interactions to further add semantic information. Under Groth and Streefkerk’s definition, many recent visualization systems that record user interactions incorporate tracking of information provenance. In GeoTime, a user’s annotations retain semantic connections to their corresponding events as well as the patterns displayed in a 3D representation [40]. Jeong et al. integrated tracking functions into a financial visualization tool and recorded semantic-level interactions that are relevant to the specific domain [63]. Heer et al. presented methods for both capturing semantic interactions within an information visualization system as well as the mechanism for reviewing, editing, and annotating on those interactions [52]. Similarly, in the Aruvi system developed by Shrinivasan et al., the user’s interactions are automatically stored into a visible history tree [106]. The user can also manually construct the state of the discovery using an interactive node-link diagram, which provides additional detail behind the user’s interactions. Lastly, Gotz and Zhou incorporated automatic tracking of information provenance in their system HARVEST (which they referred to as insight provenance) and suggested a taxonomy of various types of action-tiered user interactions [43].

Van Wijk has presented a model of visualization that describes the flow and relationship between a user and a visualization [118]. The model presented what are exchanged between the two parties, which sheds lights into what we can capture as a result. More specifically, there are two connections, I and dS/dt, between the user and the visualization. I stands for the images generated by the visualization that are perceived by the user. And the connection dS/dt represents the changes in the parameters of the visualization initiated by the user (through the use of a mouse, keyboard, or other input devices) that are applied to the visualization to generate the next sets of images I. Both of these connections can be captured directly within the visualization during a user's exploration process by performing visualization state capturing and interaction logging respectively. The captured evidences provide us a way to peek inside the blackbox - the analysis process inside a human brain.

2.3.3 Utilizing Captured Provenance

Aside from reviewing a user's interaction history, there has been little research in how either data or information provenance could be used. While all of the aforementioned systems have noted on the benefits of capturing provenance, including communication, evaluation, training, etc., the details of how provenance could be utilized to achieve such benefits is sometimes unclear. A notable exception is the Active Reports project at Pacific Northwest National Laboratory [91, 90] where the focus is not on capturing a user's interactions, but on integrating the provenance into a live reports such that sections of the reports are linked to the provenance and analysis products generated by visualization systems. Similarly, Dou et al. examined the benefits of information provenance captured in a financial visual analytical tool by comparing the captured information provenance to the original user's analysis process [37]. Their results indicated that information provenance does not equate exactly to the analysis process and the relationship between the two varied depending on the stages of the analysis.

CHAPTER 3: INCORPORATING ANALYTICAL PROCESSES INTO INTERACTIVE VISUALIZATION SYSTEMS

In this section, we introduce design considerations for incorporating analysis processes into interactive visualization systems during the design stage. We first consult the Sensemaking model [94], and several visualization task taxonomies [129, 10] to derive general analysis processes. In order to illustrate how to interact with domain users to study domain-specific analysis processes, we consult ethnographical methods well-studied in the human computer interaction (HCI) community [35, 104]. We then combine results from our interactions with target users and findings summarized by other researchers to present a list of domain-specific tasks in the field of text analytics.

3.1 General analytical processes

Taxonomies of user tasks	
Zhou and Feiner (1998) [129]	Relational visual tasks (associate, background, categorize, cluster, compare, correlate, distinguish, emphasize, generalize, identify, locate, rank, reveal, switch) and direct visual organizing and encoding tasks (encode)
Amar, Eagan, and Stasko (2005) [11]	Retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster, and correlate

Figure 1: Taxonomies of user tasks [127].

A good start to understanding general analysis process and tasks is to ground information visualization in the larger process of sensemaking. Sensemaking is the cyclical process in which human collect information, examine, organize, and categorize that information, isolate dimensions of interest, and use the results to solve problems, make decisions, or communicate findings [20, 94, 95, 102]. Interactive visualization enhances the sense making cycle by aiding the search for information, facilitating the discovery of patterns, and providing means for evaluating various hypothesis. Several high-level analysis processes and tasks could be derived from the sensemaking model, such as information gathering, discovery of patterns/outliers, hypothesis testing and validation, etc.

In the field of information visualization, several taxonomies have been proposed to categorize user tasks and analysis processes. Zhou et al. introduced a visual task taxonomy that interfaces high-level presentation intents with low-level visual techniques [129]. In particular, they characterized visual tasks by presentation intents, presenting a hierarchical taxonomy of lists of visual tasks with each list accomplishing a certain intent. For example, the process of *searching* for information using an interactive visualization interface may involve the tasks of categorizing information, clustering, comparing different items, correlating, distinguishing, ranking, etc. The tasks can then be translated into specific visual encodings and interaction techniques. The taxonomy proposed by Zhou et al. is influential in the sense that it bridges the high-level analysis process with low-level visualization designs. The taxonomy also provided a preliminary categorization of presentation intents, which are elements to describe a user's analysis processes.

As an extension to Zhou's work, Amar et al. introduced another taxonomy which further bridges interaction and high-level analytic activity [10]. More specifically, the taxonomy is comprised of ten low-level analysis tasks that largely capture people's activities while employing information visualization tools for understanding data. Drawing from large number of answers regarding different data sets from different domains, Amar et al. deducted ten tasks shown in figure 1. If considering the set of tasks in the taxonomy as analytic “primitives”, then a combination of some of the tasks support examination of other high-level questions that do not clearly fit into one category. The combinations of the analytic “primitives” also build versatile high-level analytical processes.

3.2 Know Thy User & Know Thy Tasks

In order to develop an understanding of domain-specific analysis processes, a visualization researcher must learn how to interact with domain users. Several methodologies of interacting with potential users have been widely studied in the HCI community. The most important step during the process of design is to identify target users’ needs, through understanding as much as possible about the users, their work, and the context of their work [104]. In other words, it is essential to know the users and their tasks before producing any kind of design.

Why is identifying needs important? Much has been written about the significant cost of fixing errors late in the software development cycle rather than early, during the requirements activity. For example, Davis identifies insufficient user communication, poor specification, and insufficient analysis as contributors to poor cost estimation [31]. In addition, domain users won’t adopt the final product if it is not tailored to their needs and workflow.

	interface knowledge	domain/task knowledge
novice	little to none; shallow	little to none; shallow
first-time	little to none; shallow	knowledgeable
knowledgeable intermittent	some, but not specific	knowledgeable
expert frequent	expert	expert

Figure 2: User proficiency profiles [27].

3.2.1 Know Thy User

One of the most essential steps in the design process is to understand who the potential users are. It is important because researchers can't deliver a suitable system without knowing those problems they are trying to solve. As Hansen [48] pointed out, in order to know the user, the system designer should try to build a profile of the intended user: his education, experience, physical attributes, perceptual abilities, cognitive abilities, personality and social traits, ect. User characteristics capture the key attribute of the intended user group. Two types of user knowledge are commonly collected, including information regarding users' skills with the interface and the domain. A user may be a novice, an expert, a casual, or a frequent user. This affects the ways in which visual encodings and interactions are designed. Figure 2 illustrates user proficiency profiles based on user knowledge [34]. Given the user profiles, Hansen further presented ways to accommodate multiple user profiles such as providing multi-layer, level-structured interfaces. So that novices get limited options which allows less opportunities for error, while increased proficiency enables increased functionality.

3.2.2 Identify the Tasks

In addition to understanding the target users, designers also need to know the domain. It is a process of analyzing and documenting how people perform their jobs or activities. Understanding the domain involves task decomposition, the goal of which is to find out what tasks are "atomic" and how are composite tasks put together. Understanding the domain also involves knowing the task frequency, which in turn determines navigation structure and invocation methods. Designers can focus on analyzing activities, artifacts, and relationships between artifacts and activities. Multiple methods have been developed in the HCI community to describe activities, such as developing scenarios and user cases, flow charts and workflows, as well as entity-relationship diagrams or object models.

In order to know the user and the tasks, there are multiple ways to interact with target users and gather relevant data [35, 104].

- Observation and thinking out loud. Observation of participants in their natural setting is used to understand the nature of the tasks and the context in which they are performed. This method allows designers to watch users doing activities of interest to the designers and encourage users to verbalize what they are thinking. Sometimes the observation is carried out by trained observers who record their findings and report them back to the design team, and sometimes the observation is carried out by or with a member from the design team.

- Participative evaluation. The designer can also sit and talk with users as they do their activities that of interest to the designer. The method can be considered as a relaxed version of thinking out loud in that the observer and participant can ask each other questions so a more mutual understanding could be developed throughout the course.

- Interviews. Interviews are good at getting stakeholders involved at an early design stage. Interviews can be divided into structure, unstructured, and semi-structured forms. The main difference lies in how much data of interest is predetermined by the interviewers.

- Focus groups. Focus group can be used to get at people's desires, motivations, values and experiences in the form of a group of individuals, usually ranging from 3 to 10. A focus group is relatively low cost, and a quick way to learn a lot about potential users. Focus groups are good at gaining a consensus view and highlighting areas of conflict and disagreement. On a social level it also helps for stakeholders to meet designers and each other, and to express their views in public. It is not uncommon for one set of stakeholders to be unaware that their understanding of an issue is different from another's even though they are in the same organization.

- Questionnaires. Questionnaires may be used for getting initial responses that can then be analyzed to choose people to interview or to get a wider perspective on particular issues that have arisen elsewhere. A designer can develop a list of specific questions that they want answers from domain users. A well-designed questionnaire will usually include both closed questions with range of answers and open-ended questions.

- Ethnography. This is an deeply contextual study in which a designer immerses herself in a situation that she wants to learn about. The study usually yields informative insights since behavior is meaningful only in context. Ethnography has traditionally been used in the social sciences to uncover the social organization of activities, and hence to understand work. Since 1990s it has gained credibility in HCI design, and particularly in the design of collaborative systems. A large part of most ethnographic studies is direct observations, but interviews, questionnaires, and studying artifacts used in the activities also feature in many ethnographic studies.The drawbacks of ethnographic methods include small scale to keep users group manageable and highly qualitative results.

When designing an interactive visualization system, designers often use a combination of the above methods to interact with target users to gather data relevant to both user and tasks.

3.3 Domain-specific analytical processes in visual text analytics

Visual analytics systems are usually designed for target domains, such as biology, business intelligence, intelligent analysis, etc. Each domain usually has its own workflows and existing set of analysis processes. Therefore, in addition to the general analytical processes, it's even more important for visualization designers to clearly understand domain-specific analysis processes, so that the visual analytics systems could fit right into and further facilitate domain experts' workflow.

The field of *text analytics* involves information retrieval, pattern recognition, data mining techniques, visualization, etc. The overarching goal of text analytics is to turn text into data for analysis [41]. Under this broad field, the branch visual text analytics has been gaining interests within the visualization community. Visual text analytics combines text analytics and interactive visualization to help users explore and analyze large text corpora. The benefit of coupling interactive visualization with text analytics methods are two-fold. First, the results from the text analytics techniques are often too complex for average users to consume. Interactive visualization could further process the complex output and present users with intuitive visual representations to help users interpret and examine the results from multiple perspectives. Second, despite continuing advancement, text analytics techniques are still less than perfect. The interactive visualization tools provide users with alternatives to compensate for the deficiencies in these techniques such as providing details on demand to help users understand sometimes abstract text summarization.

To summarize domain-specific analytical tasks, we surveyed papers published in the field of visual text analytics. When analyzing email archives to portray conversational relationships, Viegas et al. summarized two main questions that they want to help target users answer:

- What sorts of things do I talk about with each of my email contacts? (self-recall and review)

- How do my email conversations with one person differ from those with other people? (comparison)

The questions emerged from their experience with a previous email visualization project, in which they discovered that users were quite fascinated by the ability to look back at overall patterns of exchange in their archived messages.

In the domain of consumer analytics, Liu et al. presented an interactive visual text analysis tool - TIARA - to aid users in analyzing a large collection of text. More specifically, four high-level questions were extracted for guiding the design of the visualization tools:

- What are the major topics in the customer feedback?
- What are the most active topics during the last few month?
- What are the key concepts mentioned in the above topics?
- How have the most active topics evolved over time?

In the world of intelligent analysis, Stasko et al. presented Jigsaw to support examination of reports based on concepts and entities [113]. Entities include people, locations, and time. By allowing users to see the relationships of the entities visually, Jigsaw support investigative analysis by answering questions regarding who, when, and where. The intelligence analysts can then synthesize such information to develop a theory regarding threat plot and potential actions.

In addition to summarizing a text corpus and analyzing relationships between entities that are mentioned in the corpus, visualization researchers have also presented tools for sentiment analysis of news articles [122], customer reviews [87], and social media streams [49]. These visualization tools could help users assess how positive or negative a particular posting is, through an intuitive, visual way. As online news and social media data has been growing in an exponential rate, visual sentiment analysis tools could provide an overview of opinions at a glance.

When examining documents at a per-sentence level. Oelke et al. have presented a method to analyze the expressiveness of the language in document corpora [88]. The authors first put together 141 different text features to determine readability given a document. Through a rigorous process of filtering, the final set is comprised of 5 non-redundant and semantically meaningful features. The authors then built a tool, VisRA, for visual readability analysis. VisRA can help users visually explore the expressiveness at per-sentence level within a document.

From the summarized analysis tasks in the field of visual text analysis based on previous work, we developed a good understanding of what has already been accomplished and how other visualization researchers have transformed the tasks into visual representation and user interactions. Through studying related research, we also learned about the pros and cons of several NLP methods that can be applied to summarize and analyze large document corpora, which enables us to choose the appropriate method for our own design.

CHAPTER 4: PARALLELTOPICS

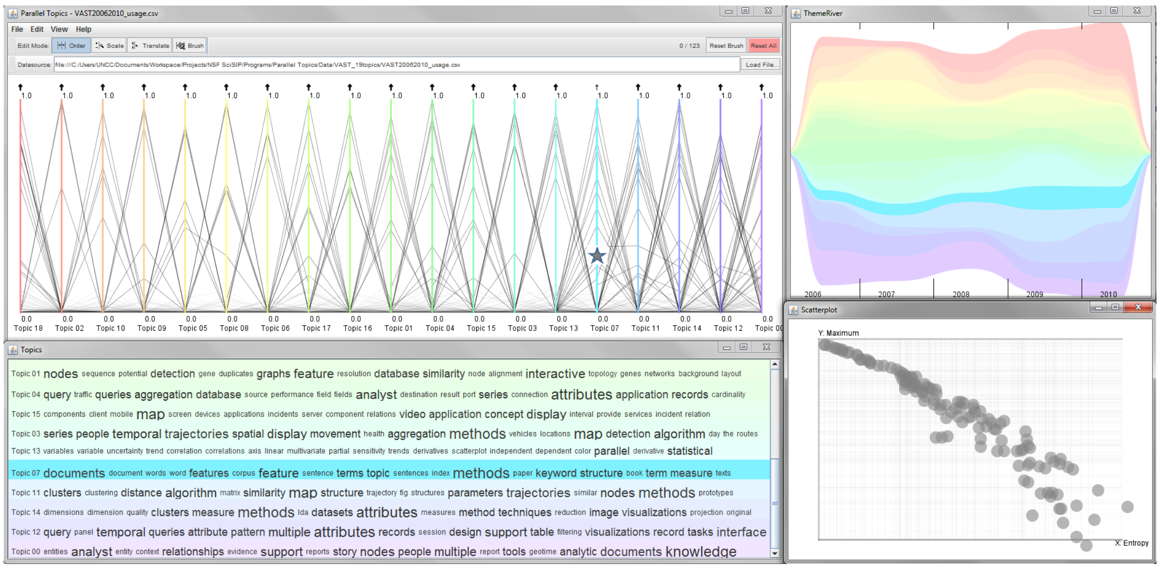


Figure 3: Overview of ParallelTopics.Topic07 is highlighted. Top left: Document Distribution view, top right: Temporal View, bottom left: Topic Cloud, bottom right: Document Scatterplot. A user is hovering the mouse over Topic 07 (light blue) in the Document Distribution view.

In this chapter, we introduce ParallelTopics, which is an interactive visual analytics system designed to help users make sense of large text corpora. To design ParallelTopics, we first performed series of interviews and focus groups with target users to understand their data and domain analysis tasks. We then categorized our findings into four tasks that are rather general in the domain of text analysis. Our effort contributes additional analysis tasks to the visual text analytics area. More specifically, we propose the tasks of allowing target users to explore relationships between documents and topics and we extend previous work to support exploration of emerging temporal patterns. Our proposal of marrying interactive visualization with state-of-the-art topic models is among the first within the VAST community.

4.1 Domain Characterization

A core part of NSFs mission is to keep the United States at the leading edge of discovery, both by funding research in traditional academic areas, including identifying broader impacts, as well as funding transformative and interdisciplinary research. In order to do the former, the program managers at NSF need to identify appropriate reviewers and panelists to ensure the best possible peer review. In order to effectively perform the latter, the program managers need to identify emerging areas and research topics for funding interdisciplinary and transformative research. In addition to making funding decisions, program managers also need to manage their award portfolios. While the program managers have done a great job in the past, they are in need of new methods to help them due to the rapidly changing nature of science, and the significant increase in the number of proposal submitted.

Through workshops and interviews, we have collected information regarding data, tasks, and workflow of the program managers. The first common task involves identifying appropriate reviewers for newly submitted proposals, they first need to divide these submissions into groups based on their topics. This task requires an understanding of the major topics covered by the submissions, and the task also involves dividing the submissions based on their perspective topics. Another task is appropriate reviewers based on their expertise. Knowing research background of potential reviewers will help to determine which area their reviews can contribute to. The third task the program managers often perform is to examine award portfolios, especially from a temporal aspect. But currently the program managers are only able to look at the temporal trend based on numerical data, such as the number of proposal submission/acceptance over the years. During the interviews, they mentioned that they would like to follow how ideas in the research have evolved over the years.

4.2 System design based on gathered analysis tasks

We categorized the domain analysis tasks into four general questions:

- Q1: What are the major topics that capture the document collection well?
- Q2: What are the characteristics of the documents based on their topical distribution?
- Q3: What documents address multiple topics at once?
- Q4: How do the topics of interest evolve over time?

Answering a combination of these questions could help program managers solve their routine tasks. For example, task 1 focuses on dividing new proposal submissions into groups based on their topics. This task requires understanding the major topics of the text corpus (Q1), and filtering sub document collections based on their characteristics over topics (Q2). Task 2 is to identify appropriate reviewers for the proposal submissions, which further involves knowing whether a submission is related to multiple topics (Q3) in order to gather the right expertise. Last, Task 3 focuses on the temporal aspect of the award portfolios which involves discovering the topical trend over time (Q4).

To help users answer these questions, ParallelTopics (figure 3) first extracts a set of semantically meaningful topics using LDA [16]. To support visual exploration of a document collection based on the topic model, ParallelTopics employs multiple coordinated views to highlight both topical and temporal features of document corpora. The novel contribution of ParallelTopics lies in the depiction of the probabilistic distributions of documents over topics and supporting interactive identification and more detailed examination of single-topic and multi-topic documents.

More specifically, to describe a corpus of documents, ParallelTopics first extracts a set of semantically meaningful topics using LDA. Unlike most traditional clustering techniques in which a document is assigned to a specific cluster, the LDA model accounts for different topical aspects of each individual document. This permits effective full text analysis of larger documents that may contain multiple topics. To highlight this property of the model, ParallelTopics utilizes the parallel coordinate metaphor to present the probabilistic distribution of a document across topics. Such representation allows the users to discover single-topic vs. multi-topic documents and the relative importance of each topic to a document of interest. In addition, since most text corpora are inherently temporal, ParallelTopics also depicts the topic evolution over time. Since making sense of large text corpora may involve exploring all of the aforementioned aspects, ParallelTopics provides rich interactions that support the coordination among all views with each representing on aspect of the document corpora.

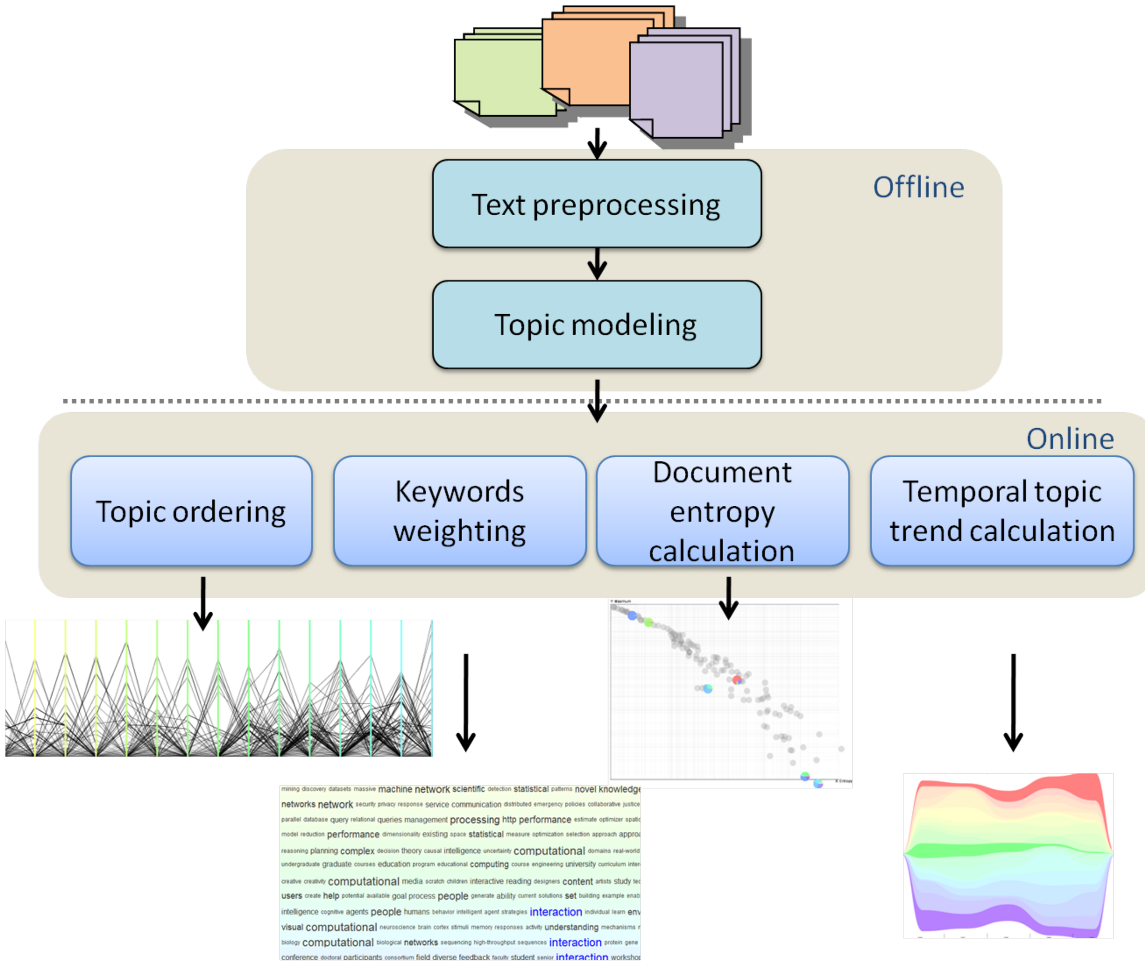


Figure 4: System architecture of ParallelTopics.

4.3 System architecture

Based on the state-of-the-art topic model, ParallelTopics employs multiple coordinated views with each view addressing one of the aforementioned questions. In this section, we describe the design of the ParallelTopics system. Figure 4 illustrates the overall architecture of ParallelTopics. Starting from the top, document processing and topic modeling are done offline. Based on the offline processing results, each online module serves one specific visualization view in the ParallelTopics. We start by introducing the topic model that underpins the ParallelTopics system.

4.3.1 Topic-based text summary

Topic models have several advantages over traditional text processing techniques. Therefore we employ a probabilistic topic model in the ParallelTopics to summarize document collections. More specifically, we used Latent Dirichet Allocation (LDA) [16] to first extract a set of semantically meaningful topics. LDA generates a set of latent topics, with each topic as a multinomial distribution over keywords, and assumes each document can be described as a probabilistic mixture of these topics [18]. To introduce the notation, we write $P(z)$ for the distribution over topics z in a particular document. We assume that the text collection consists of D documents and T topics. These notations will be used throughout the rest of chapter.

In our system, we first processed the document collection and remove stopwords such as in IN-SPIRE [5] and [46]. We then use the Stanford Topic Modeling Toolbox (TMT) [97] to extract a set of topics from the document collection. The extracted topics and probabilistic document distributions serve as input to the visualizations in ParallelTopics.

4.3.2 Interactive visual exploration of text corpora

In this section, we introduce the visual design of ParallelTopics. The system consists of four coordinated overviews: a Document Distribution view that displays the probabilistic distribution of documents across topics; a Topic Cloud that presents the content of the extracted topics; a Temporal view that highlights the temporal evolution of topics; and a Document Scatterplot that facilitates interactive selection of single-topic vs. multi-topic documents. Each of the four views in the ParallelTopics system serves a distinct purpose, and they are coordinated through a rich set of user interactions. In addition, upon selection of any documents, we provide a Detail view that presents the text content on demand.

4.3.2.1 Topic Cloud : Revealing the major topics (Q1)

To help the users quickly grasp the gist of a document collection, we present the topics as a tagcloud. In the Topic Cloud view, each line displays a topic, which consists of multiple keywords. Since each topic is modeled as a multinomial distribution over keywords, the weight of each keyword indicates its importance on the topic. To encapsulate such information in the Topic Cloud, we align the keywords from left to right with the most important keyword at the beginning. In addition, since one keyword may appear in multiple topics, the size of each keyword reflects its occurrences within all topics. An example of the Topic Cloud view is shown in figure 3 bottom left. To assist users in understanding the major topics in a document collection, we present the topics in a sequence that semantically similar topics are close by so that there is continuity when scanning over the topics sequentially. Since the LDA model does not model the relationship between topics, we reorder the topics by defining a similarity metric.

Interaction supported To design user interactions in the TagCloud view, we consult both interaction taxonomies and our domain task characterization. We enable users make sense of the topics by supporting high-level user interactions such as explore, highlight, and search. For example, hovering over a particular keyword would highlight all other occurrences in the Topic Cloud. A user may also search for a particular keyword of interest. In addition, the Topic Cloud view is tightly coordinated with all other views to promptly provide information regarding a specific topic on demand.

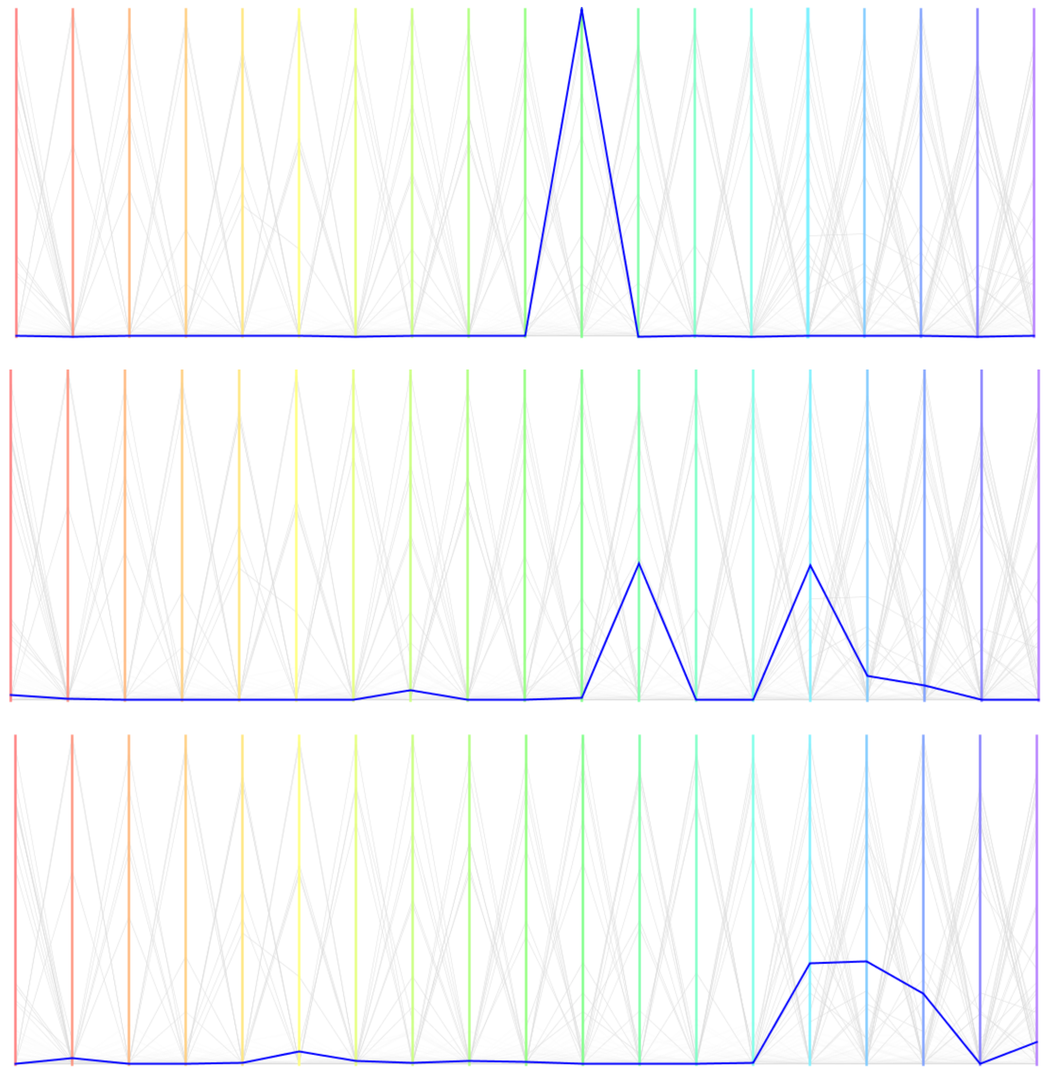


Figure 5: Document Characteristics: Top - Single Topic document; Middle - Bi-topic document; Bottom - Multi-topic document.

4.3.2.2 Document distribution: depict the characteristics of the documents (Q2)

To provide an overview of documents as mixtures of topics, we highlight the distribution of each document across all extracted topics. Our representation converts the documents which are in the form of probabilistic distributions to signal-like patterns that signify each document. More specifically, we adopt the parallel coordinate metaphor [59] with each axis denoting a topic and each line representing a document in the collection (figure 3, top left view). In our use of the parallel coordinate, all variables (topics) are uniformly spaced, and every variable shares the same value range from 0 to 1. Therefore, when viewing the document distribution view, it is not necessary to make sense of a document based on its value on each individual axis but based on the pattern across all the axes as a whole. In addition, we order the topics in a way that semantically similar topics are next to each other so that the correlation between similar topics becomes visually salient.

Interaction supported The document distribution view provides a rich set of interactions, such as brushing and highlighting, to allow users to filter and examine documents of interest. Brushing a probability range on a topic allows users to select documents that have a certain probability for that specific topic. Through synthesizing the information from both Topic View and Document Distribution View on the major topics and document characteristics, a user could effectively develop an overview of the document collection. When exploring the document distribution over topics, one can easily discover that documents present different characteristics based on the number of topics they have. Figure 5 illustrates documents that focus on only one topic, two topics, and more than two topics. Different number of topics within documents can be interpreted as distinct characteristics given a context of the text collection. For example, in a collection of scientific publications, documents with one topic denote publications on a specific research field. Documents with two or more topics are more likely to represent interdisciplinary research articles, which often integrate two or more bodies of specialized knowledge.

4.3.2.3 Document Scatterplot: Investigating documents based on their number of topics (Q3)

The document distribution view enables users to identify documents that focus on a specific topic through brushing the top range on the topic. However, identifying documents that are related to two or more topics in a large corpus is not as straightforward since they are shadowed by high probability values of the single-topic documents. To alleviate this problem, we represent all documents based on their entropy so that single-topic and multi-topic documents are easily separable. We plot each document based on its entropy and its maximum probability value over topics (normalized to [0, 1]) in a scatterplot view. In this presentation, single-topic documents (with higher max value and lower entropy) are at the top left corner within the scatterplot while bottom right corner captures documents with more number of topics (lower max value and higher entropy).

Interaction supported The scatterplot supports user interactions such as selection and filter. Upon selection, pie glyphs are shown to describe the topical contribution to a specific document. In figure 6, each pie glyph represents a selected document, with each color denoting a topic. As shown, documents with smaller entropy values appear as pie glyphs as a solid circle; whereas documents with larger entropy values appear to have multiple colors, indicating that entropy values do correspond to the number of topics in the input documents. In summary, the Document Scatterplot enables users to interactively identify subgroups of documents with desired number topics through selecting document within different regions.

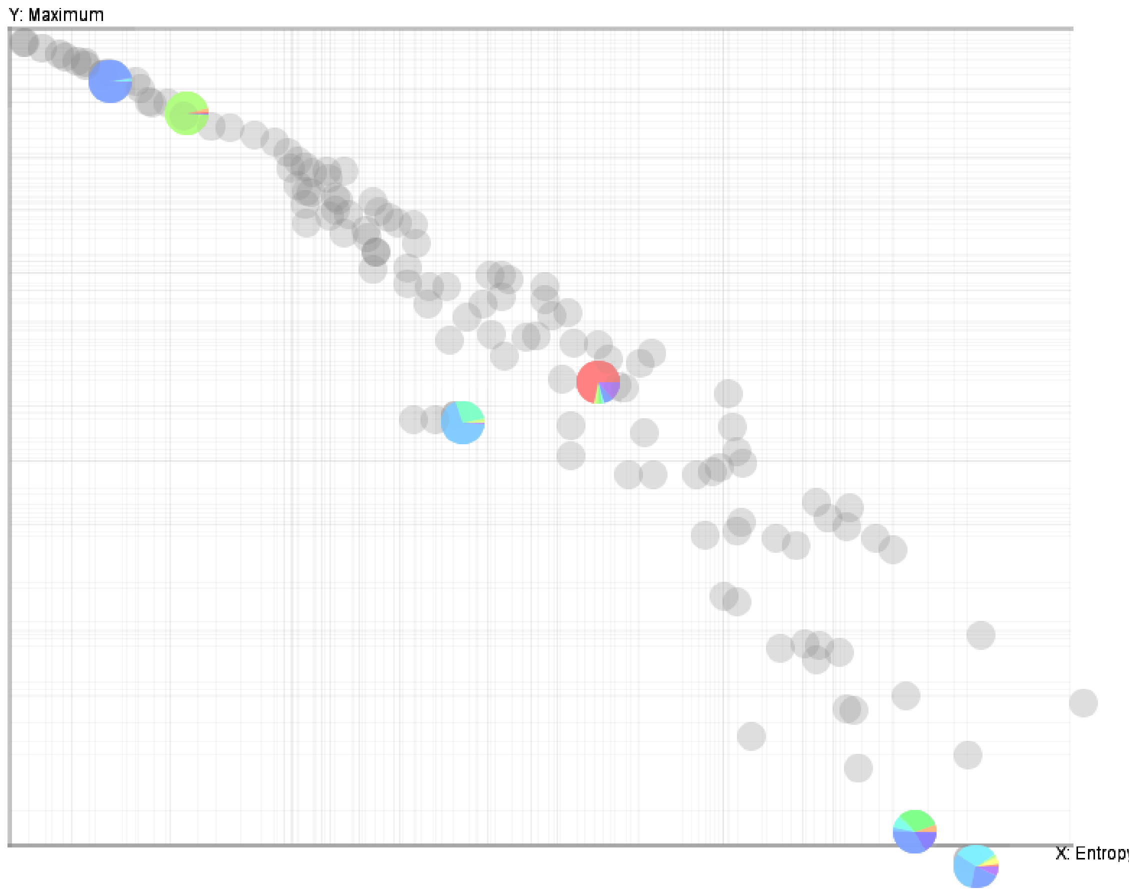


Figure 6: Document Scatterplot: the position of each document in the scatterplot correlates to its number of topics. Single-topic documents are in the top left corner while multi-topic documents reside in the bottom right corner. Each pie glyph is colored based on number of topics in each document.

4.3.2.4 Temporal View: presenting topic evolution over time (Q4)

Since most document collections are accumulated over time, it is helpful to present such temporal information to assist users in understanding how topics of a corpus evolve. Our temporal view is created as an interactive ThemeRiver [50], with each ribbon denoting a topic (figure 7). The height of the ribbon is determined by the sum of document distribution on this topic within a time period.

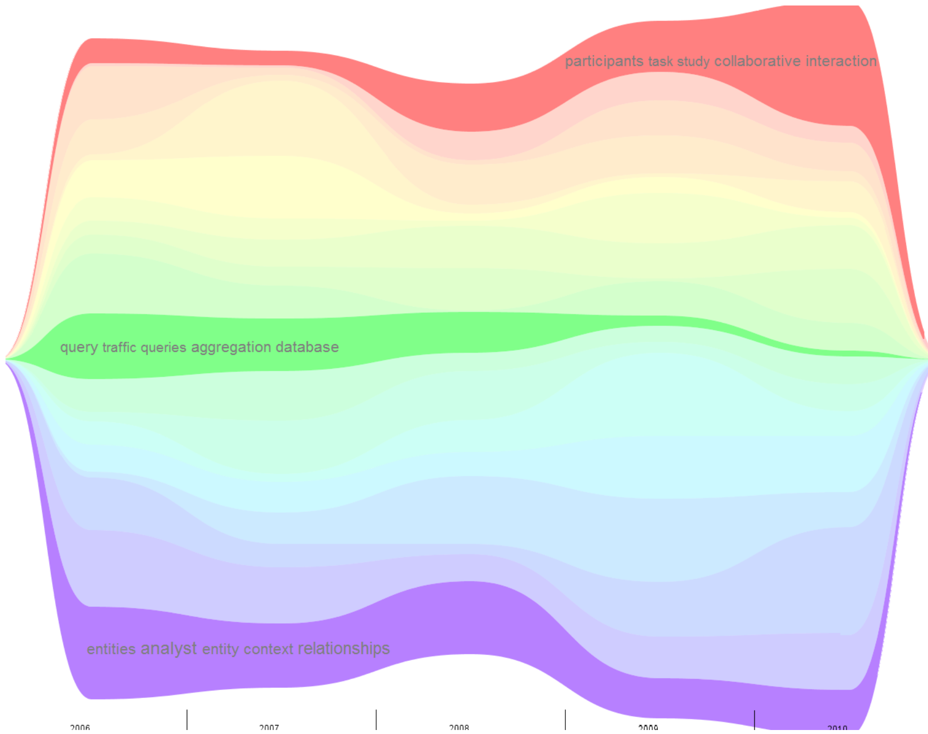


Figure 7: Temporal view with three topics highlighted. Each topic is labeled with its first five keywords. Topic in red: participants task study collaborative interaction; topic in green: query traffic queries aggregation database; topic in purple: entities analyst entity context relationships.

The order of the topics (from top to bottom) is the same as in both the Document Distribution view and the Topic Cloud. We assign the topic colors by interpolating a color spectrum using the normalized distance (Equation 1) among all adjacent topics. As a result, a more similar pair of topics is assigned with colors that are more alike. Overall, the temporal view provides a visual summary of how topics of the document collection evolve over time. The view added richness to ParallelTopics by revealing temporal information hidden in a document collection.

4.3.2.5 View coordination through user interaction

Since making sense of a large text corpus may involve the utilization of all four views, coordination among all views is carefully crafted within the ParallelTopics. On the *topic* level, hovering over a topic in any view that involves topic representation would highlight the same topic in other views. For example, if a user hovers the mouse over an axis in the Document Distribution view, the same topic would be highlighted in both Topic Cloud view and Temporal view (Figure 3). Thus the user could quickly synthesize information regarding keywords, document distribution, and temporal trend of the particular topic. In addition, the views are also coordinated by colors, with each topic being the same color in all views. On the *document* level, selection of any set of documents in the Document Scatterplot would be immediately reflected in the Document Distribution view, and vice versa. When a user selects a few documents with two prominent topics (mid-range) in the scatterplot, seeing the distributions of these documents help the user understand their topical combinations.

4.3.2.6 Details on Demand

In ParallelTopics, upon selection of any documents, we provide details of the actual text content of the documents of interest. Since any topic models are far from perfect, the function of the detail view is two-fold: first, it provides context for users to develop a deep understanding of a topic and its associated keywords. Second, the detail view helps users to validate the patterns shown in the visualization.

4.4 Case Study

To evaluate the efficacy of our system in answering the four intended questions, we applied ParallelTopics to exploring and analyzing a text corpus, namely the scientific proposals awarded by the National Science Foundation. We invited a former program manager to use the ParallelTopics system to explore the proposals to evaluate whether the system could assist her in decision-making and award portfolios management. In this section, we first describe the data we collected. We then present how ParallelTopics could assist the expert user in solving her domain tasks, which was introduced at the beginning of the chapter.

4.4.1 Data Collection and Preparation

To examine whether ParallelTopics could assist program managers in making funding decisions and managing award portfolios, we first collected the awarded proposals from 2000 to 2010 under the IIS (Information & Intelligent Systems) division, which is part of the CISE (Computer & Information Science & Engineering) directorate. The collection consists of nearly 4,000 awards, with structured data on the Award Number, Directorate, Division, Program, Program Manager, Principal Investigator, and Award Date; as well as abstract of the proposals, which is in the form of unstructured text. We processed all collected abstracts with each abstract constituting a single document in the corpus. We removed a list of standard stopwords. This gave us a vocabulary of 334,447 words. We then extracted 30 topics from the corpus using the LDA model.

4.4.2 Expert Evaluation

Since program managers at NSF are extremely busy, we invited a former NSF program manager for our expert evaluation. The participant has two years of experience working as a program manager at NSF. At the beginning of the evaluation, we spent 30 minutes demonstrating the system design and functionality of each visualization. Then we asked the participant to perform the following three tasks using the

ParallelTopics system.

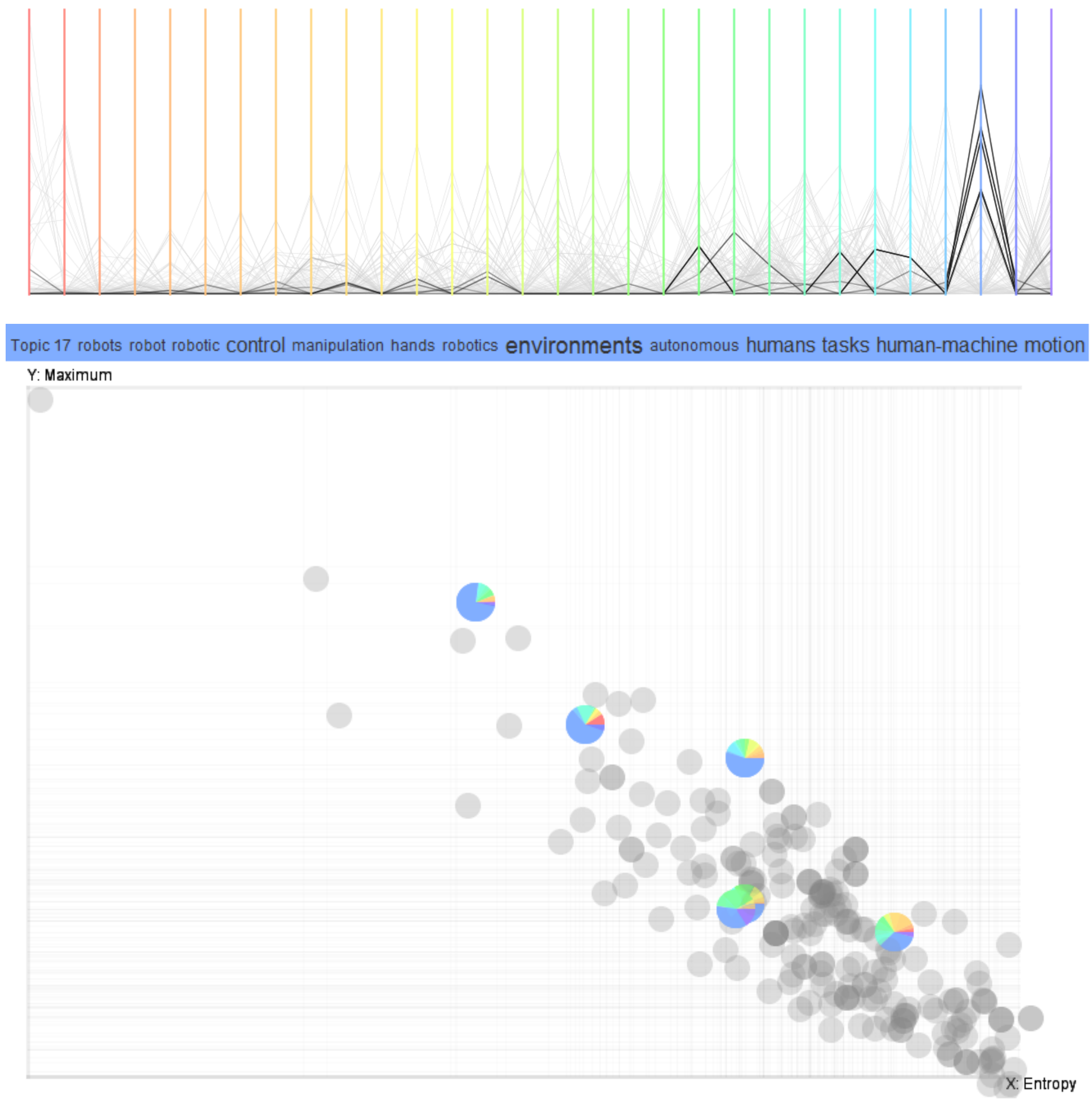


Figure 8: Exploration scenario. Top: Selection of documents on the topic "robotics". Bottom: Pie glyphs in the Document Scatterplot show the number of topics for each selected document respectively.

1. **Task1: To group 200 newly submitted proposals based on topics** : Starting with the Topic Cloud, the participant quickly scanned the extracted topics to gain an overview of the newly submitted proposals. Since the participant was responsible for proposals in the areas of robotics and computer vision,

she quickly focused her attention on these two topics. Upon selection of the proposals that focus on the topic regarding robotics (figure 8), the participant quickly glanced over the titles in the detail view to validate their relevancy. While the participant was making sure that each selected proposal is relevant, she also noticed that the positions of the proposals are scattered in the Document Scatterplot. Since the proposals in the lower right positions are more likely to contain two or more topics, the participant was interested in knowing what other topics these proposals relate to. Through further filtering the proposals that appear to be more interdisciplinary in the Document Scatterplot, the participant found that they are related to other fields such as neuroscience and social communication.

2. **Task2: To identify appropriate reviewers** : For the purpose of identifying reviewers, the participant first wanted to roughly divide the proposals into groups. Based on the initial exploration, the participant concluded that there are roughly two groups of proposals: one group that focuses on the core of robotics area, and the other that utilized a body of knowledge from other fields such as neuroscience and social communication. To identify reviewers for the two groups of proposals, the participant would like to find PIs from previously awarded proposals. Through examining the historic data, the program manager located the topic regarding robotics in the Document Distribution view. She then brushed the top range of the axis to select proposals pertinent to the topic. Finally, the participant turned to the detail view to look for PIs that

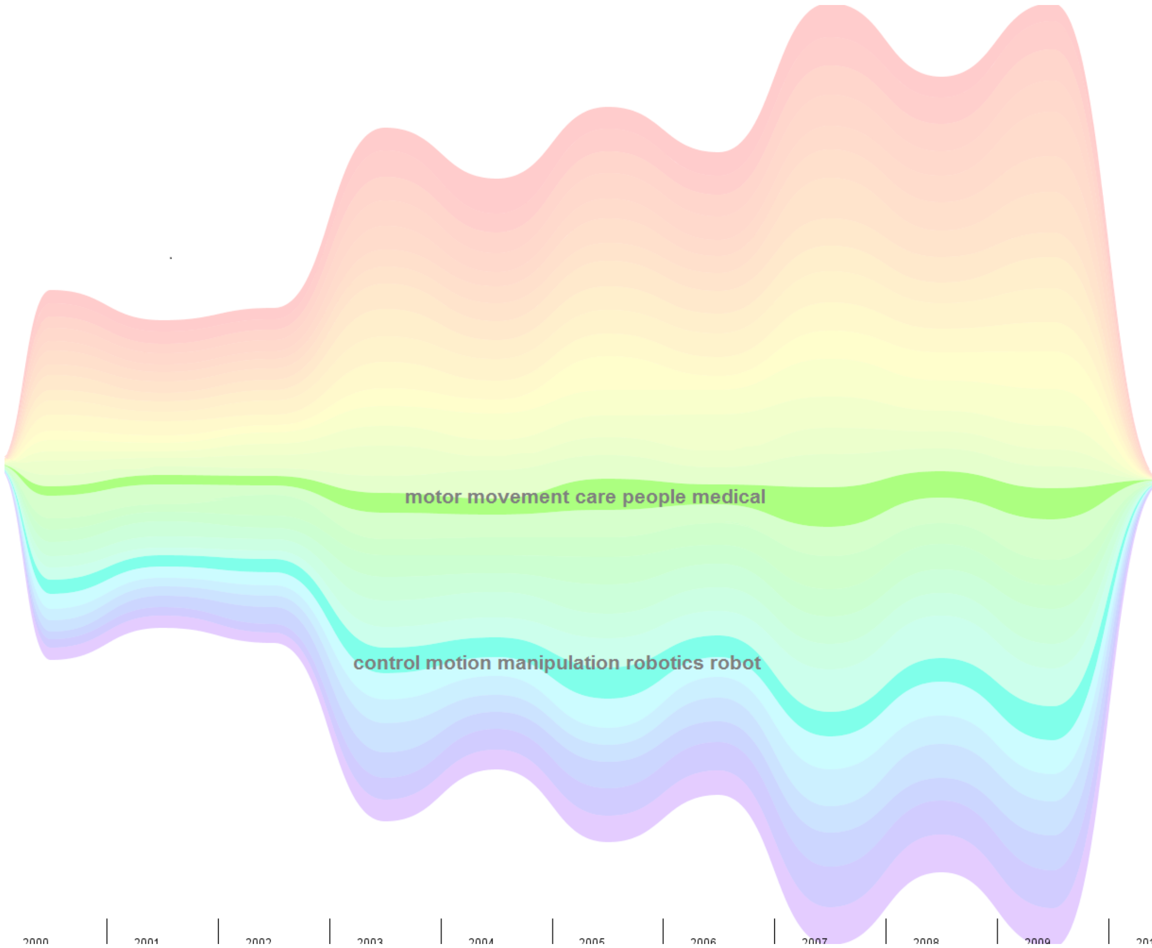


Figure 9: Exploring the temporal aspect. Although the total number of proposal grew continuously over the years, the awarded proposals regarding topic “robotics” remained steady (light blue). In contrast, more proposals related to “using interfaces to help people with impairement” were awarded over the years (green).

were previously awarded in the robotics area. For interdisciplinary proposals in group2, the participant went through similar processes to identify additional experts from other related fields (e.g. neuroscience) to serve on the review panel to ensure the best possible peer review.

3. **Analyzing temporal trend of award portfolio**: On a portfolio level, the former program manager was interested in seeing the temporal trend of the areas she is in charge with over years. Through exploring the Temporal View,

the participant discovered that the trend of awarded proposals in the field of robotics is steady, although the overall number of proposal awarded grew during year 2006 and 2009. Unlike the steady trend of robotics, the number of awarded proposals on the topic of using technology to help people with disability grew over the years (Figure 9). The former program manager commented that this view is valuable to her since it enabled her to see funding trends regarding different topics that are otherwise hard to discover.

4.5 Summary

In this chapter, we first presented a list of domain-specific analysis questions gathered from target users regarding analyzing large collections of research proposals. Our list of questions contributes to the domain analysis of text analytics. Based on the gathered analysis tasks, we present a novel visual analytics system, ParallelTopics, to enable target users to interactively explore large text corpora. To enable users to grab the gist of the text corpora, the ParallelTopics system utilizes a probabilistic topic model to summarize the textual information. However, the state-of-the-art probabilistic model is only a start, the visualization components in ParallelTopics are tailored to the analysis processes of examining large collections of textual information. Therefore additional processing steps were taken to help users answer questions regarding temporal topical trend, topic similarity, and relationship between topics and documents. Because the information gathered regarding the analysis process was properly transformed into the design of the interface, the ParallelTopics is highly valued by our expert users.

CHAPTER 5: ARCHITECTURE FOR LARGE-SCALE TEXT ANALYTICS

Domain Challenges: During expert evaluations, ParallelTopics is considered helpful in supporting analysis of collections of texts. However, in retrospective analysis of our work with ParallelTopics, we found that one of the biggest issue preventing the system to be useful for a broader audience is its scalability. As the amount of textual information available keeps increasing over the years, without a more scalable approach to deal with large text corpora, we cannot provide a good solution to domain tasks in text analytics. Since the amount of information exceeds the current processing power of a single computer, without parallelized computation, we can't summarize a large text corpus into meaningful topics, let alone support further exploration based on the topical results.

Therefore, we present a general visual analytics architecture to effectively analyze unstructured data on a large scale. Pipelined based on a high-performance cluster configuration, MPI processing, and interactive visual analytics interfaces, our architecture, I-Si, closely integrates data-driven analytical methods and user-centered visual analytics. It creates a coherent analysis environment for summarizing large text corpora, identifying temporal patterns on a topical level, and key indicators of emerging events. Such an environment can support monitoring, analyzing the latent information extracted from the text corpora. We have currently applied the I-Si architecture to collect data from social media, analyze the data on a large scale and uncover the latent social phenomena. To demonstrate the efficacy and applicability of I-Si, we describe several use cases in multiple domains that were evaluated by experts. The use cases demonstrate that I-Si can benefit a range of users by constructing meaningful event structures and identifying precursors to critical events within a rich, evolving set of topics. To showcase that I-Si architecture work on a wide range of text collections other then scientific proposals or publications, we apply the architecture to more fragmented textual information such as data collected from social media.

5.1 The "Big Data" Problem

We are moving toward a ubiquitous social era, in which mobile communications, social technologies and sensor-based services connect people, the Internet and the society into one immensely interconnected community. With the rapid growth of such ubiquitous communication infrastructures, we are living in a world where nearly everyone is connected in real time. Our society as a whole is being greatly influenced by such intimate connections, affecting every aspect of people's social behaviors. Moreover, the evolution towards such interconnected, real-time social discourse has changed the way people organize and respond to social events (e.g., happenings, protests or campaigns), enabling people to form, share, discuss, and react to social activities instantaneously.

As a result of all this personalized, digital communication, massive amount of data, including both textual and multimedia data, are collected in real-time regarding who we are, where we are, and what we are talking about. Particularly, the emergence of microblogging has yielded an overwhelming amount of such data, ranging from status updates on Twitter and Facebook, to extended comments on Google+, all often accompanied by images and more and more by video. As one example of the explosive growth, Twitter rose from about 6 million visitors per month in January 2009 to over 37 million per month as of November 2011[110]. Based on multiple estimates, on an average day, users globally submit 140 million “tweets” on Twitter; and for each month, users share about 30 billion pieces of content on Facebook.

These massive, large scale social media datasets bear extremely rich information that, if revealed, can lead to a profound impact on depicting patterns for emerging social events (and their underlying topics). This can contribute in new, important ways to the understanding of social phenomena. The need to assess the related social phenomena in a systematic way has increased for both citizens and government (e.g. emergency responders and law enforcement). Analyzing this rich social media data gives them the ability to understand and even predict people's interests, and to further depict the shifts and turns of social activity at the individual, group and global level. For example, analysis of social media could give government valuable information on how to effectively transmit problematic situations (natural disasters or chaotic scenes). Citizens or citizen groups could learn about the development, history, and spread of ideas of social movements (e.g., Occupy Wall Street).

5.1.1 Major Challenges and Opportunities

Despite continual efforts, analyzing such large-scale, loosely structured, and less-contextual social media data to support analytical reasoning remains extremely challenging. There is a scarcity of methods to extract the latent semantic information in the massive text corpora. Specifically, the challenges are two-fold:

5.1.1.1 Motivating Challenges I: Depicting Latent Social Activities

At an individual level, as streams of diverse information constantly arrive to users, it is difficult for them to keep up with, let alone harvest important and interesting messages. In addition, a user might want to identify useful content outside of her selected focus, or to discover trendy topics that other people have been discussing on social media. This task involves not only a meaningful summary of vast information streams, but also the support for interactive exploration of content according to individual interests.

At an organizational level, analyzing social media data streams allows institutions to grasp up-to-date topical trends and identify critical events that may require appropriate action. For instance, a commercial organization might be interested in reviewing consumer responses to products or the company's general image. Analyzing relevant information from social media may be a better way to gather honest opinions from possible customers than conducting targeted surveys on a sample population. Similarly, campaign strategists might be interested in knowing people's general opinions towards different parties and politicians. Social media is a perfect place for collecting such data since large groups of users voice a rich set of attitudes over time and respond to events through Facebook or Twitter.

In addition to the above examples, valuable information extracted from the noisy social media data could inform other entities such as emergency responders and police departments about future events that are being organized or current events as they unfold.

5.1.1.2 Motivating Challenges II: Establish Meaningful Social Event Structures

Nowadays people don't need to be powerful to launch a successful media campaign thanks to social media. A 27-year old art gallery owner started a national movement "Bank Transfer Day" against big banks though one Facebook post [73]. The movement led more than 1 million customers (estimated) to transfer their cash out of big banks to credit unions. An ongoing national campaign "Occupy Wall Street" used social media abundantly to spread nationwide [8].

People are amazed by the scale of such movements, but little is known regarding how they were initiated and organized. Through analyzing information related to the occupy movement in social media, for example, one can construct a timeline or even an event structure to investigate the progression of the movement and answer questions such as who were the initial organizers, who joined the campaign at what time, when exactly did the movement start, what ideas and issues developed, and which events might have led to this massive national campaign.

5.1.2 Introducing I-Si Architecture

In response to these challenges, we have developed a general visual analytics architecture to support topical-level investigative analysis of social media data. Our architecture, I-Si, is centered on the combination of data-driven topic modeling approaches with human-centered visual analytics techniques; topic modeling is enhanced by interactive visual interfaces, providing results that can be explored, filtered, and

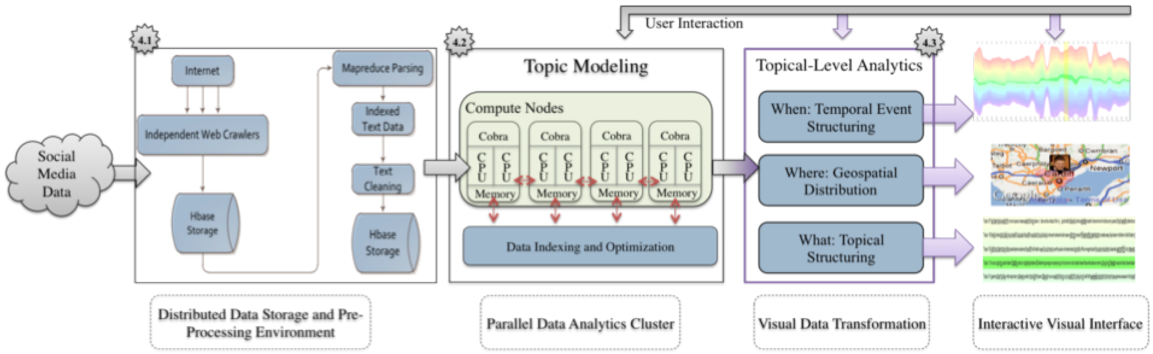


Figure 10: An overview of I-Si architecture. There are 4 major components in the I-Si architecture: Distributed Data Storage and Pre-Processing (Section 5.3.1), Parallel Data Analytics Cluster (Section 5.3.2) ; Visual Data Transformation; Interactive Visual Interface (Section 5.3.3).

managed by users. I-Si creates a coherent analysis environment for identifying event structures, geographical distributions, and key indicators of emerging events. In other words, I-Si can help analysts identify and follow social phenomena as they emerge, evolve, and mature.

On the high-level, as shown in Figure 10, the data analytics capability of I-Si comes from leveraging High-Performance clusters to apply automated topic modeling to social media data such as large collections of tweets. The visual analytics component of I-Si shares similarity to previously developed topic-based text methods [36, 38, 124], creating an investigative visual analytics environment [120] that not only provide a summary of "what happened" in terms of meaningful topics, but also allow inferences of causal relationships between a critical event and the effect of the event. In particular, our architecture could represent the progression of social events through underlying latent common themes over time, which allows users to discover the overall trend as well as the rise and fall of individual social activities.

We have currently applied the I-Si architecture to collections of unstructured social media data (e.g., Twitter data), analyzing the data on a large scale and uncovering the latent social phenomena (who-when-where-what-why). The results bring forth meaningful semantic information that is otherwise hidden in the large aggregation of noisy tweets. The results contain topics summarized based on the tweets; dynamic patterns of topics, and emerging events.

We have reached out to multiple user communities, as detailed in Section 5.5. During this outreach, I-Si has been demonstrated to and evaluated by multiple users, including political campaign strategists and law enforcement people. Feedback from these experts suggested that our I-Si architecture contributes to the social media analysis in the follows aspects:

- Analyzing social media data on event/topical level instead of keyword level. The cohesive themes from the otherwise noisy social media data are nicely summarized and presented to users. The purpose is to analyze and ultimately predict social activities/behaviors.

- Ability to handle large amounts of data. Studies have seldom focused on analyzing social media data on a large scale. We have utilized parallel computing methods to handle billions of microblog messages at once.

- Straightforward identification of critical events and even precursors to the events via temporal and geospatial visualization. In addition, through providing interactive exploration capabilities, the I-Si architecture enables users to perform investigative analysis regarding certain events/topics and answer who-what-where-when-why questions.

5.2 Background

There is a wide range of research on social media analysis, especially on Twitter data because of the public nature of tweets.

5.2.1 Analysis of space and time in Social Media

A large portion of the published research about Twitter has focused on questions related to Twitter's spatial and temporal properties with little or no semantic analysis on the textual content of tweets. For example, Java et al. [62] studied the topological and geographical properties of Twitter through constructing a social network based on users and their "friendship" information without considering the content of tweets. More recently, MacEachren et al. [80] has developed SensePlace2 - a geovisual analytics system that supports situational awareness for crisis events using Twitter data. SensePlace2 focused on extracting explicit and implicit geographic information for tweets, and combining geospatial with temporal information to promote understanding of situations evolving in space and time.

5.2.2 Topical Analysis of Textual Content in Social Media

Other work has presented analysis of the textual content of tweets using probabilistic topic models. Ramage et al. maps the content of the Twitter feed into dimensions using Labeled LDA [96], with the four dimensions corresponding roughly to substance, style, status, and social characteristics of posts (4S). One limitation of the work comes from the authors manually assigning the predetermined labels (4S) to learned topics without leaving room for users to explore and attach other meanings to the topics. Ritter et al. have applied LDA and other unsupervised approaches for the purpose of modeling conversations within Twitter streams, as the sequential dialogue reflects the shape of communication in the online platform. More recently, Sizov proposed a framework, GeoFolk [111], which combines textual content with spatial knowledge (e.g. geotags) to construct better algorithms for content management, retrieval, and sharing.

5.2.2.1 The Use of Topic Models

Lots of aforementioned content analysis was performed using Latent Dirichlet Allocation (LDA) or extended version of LDA, which is introduced by Blei et al. in 2003 [16]. The aim of LDA is to discover the hidden thematic structure in large archives of documents. Since the debut of the first topic model, a number of variations have been developed to extend the capabilities of LDA. One distinct advantage of topic models over previous vector space models such as tf-idf and latent semantic analysis is that each topic is individually interpretable, providing a probability distribution over words that pick out a coherent cluster of correlated terms [15]. The LDA model postulates a latent structure consisting of a set of topics; each document is produced by choosing distribution over topics; and then generating each word at random from a topic chosen by this distribution. The extracted topics capture meaningful structure in the otherwise unstructured data.

5.2.3 Visual Analysis of Social Media Data

Most of the aforementioned [62, 96, 111] work only focuses on data-driven techniques with a limited scale. Therefore, their objective and approach differs from our approach of combining both topic-level analysis and human-centered visual analytics methods. In the realm of interactive visualization, aside from SenseSpace2 [80], researchers have presented systems to track on-going social events and to support the use of social media as supplemental information sources for journalists [36, 33].

Dork et al. introduced Topic Streams, a web-based inter-active visualization system to follow and explore conversations on Twitter about large-scale events [36]. Topic Streams provides coordinated views which support visualizing topics over time, participants' activities, and popularity of event photos. The authors also provides several design goals such as summarizing the conversation, providing flexible time windows, etc, which are quite informative for future design. In addition, Diakopoulos et al. [33] presented a visual analytics tool, Vox Civitas, to help journalists extract news value from social media content around broadcast events such as televised debates and speeches. The visualization component of the I-Si architecture differs from Topic Streams and Vox Civitas in defining topics. Dork et al. treated each of the most frequent words as a topic in their visualization while Diakopoulos extracted keywords from each time window; we extract topics using LDA to pick out stronger and more cohesive themes from the entire social media corpus.

5.3 I-Si: Scalable Architecture for Topical Analysis of Social Media Data

In this section, we present our architecture and its implementation. Pipelined based on Hadoop servers, a high-performance cluster configuration, MPI processing, and visual analytics interfaces, our architecture closely integrates data-driven analytical methods and user-centered visual analytics. It creates a coherent analysis environment for identifying event structures, geographical distributions, and key indicators of emerging events. The core components of our architecture include Data Collection, Data Cleaning, Topic Modeling, and finally Interactive Visual Analytics Interfaces. As shown in the overview pipeline (Figure 10), the benefit of our componentized modules is that the structure can incorporate more efficient and advanced analysis components to enrich the analytic capability of the architecture.

5.3.1 Distributed Data Storage and Pre-Processing Environment

```
Input: Text data from HBase storage
Output: Cleaned data, frequencies of each distinct word in every
document with its associated document ID
Map stage:
  Create stop words hash table;
  Read input from HBase;
  Tokenize each line;
  Remove stop words;
  Output <docID+word, 1>;
Reduce stage:
  Input <docID+word, [1, 1, …]>;
  Count the number of elements in the value of input value array
  [1,1,…] ;
  Output <docID, [word1+freq, word2+freq …]>;
```

Figure 11: Map-Reduce Process for Data Cleaning

As shown in Figure 10, our architecture aims at incorporating multiple sources of social media data such as Twitter updates, editorial news, and blogs. The heterogeneous and streaming nature of these data sources poses a significant challenge in data management schema and data cleaning.

Given the scale of data that our architecture focuses on, standard SQL data crawler and management schema are not optimized to handle the I/O of different kinds of social media data. Specifically, considering the intrinsically fragmented and loosely structured nature of tweets, our architecture requires a powerful distributed database processing approach to achieve sufficient data access and effective data processing. After experimenting with several nuance NoSQL structures (e.g. Cassandra [1], MongoDB [4]), we adopted the MapReduce framework [32]. MapReduce is a framework for processing large datasets in a highly distributed fashion using a large number of computers, collectively referred to as a cluster (if all nodes use the same hardware) or a grid (if the nodes use different hardware). We have adopted a recent open source implementation of MapReduce, which is Hadoop, to establish our data analysis infrastructure. Hadoop parallelizes data processing across many nodes (computers) in a compute cluster, speeding up computations and hiding I/O latency through increased concurrency. Hadoop is especially well-suited to large data processing tasks (like searching and indexing) because it can leverage its distributed file system to cheaply and reliably replicate chunks of data to nodes in the cluster, making data available locally on the machine that is processing it. We utilizes Hadoop's native implementation in Java, and extended using our own shell and python script to conform to a data process pipeline. HBase, an open-source realization built on the Hadoop File System (HDFS), is used in this platform to store the collected social media data.

Besides providing a stable data management platform, we also utilized Hadoop to achieve a robust parallel data crawling and cleaning system. As shown in Figure 10, the crawling system is interfaced with the Internet through multiple independent crawlers. Each of the crawlers constantly collects social media data from various public domains and dumps it into HBase. Specifically, our crawler taps into Twitter's public API to collect tweets. It acquires such information on the "Garden-hose" level, which constantly delivers 10% of Twitter messages with a "statistically significant sample" of all tweets [6].

Concurrently, such textual data is being cleaned, parsed, and prepared for topic modeling through multiple Mapreduce jobs that perform these analytics tasks in the back-ground. During these tasks, noise symbols and stopwords are removed. Basic statistical analysis, such as word count, is also performed over the data, preparing for the topic modeling procedure.

The implementation of both the data cleaning and basic statistical tasks contain two stages (i.e. map and reduce stages), detailed in figure 11. In the map stage, data is distributed into working nodes for intermediate computation; the output of the map stage follows the $< key, value >$ pair format. After merging all the values with identical keys into an array, the merged intermediate results are collected for further computation in reduce(s). The final output of the reduce(s) has the same $< key, value >$ pair format as map's. Extracted data is then stored into HDFS data repository and is distributed across multiple nodes within our Hadoop cluster to guarantee reliability.

5.3.2 Parallel Topic Modeling using High-Performance Computing Cluster

In order to have a comprehensive understanding about the latent social media data, one needs to extract and correlate information from massive amounts of data. With new tweets reaching a billion every five days, performing such analysis is beyond the scope of computing power of any single-node configuration, either it would be impossible to process the data (i.e. memory issues) or it would take too long to obtain analysis results. This suggests yet another significant scalability challenge in social media analysis.

To alleviate such scalability issues, our architecture incorporates the use of a high-performance cluster to strengthen the analytical capability over the social media data. In particular, once the data has been cleaned and stored in HDFS, it is then ready to be processed by parallel computing clusters for topical-level analysis. Such a process has its most notable performance bottleneck at the learning and inference stages [84]. In order to reduce the time to complete this stage, we extend on Google's PLDA MPI implementation [3]. The algorithm is a general implementation of LDA with parallelization implemented into key portions of the algorithm. Such process utilizes Gibbs Sampling, a Monte Carlo approach, to compute the result towards a conversion point as the number of iterations increase.

During this process, we use Portable Batch System (PBS) to schedule the jobs and the Message Passing Interface (MPI) is used to make parallel use of the cluster nodes. As shown in Figure 10., each node of our cluster has 12 cores and a total of 36GB of memory, with a fast Gigabit Ethernet to communicate results. Our cluster converts the input data (e.g. tweets) into output data (topic-based probabilistic information), using LDA to create a probabilistic model that uses the documents, the words, and their utterances to build the topic model.

The benefit of using such infrastructure is two-fold:

- We can now process the social media on a scale that a single node computer wouldn't be able to handle. Such infrastructure and its parallelized algorithm granted us to capability to peek into the topics that are embedded in the large unstructured text corpus. For example, we have tested the I-Si architecture to investigate topics from 17,651,186 tweets, roughly around 1.8Gb data over the course of 5 weeks.

- This setup reduces the processing time for data between 150mb to 300mb, providing our architecture the iterative capability to searching most interpretable topic modeling results. This would be important for certain critical response situations such as presented in Scenario I and III (see section 5.1 and 5.3).

5.3.3 Visual Data Transformation and Interactive Visual Interfaces

To support the analysis needs from different user communities, we designed a coordinated multiple-view interface to create an interactive visual analytics environment. Each view is designed via transforming the output from topic models to showcase one distinct aspect of the underlying social media data.

As shown in Figure 10, the I-Si interface is designed to support understanding of spatial and temporal patterns of social activities, identification of event structures, and topical trends through analysis of the growing social media database. A key goal for these interfaces is to permit users to interactively explore, characterize, and compare the space-time aspects associated with topics in tweets. The default interface includes a topic cloud view, temporal and geospatial views, and detailed text view. Tight coupling between these views via interactive techniques permits this interface to effectively visualize highly dynamic and fragmented social media data. The four primary display views are dynamically coordinated. Each view is introduced be-low and their coordination is further discussed.

1. **Topic Cloud: revealing major topics**: We present the topics as a tagcloud for quick overview/summary of the social media corpus. In the topic cloud, each line displays a topic, which consists of multiple keywords. The order of the keywords within a topic indicates their importance to the topic. In addition, since one keyword may appear in multiple topics, the size of each keyword reflects its number of occurrences within all topics.

2. **Temporal View: presenting topic evolution**: The temporal view is created as an interactive ThemeRiver [50], with each ribbon representing a topic. The length of the time frame in the themeriver can be changed by users based on their investigative needs. After the time frame has been chosen, the tweets are divided into corresponding time units based on their time stamps; then the height of each ribbon is calculated by summing the number of tweets in each time unit.

3. **Geospatial View: displaying geographical distributions**: We utilize Google Map [2] to provide users with interactive geospatial analysis (see Figure 10). By placing the tweets with geo-tagging (i.e. GPS location associated with tweets) onto the scalable map, detailed geographic relationships and patterns immediately become apparent. In addition, we extended Google Map to effectively display the topical distributions of the social media data. The geospatial view incorporates a client-side clustering algorithm to overlay large amounts of geo-coordinated tweets over the map, creating a density Heatmap [112] to show the tweet clusters.

4. **View coordination and interactions**: Since investigative analysis of social activities may involve the utilization of all views, coordination among the views is supported. On the topic level, hovering over a ribbon in the temporal view would highlight the corresponding topic in the topic cloud so that users could quickly synthesize information regarding topic content and temporal trend. On the temporal level, filtering tweets that were posted within a certain time period is supported. A user could further filter tweets by geospatial area and topics. For instance, clicking on an intersection of a topic ribbon and a time frame in the temporal view would lead to the selection of tweets that are highly related to the topic and posted during the time period. These selections support detailed examination of topical trends and events.

5.4 Case Study

To demonstrate the efficacy and applications of I-Si, we describe three scenarios based on analysis of social media data. In these scenarios, the I-Si architecture supported summarizing a large amount of social media information and interactive exploration of the generated topical trends. More specifically, the scenarios demonstrated that the interactive analysis could distill meaningful and otherwise hidden information from noisy social media data, such as revealing critical events and pre-cursors of such events.

5.4.1 Scenario I: Depict Meaningful Event Structures

The Occupy movement is an ongoing series of demonstrations and is known for using social media to attract more protesters. The Occupy movement is long-lasting and widely spread; people in almost every major city within the U.S. and around the world have joined it and created other related protests such as Occupy Seattle, Occupy London, etc. The challenge in understanding such a movement lies in distilling the main topics and trends from a movement with massive participation and a wide range of goals such as more and better jobs, more equal distribution of income, bank reform, and a reduction of the influence of corporations on politics [101]. Given the prominent use of social media in organizing the Occupy movement, it should be possible to summarize and analyze how the movement unfolded through analysis of these media.

5.4.1.1 Data collection and preparation

Since we want to focus on the Occupy movement in this scenario, we further filtered for all tweets with hashtag #occupy from our tweet collection. A hashtag is a Twitter convention used to simplify search and indexing. Users include specially designed terms starting with # into the body of each post. The resulting dataset includes more than 100,000 tweets starting from Aug 19 to Nov 01. Our architecture then automatically removed stopwords and performed topic modeling. Such automated process enables us to experiment with different numbers of topics, and results in the choice of 15 topics for interpretability.

5.4.1.2 Investigating the Occupy movement

Exploring the unfolding of Occupy movement. Our analysis environment enables users to explore and follow the evolution of the movement. As the user, a campaign strategist, inspected topics in the temporal view, she noticed that this

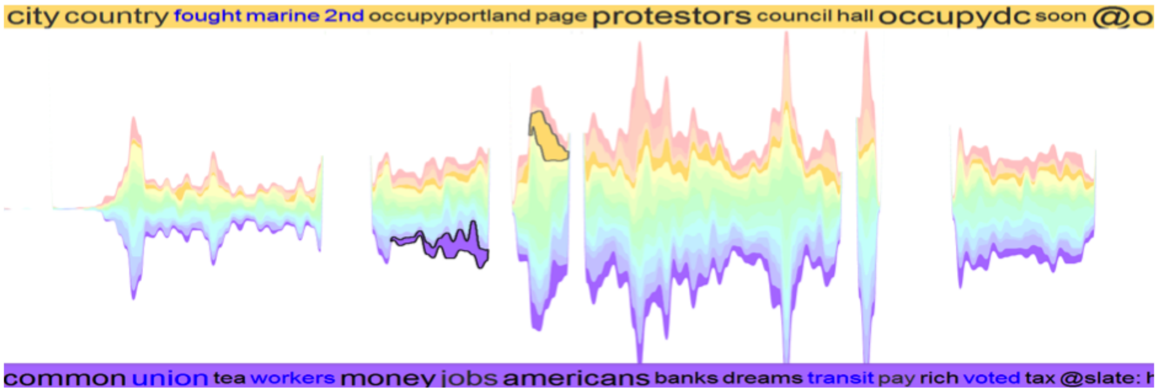


Figure 12: Difference forces joined the Occupy Wall Street movement. Highlighted portion of the yellow topic -marine joined to movement to protect the protesters from the police. Highlighted portion of the purple topic - union workers voted to support the OWS movement.

movement had been evolving gradually over the course of two months. This trend had been exemplified by two significant forces joining the Occupy movement. Specifically, as shown in figure 12 (yellow topic), marines joined the Occupy Wall Street (OWS) movement to protect the protestors from the police on Oct 1. The user reached this conclusion by selecting tweets related to the topic of interests within the burst of topic volume for this event (see Figure 12). She noticed that people were shouting out on twitter about this event:“...the marines coming to protect protesters” and “marine - 2nd time fought for my country time 1st time I've know my enemy”.

A similar pattern was also seen for another topic, which suddenly gains momentum as the NYC transit union workers joining OWS (shown in figure 12, purple topic). People sounded excited on twitter about the event: “200 000 transport workers union votes support!!!”, “new york transit workers union voted unanimously support #occupywallstreet. 38 000 active march oct. 5”. More interestingly, based on reading the last tweet, the user suspected there might be an organized march on Oct 5. Indeed,

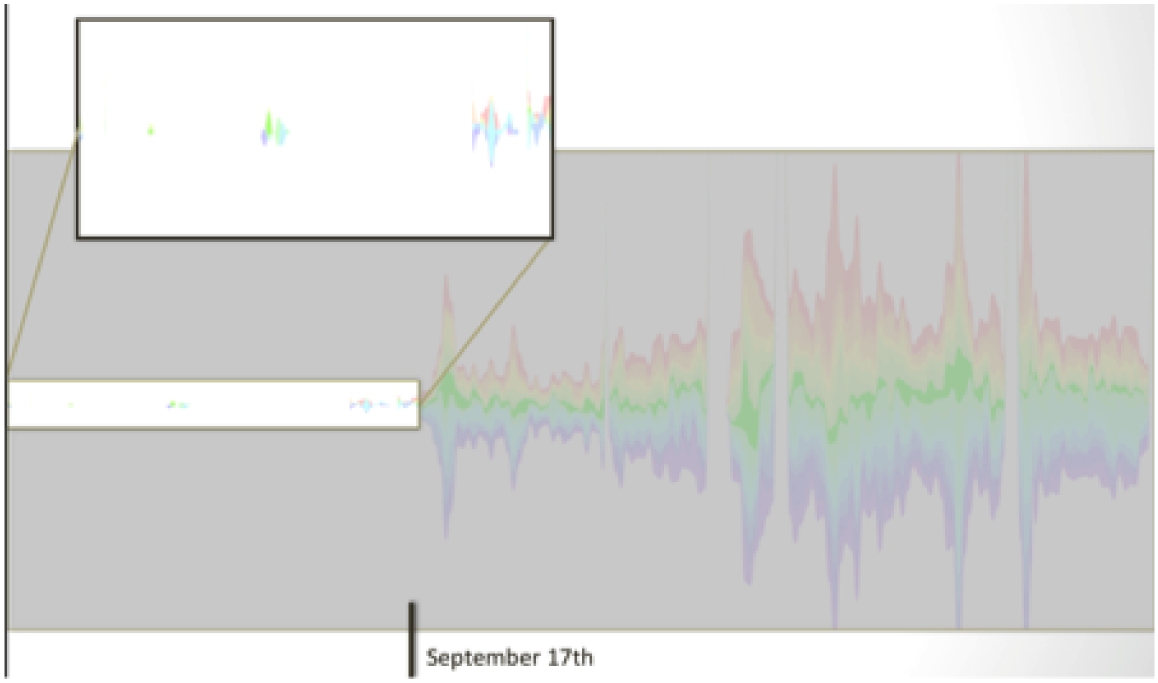


Figure 13: Precursor to the Occupy movement. People started to organize and advertise the event way before the official beginning date - Sep 17.

another big increase in volume of the same topic occurred on Oct 5, and the tweets were related to the march though the Financial District of Wall Street, which was joined by thousands of union workers.

Identifying pre-cursor to the Occupy movement. In addition to identifying meaningful events based on the sudden topical volume change in the themeriver, our analysis environment also enables users to construct a comprehensive story by looking at the overall movement. As shown in figure 13, the overall volume of tweets with "#occupy" became significant around Sep 17, 2011, which is the protestors' self-proclaimed start date of the movement. However, our temporal view clearly indicated relevant tweets were posted well before Sep 17, dating all the way back to Aug 19, 2011 (highlighted region in figure 13).

This unique pattern could suggest a pre-cursor to the Occupy movement, and motivated the user to look further into the details of the tweets. With our coordinated views, she was able to directly click on each time frame to inspect tweets one time step at a time. Upon reading the tweets, the user immediately realized that the OWS was a well-organized event. Specifically, organizers had been using Twitter to advertise the upcoming event and to raise media attention as early as Sep 11, with tweets stating: "trainings! medic! legal support! communication training. facilitator trainer #occupywallstreet #sept17". As the actual event (Sep 17th) drew near, the organizers were giving more specific instructions, as they posted on the 14th :"bringing tent sleeping bag food water to new york this weekend!" And even on the early morning of the 17th, protestors were provided maps of how to get to Zuccotti Park: "hashtags user-friendly time-table map uploaded". At this point, it became obvious that the initial protest was orchestrated by a group of organizers.

To seek the possible origin of the movement, the user kept retracing the tweets published earlier than the 14th and noticed that there were other hashtags that frequently co-occur with #occupywallstreet during the first few days of the movement. These hashtags include: #usdayofrage, #yeswecamp, #nyccamp, etc. At this point, the user could carry the investigation further by looking into tweets with these hashtags around that time. Such observation was validated by recent Wikipedia's updates on Occupy movement where the U.S. Day of Rage (#usdayofrage) was considered the governing body of the OWS group [9].

In summary, the I-Si framework supports the analysis of social media data regarding the Occupy movement. The backend topic modeling plus frontend interactive visualization supports investigative analysis of the otherwise unorganized and noisy information, and enables the answering of questions such as how did the movement evolve, which forces joined the movement at which time, are there any precursors to the Sep 17th protest, etc. As detailed in Section 5.5.2, the implication of this finding can be significant to public safety personnel in that, if they are able to acquire the key indicators hours/days before the protest, they can develop a better oversight and management strategy.

5.4.2 Scenario II: Establishing Investigative analysis

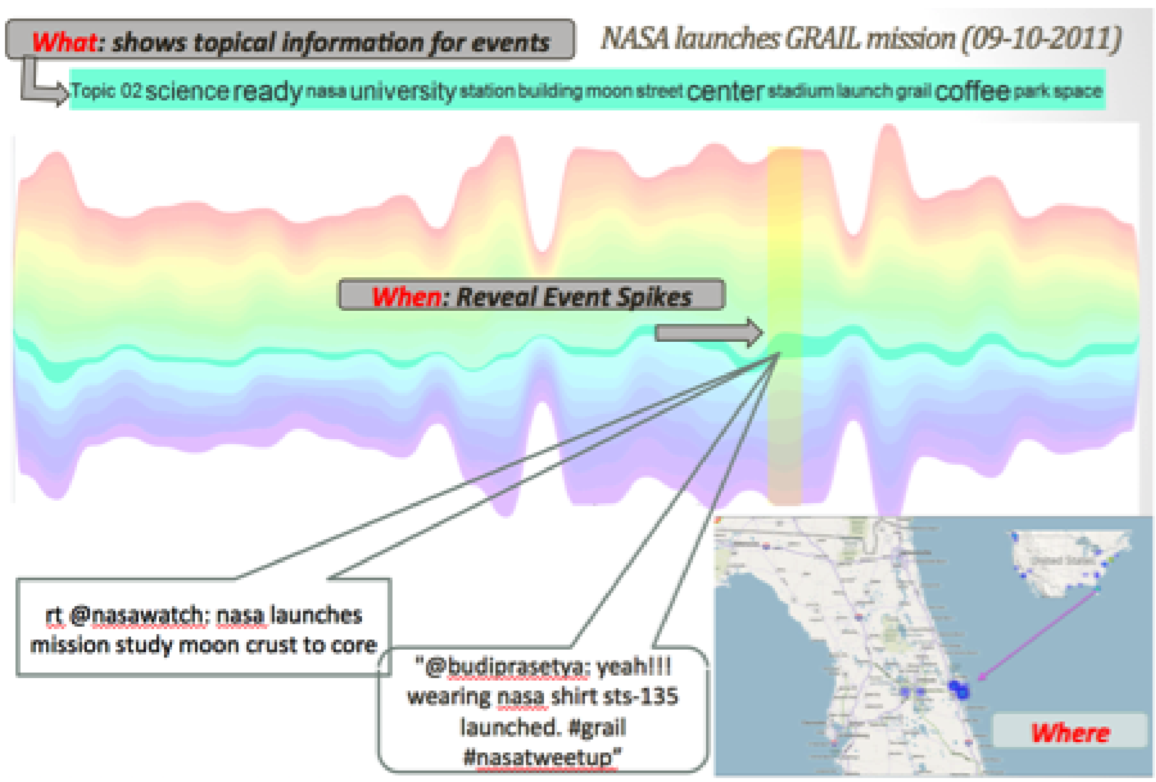


Figure 14: Topical burst indicates NASA launching GRAIL mission to explore the moon.

We use this scenario to demonstrate the scalability of our architecture, demonstrating that the I-Si interactive analysis environment allows users the capability to tap into relevant data on a large scale. Over 12 million tweets were examined over the course of three weeks, with 30 topics extracted for interpretability. Unlike the previous scenario, these tweets were not filtered by hashtag. Thus exploration of the dataset will permit it to tell the user what it is about.

In this scenario, a summer research intern began by examining this tweet collection to discover interesting events he might have missed during the past few weeks. Upon highlighting different ribbons in the ThemeRiver (When) view, he notes that the cyan ribbon (see Figure 14) exhibits a unique temporal pattern. A closer examination of the time-line reveals a volume burst around Sep 10, suggesting more tweets were related to this topic within that time period.

The student then associates this timeline view with the topic cloud view (What), and finds that the topic refers to science and NASA. This becomes very interesting to the student, who happens to be an enthusiast in astrophysics. Quickly he references the tweets with their geospatial location on the map (Where) view. He observes that the mention of such event is mostly centered in Orlando, FL, where one of NASA's launching sites is located. Further browsing through the actual tweets suggests that people (Who) across the country are excited about this event. At this point, the student has linked all these investigative hints together and checked into the news database to find media coverage. He then correctly concludes that the event is the NASA GRAIL launch on Sep 10 to study the moon from crust to core.

The investigation in this scenario included all four of the W's and ended in the correct hypothesis of how a single space event could stir discussion and may inspire people toward scientific activities. It uncovers the cause (an event) of a certain volume burst in large-scale social media data and also makes clear the trend and pattern of social phenomena.

5.4.3 Scenario III: Identifying Epidemic Spread

In this scenario, we demonstrate how I-Si can support investigation of the spread of an epidemic and pinpoint when the epidemic happened by analyzing microblog messages.

5.4.3.1 Data source and data preparation

In contrast with the other two scenarios, the dataset in this case is provided by the VAST Challenge 2011 committee [7]. The dataset contains more than a million microblog messages collected within a major metropolitan area, over the course of a month. While synthetic, this dataset can be a great benchmark for its careful integration of the epidemic theme, the investigation of which requires robust analysis capability since the epidemic is buried in a mass of irrelevant microblog data. This VAST challenge is also a great evaluation of our I-Si environment, since it permits comparing the patterns observed from the visualization interface with the ground truth that comes with the dataset.

Relying on the robust text analytics capability in I-Si, we were able to effectively perform topic modeling over this textual corpus, with a vocabulary of 13,284 unique terms. 10 topics were extracted and visualized from the corpus after several experiments of topical interpretability.

5.4.3.2 Characterizing Epidemic Spread

As shown in Figure 15, the temporal patterns of the ten extracted topics are presented in our ThemeRiver view over the course of a month. Each time unit in the

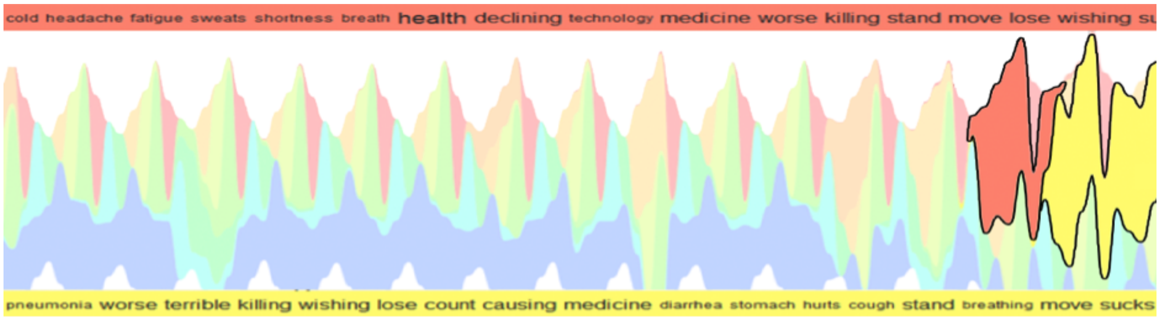


Figure 15: Identifying the start of an epidemic spread. The orange topic captures flu-like symptoms. The topic bursts into being on day 1. The yellow topic shows on the next day, the symptoms have evolved to more severe ones such as pneumonia and diarrhea.

figure denotes 4 hours, which is adjustable to support inspection of different temporal granularities. Upon exploration of the temporal view, one can easily discover that multiple topics share a repetitive characteristic, such as the repeating mentions of TV shows every night (un-highlighted topic in red).

What really attracted the users' attention, however, is the sudden disappearance of the repetitive patterns during the last 3 days. Instead emerging topics in that time frame were about flu-like symptoms, such as "cold, headache, fatigue, sweats, etc." (see orange topic in Figure 15) and "pneumonia, diarrhea, cough, etc." (see yellow topic in Figure 15). These two topics signify exactly when the out-break has begun. Moreover, our temporal view clearly suggested a progression of the illness from cold and headache to more serious symptoms, such as pneumonia, diarrhea and difficult breathing, since the orange topic stream appeared before the yellow topic stream. The results conform to the ground truth provided by the challenge committee. Finally, with the ability to pinpoint when the epidemic has begun, one can further conclude that the disease did not seem to be contained based on the volume of the yellow topic in the last day. Therefore, if the microblogs were collected and analyzed as the epidemic unfolded, the results could inform emergency responders to take actions to prevent the disease from spreading.

In summary, the I-Si architecture supports processing and analysis of the microblog messages. Through interactive exploration of the visualization results, a user could successfully identify latent information regarding the outbreak and depict temporal patterns of the epidemic spread.

5.5 Preliminary User Feedback

Compared to typical visual analytics evaluations, we recognize the challenges in conducting thorough evaluation of the I-Si architecture. This evaluation process, to bring it to completeness, may require experts from multiple research domains to collaboratively examine the efficiency and effectiveness of the scalable architecture. We believe that any findings from such evaluation would be of tremendous value; yet, conducting it can be a longitudinal process that needs more strategic consideration and explanation that beyond the scope of this chapter.

Instead of focusing on evaluating the architecture as a whole, we seek users' understanding about the analysis environments that our architecture enables. In this section, we report user feedback based on several preliminary user evaluations with our investigative visual analysis interfaces. These evaluations were conducted to assess the effectiveness and efficiency of such an interface in supporting understanding latent social phenomena such as the three case studies shown above.

In particular, we report our interactions with three groups of experts in political campaign planning (CP) (5 experts), finance (4 analysts) and emergency response (3 lead experts). While the evaluations were conducted informally, these outreach activities granted us a sufficient amount of time to introduce our architecture and its visual interfaces as well to gather their feedback. First we presented our system by demonstrating the investigative scenarios described in the previous section. Then the experts were given some time to ask questions regarding the system and the interface. Finally, we concluded the evaluation by asking them to give feedback and comments. Given privacy concerns, we are removing all these experts' affiliations. However, they all agreed to have their comments published in this section.

5.5.1 Monitor and Analyze Social Phenomena

One of the benefits that all these experts see in the I-Si architecture is its capability in helping to depict latent social phenomena that are otherwise hidden in the data. Especially to CP strategists, who are responsible for analyzing hundreds of political blogs and news on a daily basis, the capability to identify and summarize the latent topics from their data is of great value. One of the experts mentioned that, "this tool is very exciting in that it could give me a way to effectively assess what people are talking about with regard to political events." He further commented that this would provide a great baseline analysis for their strategic planning work, "where [their] line of business is about finding the right people and talking about the right things".

While the analysis environment was well received, one of the CP experts pointed out that trust issues and uncertainty might affect analysis outcomes. Given the accuracy needed in the CP's work, they were interested in learning how they can effectively validate the outcomes of the overall analysis architecture. The validations and quantification of analysis outcomes is certainly a crucial future direction as we continue enriching our architecture. Of course, it would be possible to validate a particular topical outcome by merely reading enough of the blog entries organized around that topic, but it would be good to have the visualizations of the automated results show this at once or with limited probing of details. Otherwise the approach is not scalable.

5.5.2 Potentially be Proactive to Key Event Indicators

As mentioned in scenario I (section 5.1), the I-Si architecture helped depict key event indicators for the Occupy movement. This capability is highly appreciated by emergency responders; and our results demonstrated great potential in facilitating their duties. The ER experts we interviewed were very excited to see the system in action. One of them indicated that the potential of having I-Si in their working environment could not only help them "follow up with what they knew", but also raise their awareness on "what they didn't expect". One usage case they pictured to use our system is for proactive measures for political events. They would like to utilize our tool to deploy their manpower in more targeted directions.

Due to the "limited resources (financially and personal-wise)", an ER manager mentioned that, "we can't respond to every small indicator that the system provided us." This requires our architecture to be able to perform more comprehensive event structuring, producing more a hierarchy with key indicators. His comment is well received, and we are working extensively on researching a quantifiable event structuring metric for ascertaining where attention is needed and resources should be deployed.

5.5.3 Follow the Influence of Social Events

Based on our interaction with marketing experts, one of their strongly emerging interests is utilizing the social media data to depict marketing impacts that are generated by certain social events. They see our architecture could potentially help them to follow their customer base, and understand their interests. As summarized by one of the experts, "this system ties the marketing loop back to us...It could help us to find a targeted audience and pursue that market with customized approaches".

While we demonstrated our data connection between structured data (e.g. GPS locations) and unstructured text, the further fusion of heterogeneous data is another important aspect for which these experts would like to have further evaluation. In particular, they are interested in learning how we could effectively associate information from different text corpora. This should certainly be doable at the topic level.

5.6 Limitations of the I-Si Framework

There are limitations to this research that need to be addressed. The current analytics capability of our architecture is limited because this research was conducted within the specific Natural Language Processing area of topic modeling. We attempted to mitigate this limitation by componentizing our architecture, which opens up opportunities to incorporate other text analytics methods such as sentiment analysis and named entity recognition. Nevertheless, different characteristics, other natural language processing algorithms, and their scalability constraints could engender different analytical environments.

In addition, we undertook this architectural research to depict information of social media data from an analysis perspective. Our data management schema, which resides in HADOOP clusters, is still preliminary. We are in the process of determining more comprehensive data collecting and integration schema to handle the ever-growing complexity and scale of social media data. An important benefit of integrating an optimization process into our architecture is the potential for improved efficiency of the infra-structure, allowing informed resource management, avoiding replicated work.

The presented architecture illuminates the strong role that a combined approach of data-driven modeling algorithms and user-center visual analytics plays in revealing the latent phenomena within complex social media. It is our hope that by identifying these system limitations, the research fields of visual analytics, parallel computing and databases might be brought together, providing scalable solutions for social media analysts and new techniques for revolutionizing the analysis environments.

5.7 Future Work and Conclusion

In the future, we would like to enrich each component within the I-Si architecture. For the cluster computing, we would like to optimize LDA parallel processing algorithm and further improve its scalability and efficiency. As for the data analytics stage, we would like to add in techniques such as sentiment analysis and named entity recognition to automatically extract more information other than semantic topics from social media data.

In this chapter, we presented a visual analytics architecture, I-Si, to effectively analyze unstructured social media data on a large scale. I-Si integrates data driven analytics methods such as topic modeling with human-centered visual analytics, via an interactive visual interface. We demonstrate multiple investigative visual analysis environments that I-Si is able to provide for monitoring, analyzing, and potentially enabling response to latent topical information extracted from social media. The I-Si architecture empowers users to summarize massive amount of unstructured information and further enables users to explore to answer questions regarding topics and temporal patterns. Thus the architecture supports domain users to perform their analysis tasks without being limited by the data scale through providing solid back-end data processing capability.

CHAPTER 6: EXTRACTING INDIVIDUAL'S ANALYTICAL PROCESSES

As the previous chapters have described, understanding target users' analysis processes is essential to a well-designed visual analytics system. Visual analytics systems facilitate domain users' analysis processes through interactive, visual means. After a system has been designed, there is always potential room for improvement. And the improvement could come from incorporating an individual user's own analysis process into the existing system. In addition, the extracted analysis processes could further be used for the purpose of self-recall, reporting, knowledge sharing, etc. But how do we capture an individual's analysis process while using an visual analytics system?

To answer the question, we turn to van Wijk's operation of visualization model [118] and examine how a user interacts with a visualization. Based on the model, we propose that there are in fact two separate modes of capturing: internal and external capturing to the visualization. Internal capturing within the visualization includes methods such as screen capturing and interaction logging; whereas capturing external to the visualization includes the use of eye trackers, video camcorders, or advanced machineries such as EEG (Electroencephalography) and fMRI (functional Magnetic Resonance Imaging) . These two modes together represent all possible methods of capturing that are available today, but choosing the appropriate methods will depend on the goal and context in which the visualization is used.

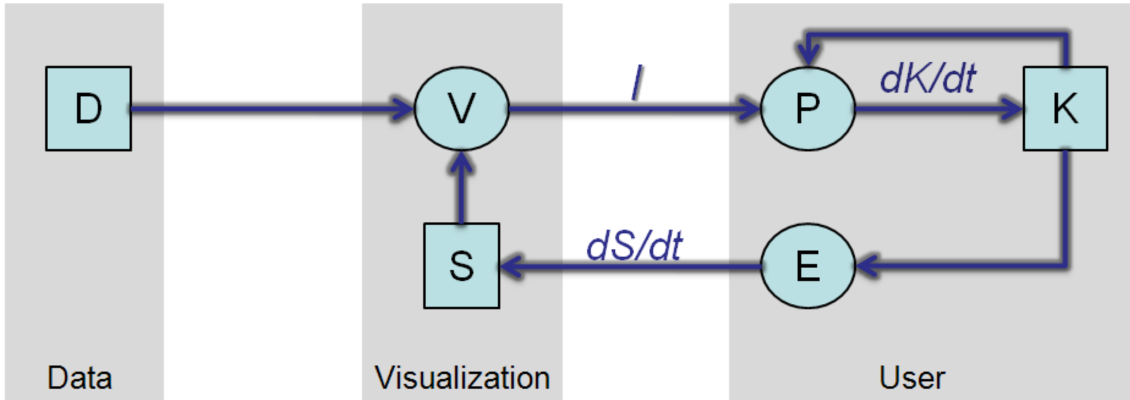


Figure 16: A model of visualization proposed by van Wijk.

The van Wijk operational model (Figure 16), although simple, distinctively depicts the flow and relationship between the user and the visualization. Specifically, there are two connections, I and dS/dt, between the user and the visualization. I stands for the images generated by the visualization that are perceived by the user. And the connection dS/dt represents the changes in the parameters of the visualization initiated by the user (through the use of a mouse, keyboard, or other input devices) that are applied to the visualization to generate the next sets of images I. Both of these connections can be captured directly within the visualization during the user's analysis process by performing screen captures and interaction logging respectively. We refer to these two methods collectively as "internal capturing" (Figure 17(A)).

In real life, however, solving a complex task is not restricted to only using a visualization. The user could jot down discoveries on a piece of paper, or watch the news on the web to gather up-to-date information [78]. In order to fully capture a user's analysis process in solving a task, the user's activities outside of the visualization need to be captured and collected as well. We further categorize the capturing of these activities into two groups: externalization and observation. In externalization, the

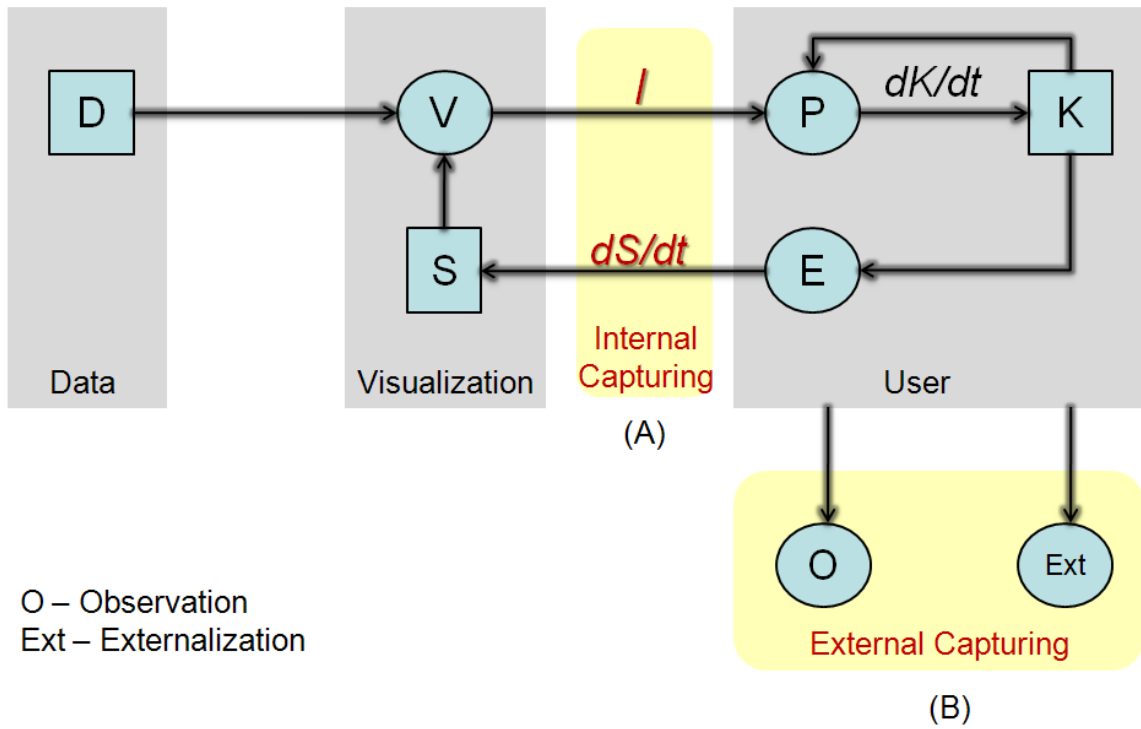


Figure 17: A model for capturing user's analysis process based on van Wijk's model of visualization. The yellow boxes (A) and (B) represent internal and external capturing methods respectively.

results that are explicitly externalized from the user of the reasoning process are collected and stored. These include the notes taken by the user during an investigation, or dictations taken using a voice recorder. In observation, information around the user is captured through the use of additional hardware and machinery. For example, eye trackers can track the user's focus, and a video camcorder can record the user's activities in an environment. In addition, advanced technologies such as EEG and fMRI can be used to monitor the user's neural activities. Together, externalization and observation are referred to as "external capturing" (Figure 17(B)).

6.1 Internal Capturing

As shown in van Wijk's operational model of visualization, the relationship between a visualization and its user can be succinctly summarized with two variables, I and dS/dt. These two variables are the input and output of a user's process in using a visualization, and in many cases can be thought to be directly related. As the model demonstrates, I, or an image, is generated by a visualization given a visualization state S. In using a visualization, a user's interactions can be thought of as the means to modifying the visualization state (dS/dt) to create the images that would lead to the user solving a specific problem. Therefore, it is not difficult to see in this model that by capturing the three variables, I, S, and dS/dt, the system can later faithfully reconstruct a user's session in using a visualization.

In practice, however, I and S can be thought to contain the same information depending on whether or not replaying the user's session involves running the visualization. In fact, Jankun-Kelly et al. have proposed a formal model for capturing S accurately [61], and Bavoli et al. have shown that the captured visualization states S can be used to generated large numbers of I efficiently [11]. The main advantage of storing a series of visualization states S over storing a series of images I is that storing the visualization states often requires less disk space than images. However, in cases where the visualization itself could take considerable amount of time to generate I, storing I could still be more efficient.

Capturing and storing the user's interactions as dS/dt, on the other hand, can be very different depending on the purpose of the capturing. If the purpose of capturing is to faithfully reconstruct a user's session with a visualization, dS/dt can be captured as "events" generated by most operating systems (e.g., MouseClick event or a Keystroke event). Typically, visualizations that gather data provenance would record dS/dt in this fashion (see section 2.3.1). The advantage of storing a series of dS/dt is to further reduce the disk space requirement over storing individual visualization states S since consecutive visualization states often contain duplicate and redundant information.

However, if the purpose of capturing is to reconstruct a user's analysis processes, storing dS/dt as low-level events is inadequate. As Hilbert and Redmiles noted, such events do not carry enough information on their own to allow their significance to be properly interpreted [55]. Most visualization systems that capture information provenance (see section 2.3.2) therefore capture the user's interactions at a higher level that include additional contextual information. In some systems, the additional contextual information are semantically related to the specific data or application [37, 106, 99]; whereas some projects categorize the user interactions according to structures that are relevant to the domain or task [43, 72, 52]. In either case, it is clear that storing only low-level user interaction is not enough for gathering information provenance, and subsequently not sufficient for reconstructing a user's analysis process.

6.2 External Capturing

As mentioned before, in most real-life analysis tasks conducted using visualization systems, not all of the analytical activities actually take place within a visualization. Since these activities are not directly part of the interaction between a user and a visualization system, van Wijk's model is no longer sufficient in describing these activities in relation to the use of a visualization. However, in order to fully understand an analyst's analysis process, it is still very important to consider these activities as they have immediate affects on how a user utilizes the visualization.

We propose that these activities that are external to the visualization can be categorized into two types: externalization and observation. Externalization denotes the methods to capture the artifacts users actively externalize during an analysis process. Examples of externalization include recording user's think-aloud and saving the notes a user jots down, both of which intimately reflect the analysis process at a semantic level. Research in visualization has often relied on externalization mechanisms to understand the behaviors of the user. In particular, the think-aloud protocol is frequently used in evaluations and has been found to be effective in reflecting the user's analysis process [28, 37].

On the other hand, observation represents methods that monitor a user using a visualization during an analysis without requiring the user to actively externalize his thoughts. Besides human observers taking notes of what happens during the analysis, devices like eye tracker, video camcorder, EEG (Electroencephalography) and fMRI (functional Magnetic Resonance Imaging) could also be used to record information about user's eye movement, physical motions, neural activities, etc. As noted separately by Huang [28] and Convertino et al. [56], eye-tracking data offers additional insights into typical strategies used for accomplishing given tasks within a visualization environment. Although other devices like EEG and fMRI have seldom been used in experiments related to visualization systems, they are effective for studying brain functions during experimental tasks [23].

Relying solely on external capturing methods has been shown to be effective in discovering a user's (qualitative) mental model when interacting with visualizations for analysis tasks. In an experiment by Trafton et al. [116], video camcorders recorded both the physical environment in which the experiments took place, as well as how the participants interacted with the various computers in the environment (observation). At the same time, the participants were requested to provide think-alouds of their thoughts and take notes of their incremental discoveries (externalization). By combining and manually analyzing these recordings, Trafton et al. demonstrated that they were able to identify how the participants formed mental models of the analysis task and applied the models to solve problems.

6.3 Internal Vs. External Capturing

Figure 17 illustrates van Wijk's operational model of visualization after integrating both the internal and external capturing mechanisms. According to this model, the four capturing methods are distinguishable and independent from each other. However, in practice, the lines between the methods are sometimes blurry depending on the implementation of the visualization or the physical environment where the analysis takes place. As visual analytic systems become more mature, some analysis that has traditionally been performed outside of a visualization can now be done directly within the visualization. A prime example of this is the inclusion of annotation techniques [47, 54, 40] or "shoeboxes" [126, 106, 121, 91]) in visual analytical tools. Traditionally, such annotations are written on a piece of paper that would have to be collected externally, but in many recent visual analytics systems, annotations have become a part of the visualization that can be captured internally within the visualization.

Practically, one distinguishing factor that separates internal and external capturing can be described based on how intrusive they are to the analyst. Internal capturing methods can be implemented directly within the visualization, and are mostly transparent to the user. External capturing methods, on the other hand, often require physical devices or mechanisms that would alter the physical analysis environment and potentially change the analysis process. Certain eye trackers require the user to wear additional hardware [39] that could be cumbersome. Requiring analysts to perform think-alouds during their analysis could be an annoyance to other analysts [85], just as the use of EEG or fMRI are most likely unfeasible due to the monetary cost of the machinery and the cost of time in the setup process prior to use. Even more importantly, in most cases involving external capturing, the fact that the analysts are externalizing their thoughts (e.g., via think-alouds), or are reminded of potential observers (e.g., in the case of being recorded on video) could change their behavior significantly. As noted by Shapiro, performing the think-alouds protocol may slow down a participant's task performance and even alter the process of interest [103]. Similarly, the use of observational tools could solicit an effect known as social facilitation and inhibition in which the participant would either over perform or under perform depending on their confidence in performing the task [13, 128].

6.4 Summary

Based on the analysis of the cost and benefit of both internal and external capturing methods, internal capturing seems to be the winner because of its un-intrusiveness and affordability. Therefore, the rest of the dissertation will focus on how to capture the interaction steps and visualization state during the use of a visual analytics tool. Since visualization states could be easily recovered given interaction steps, we will be mainly evaluating capturing user interactions with a visual analytics system.

First, from a theoretical perspective, studying the outcome of the analysis process under different interaction constraints validates the importance of interaction to a user's analysis process (chapter 7). Experimental results have shown that constraining the way a user could interact with a problem significantly affects the outcome of the problem solving process. Interaction connects users and the visual representations to enable human-information discourse, and at the same time interaction dictates the analysis process.

Then from a practical viewpoint, we conducted user studies to capture and analyze sequences of interactions conducted by a user with an interactive visual interface. This may provide insight into the user's reasoning process. The result of our experiment suggests that at least 60% of the high-level reasoning process can be recovered from merely analyzing the interaction logs. The recovered reasoning process could be documented and potentially be used for recall, reporting, knowledge sharing, and training (chapter 8).

CHAPTER 7: EVALUATING THE EFFECT OF INTERACTION CONSTRAINTS ON PROBLEM-SOLVING PROCESS

Given the goal of extracting a user's analysis process from user interactions, we first need to understand the relationship between the high-level analysis process and user interactions. In other words, we want to evaluate the effect of user interactions on the analysis process. If the process changes under different constraints enforced upon user interaction, then there is indeed an intimate relationship between the two. To meet this aim, we designed an experiment to study the outcome of analysis process under various interaction constrains.

Interaction and manual manipulation have been shown in the cognitive science literature to play a critical role in problem solving. Given different types of interactions or constraints on interactions, a problem can appear to have different degrees of difficulty. While this relationship between interaction and problem solving has been well studied in cognitive science literatures, the visual analytics community has yet to exploit this understanding for analytical problem solving. In this chapter, we hypothesize that constraints on interactions and constraints encoded in visual representations can lead to strategies of varying effectiveness during problem solving. To test our hypothesis, we conducted a user study in which participants were given different levels of interaction constraints when solving a simple math game called Number Scrabble. Number Scrabble is known to have an optimal visual problem isomorph, and the goal of this study is to learn if and how the participants could derive the isomorph and to analyze the strategies that the participants utilize in solving the problem. Our results indicate that constraints on interactions do affect problem solving, and that while the optimal visual isomorph is difficult to derive, certain interaction constraints can lead to a higher chance of deriving the isomorph.

7.1 The number scrabble problem

The original Number Scrabble [109] is a game played by two people with nine cards: ace through nine. The cards are placed in a row, face up. The players draw alternately, one at a time, selecting any one of the unselected cards. The objective of the game is for a player to get three cards which add up to 15 before his opponent does. If all nine cards have been drawn without either player having a combination that adds up to 15, the game is a draw.

The main reason we chose to use the Number Scrabble game is that there is a known visual isomorph of the problem called the "magic square" (figure 18). Since the magic square visually represents all possible combinations of three numbers that can be added up to 15 in a succinct manner, it can significantly help a player to perform well at the game. In other words, once this visual isomorph is identified, the Number Scrabble problem is turned into a much simpler tic-tac-toe game, which is played by two players who take turns marking the spaces in a 3*3 grid. The number scrabble game represents a large number of well-defined problems that show how visual isomorphs can make evident what was previously true but obscure [109].

2	7	6
9	5	1
4	3	8

Figure 18: 3x3 magic square

7.2 Isomorphs and diagrammatic reasoning

Simon defined problem isomorphs as problems whose solutions and moves can be placed in one-to-one relation with the solutions and moves of the given problem [109]. The key to isomorphism is that even when two representations contain the same information, they can still provide very different sets of operations for accessing and inferring about that information, which can make a given problem easier or harder to solve [76]. In our example, the magic square and number scrabble are isomorphs of the same problem in that they both contain all the information needed to play the game. However, in number scrabble, the operations provided to the player to access important information about the game—such as whether your cards contain a winning combination—are mathematical. In the magic square case, that information is contained in a visual operation: seeing whether the cards form a line across the magic square grid. Since the brain processes such visual operations faster than mathematical ones, the visual isomorph is more efficient in this case.

The idea that visual representations make certain operations more efficient to perform is at the core of the theory of diagrammatic reasoning [26, 76]. However, efficiency is not the only measure of interest in visualization; our goal is to make information not just accessible, but understandable. The distinction between these goals is highlighted by Carroll et al. [21], who had participants solve a design problem presented as one of two isomorphs: a spatial arrangement problem and a temporal scheduling problem. The spatial isomorph was easier and faster for participants to solve and led to fewer failures to understand the problem. That is, in the temporal case there were several participants whose solutions did not follow the requirements of the task. Interestingly, when participants in both cases were provided a simple graphical representation (a grid) in which to work on their solution, the temporal case was as easy to solve as the spatial one, but participants in the temporal case remained more likely to fail to understand the problem requirements. The authors took this to mean that appropriate graphical representations can make problems easier to solve, but not necessarily easier to understand.

Another way to interpret this is that there is more to designing a visual isomorph than making information more efficient to access. Much of the power of visual representations comes from how they set constraints on interpretation and reasoning. Constraints inherent in visual isomorphs can encode constraints on the information they represent, leading to a more direct preservation of information structure [89]. As Stenning and Oberlander [114] argue, these constraints inherent to visual representations help to meaningfully restrict the number and kinds of inferences that can be made about a problem, focusing processing power on only valid cases. In this way, visual isomorphs can not only make operations more efficient, but can also model the constraints of a problem directly. This can affect the difficulty of solving a problem by reducing the cognitive load of remembering rules [71] or by encouraging different types of strategies [51].

7.3 Interaction and problem solving

While visual representations can aid problem solving significantly on their own, they gain even more power to model a problem when interaction is introduced. Interaction is increasingly seen as central to the process of reasoning with visualization [79, 93, 117]. Lending weight to the intuition that interaction improves reasoning, Hundhausen et. al [57] found that interacting with an algorithm visualization produces better understanding than viewing an equivalent animation.

We use the term "interaction" in the broad sense defined by Yi et al.: "the dialogue between the user and the system as the user explores the data set to uncover insights" [127]. In this sense, the relationship between interaction and problem solving has been the subject of much research by cognitive scientists in the field of distributed cognition [58]. In particular, David Kirsh has extensively argued that projection and interaction with external representations are fundamental to human reasoning [65, 66, 67, 68, 69]. Kirsh points to the pervasive use of external representations and interaction with the world in everyday problem solving, and identifies several functions performed by interaction in the reasoning process [66]. Of these, most relevant to our work is *reformulation*, or the ability to restate ideas. Kirsh sees reformulation as a process that is frequently too complex to perform entirely in memory, and so is often managed with external tools. Since reformulation is closely related to identifying different problem isomorphs, we argue that this process can also be made easier through certain types of interaction.

7.4 Hypotheses

Our research objective is to investigate the question of how constraints on interaction affect problem solving through the derivation of visual isomorphs. We propose that in developing a strategy for playing a game like Number Scrabble, participants will tend to derive an isomorph for the problem that is easier for them to use than the representation in the original game, and that the availability of different levels of interaction while strategizing will lead to different types of isomorphs. If this is the case, it can help to clarify the relationship between interaction with visual representations and reasoning. To what extent does the nature of a visual representation, and the type of interactions a user is allowed to perform upon it, affect the kind of strategy that user develops for solving a problem?

We therefore designed a study based on the aforementioned Number Scrabble game due to its known optimal visual isomorph, the magic square. In our study, we developed 5 different interaction conditions, ranging from free-form to very restrictive, and studied how strategizing under these conditions affects problem solving and the development of isomorphs. In particular, we propose three interrelated hypotheses concerning interaction, problem solving, and isomorphs:

1. **Interactions and Problem Solving**: We hypothesize that different types of interactions will affect the participants' performance in playing the Number Scrabble game. Specifically, we hypothesize that more constrained interactions can encode more information, and will therefore lead to better problem-solving.

2. **Interactions and Isomorphs**: We hypothesize that the different constraints

on interaction will affect the isomorphs generated by the participants. With higher constraints on interaction, a participant will be more likely to derive the optimal visual isomorph (the magic square).

3. **Isomorphs and Problem Solving**: We hypothesize that not all isomorphs developed by participants will be visual, but that visual isomorphs will be more effective for playing the Number Scrabble game.

7.5 Experiment Design

The main factor of interaction constraint had five levels (no interaction, pen and paper, single set of cards, multiple sets of cards, and boundary). Details of each constraint and design rationale will be discussed in section 7.5.3. We used a between-subjects design with repeated measures. Each subject is randomly assigned to one of the five interaction constraint conditions which determines what interactions are available to them during their strategy session. Qualitative measures in our experiment are the types of isomorphs our subject derived during their strategy session. Quantitative measures involved response time and scores on Number Scrabble games played against a computer, using the game interface shown in Figure 19. The computer was programmed to play the game optimally so that it never loses. While our subjects played the game against the computer, we recorded number of games tied or lost and the time it took them to figure out the next move for response time. We alternate who makes the first move between the subjects and the computer for every game played.

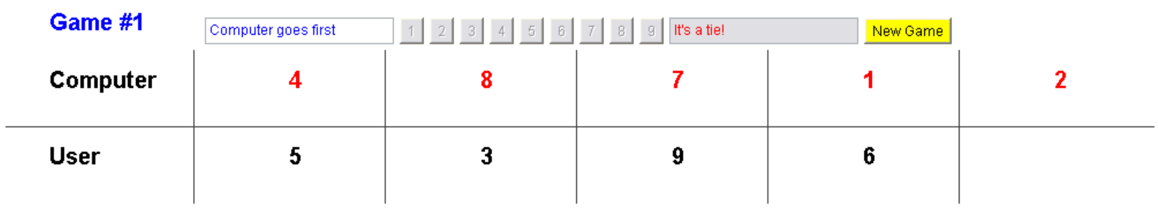


Figure 19: Number scrabble game interface

7.5.1 Participants

We recruited a total number of 117 participants (86 Male, 31 Female) from introduction to computer science courses at our university. Participants' age ranged from 18 to 40 with median of 25. Students were primarily undergraduates, and 80% were in computing-related majors.

7.5.2 Task

The experiment begins with investigators introducing the Number Scrabble game to the subjects based on a training script. The investigators were asked to play the game with the participants until they fully grasped the rules. Next, the participants fill out a demographic form on age, gender and experience with mathematical courses through a web interface. The rest of the experiment can be divided into four major sessions: pre-test, strategizing, externalizing isomorph, and post-test.

1. **Pre-test**: During the pre-test session, the participants were asked to play the Number Scrabble game six times against the computer. To make sure that our participants do not start developing strategies during the pre-test, we enforced a maximum time limit of 18 minutes to finish all six pre-test games. Failing to meet the time limit resulted in a participant's data being dropped from analysis.

2. **Strategizing**: During the strategizing session, the subjects were given 20 minutes and allowed to interact with the materials we provided under different constraints and are told to look for a strategy that can help them play the game better.

3. **Externalizing isomorphs**: At the end of the strategizing session, all participants were given 2-3 minutes to make a “cheat sheet” out of the strategy they developed so that they can refer to it during the post-test session when they play Number Scrabble again. This cheat sheet was a single sheet of paper onto which participants were told they could write anything they felt would help them play the game. (In the case of the pen and paper condition, this was a separate sheet from those they wrote on during the strategizing session.) This gave us a record of the isomorph used by participants in forming a strategy and reduced the cognitive load on participants during the post-test. We only gave them a very short amount of time to make their “cheat sheet” so that they could not continue elaborating on it after the end of the strategizing session.

4. **Post-test**: During the post-test session, participants were asked to play the Number Scrabble game six more times against the computer while consulting their “cheat sheet.” To be consistent with the pre-test and also to make sure that the participants do not refine their isomorphs during the post-test, 18 minutes was set as the upper limit for playing all six games. As in the pre-test, failing to meet the time limit resulted in a participant’s data being dropped from analysis.

After the post-test session, participants were asked to fill out a questionnaire regarding how they arrived at their strategy and their experience during the strategizing session. The investigators collected all the participants' "cheat sheets" for further analysis of the isomorphs they derived during the experiment. In addition, the strategizing sessions were video recorded, which allows us to examine how the interaction constraints affected our participants' behavior during the process of searching for an isomorph.

7.5.3 Interaction constraints

We went through multiple rounds of a refining process to design the interaction constraint conditions used in our study. Our goal was to design constraints that ranged from placing no limit on the interaction to restricting the interaction a great deal.

- Constraint #1 (no interaction): The participants were asked to think about the problem in their head during the strategizing session to develop a strategy to help them play the game better. The participants were not allowed to interact with any materials.

- Constraint #2 (pen and paper): The participants were provided with pen and paper to work out their strategy for the Number Scrabble problem.

- Constraint #3 (multiple sets of cards): The participants assigned to this constraint were provided with multiple sets of cards, with each set consisting of the numbers one through nine. Each card is square in shape and made from paper with the numbers printed on them. Within the strategizing session, the participants were encouraged to organize the cards freely.

- Constraint #4 (single set of cards): The participants were further limited to interact with only one set of cards labeled with the numbers one through nine.

- Constraint #5 (boundary): This is the most restrictive case. Participants were presented with nine cards and a square space only large enough to fit the cards in a grid, and were told to confine their interactions to that space. Figure 20 shows this condition.

Our conditions are designed so that "no interaction" serves as a control group, and "pen and paper" represents no limit on user interaction. Then, based on both the original description of the Number Scrabble problem and the optimal visual isomorph, we derived the other three interaction constraints from "multiple sets of cards" to "boundary" by adding more constraints on interaction each time, all of which encode some information about the optimal visual isomorph of the problem.

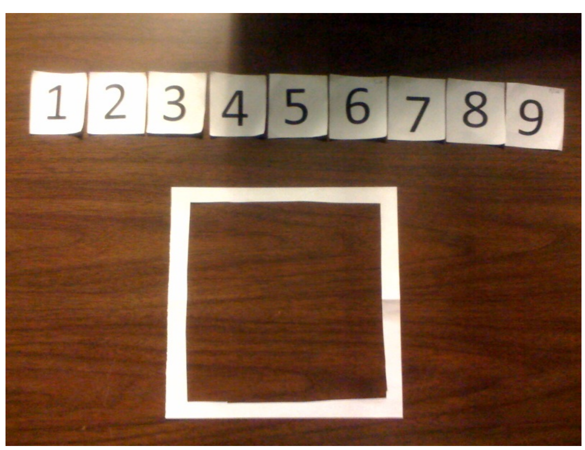


Figure 20: Cards and Boundary

7.6 Results

When analyzing the experimental data, we were concerned with the impact of outliers due to random responses. Therefore, we trimmed out the data of four participants whose response times were unusually fast during the pre-test. In addition, 11 of our participants reached the 18-minute time limit during either pre- or post-test, thus their data are automatically dropped since their missing data made it impossible to fairly compare pre-test and post-test scores. As a result, we have valid data from 100 participants with 20 subjects under each interaction constraint.

7.6.1 Isomorph vs. Interaction constraint

Based on the strategies recorded on their cheat sheets, our participants developed a wide range of problem isomorphs during the experiment. Some of these are visual while the others are either mathematical or purely descriptive. We classified these isomorphs into five different categories:

1. **Magic square (Visual)**: The magic square isomorph.

2. **Partial magic square (Visual)**: Same layout as the magic square isomorph with different ordering or numbers.

3. **Other visual isomorph**: Visual isomorph but numbers are not organized in a 3*3 matrix manner.

4. **Permuted isomorph**: All possible combinations of 3 numbers adding to 15.

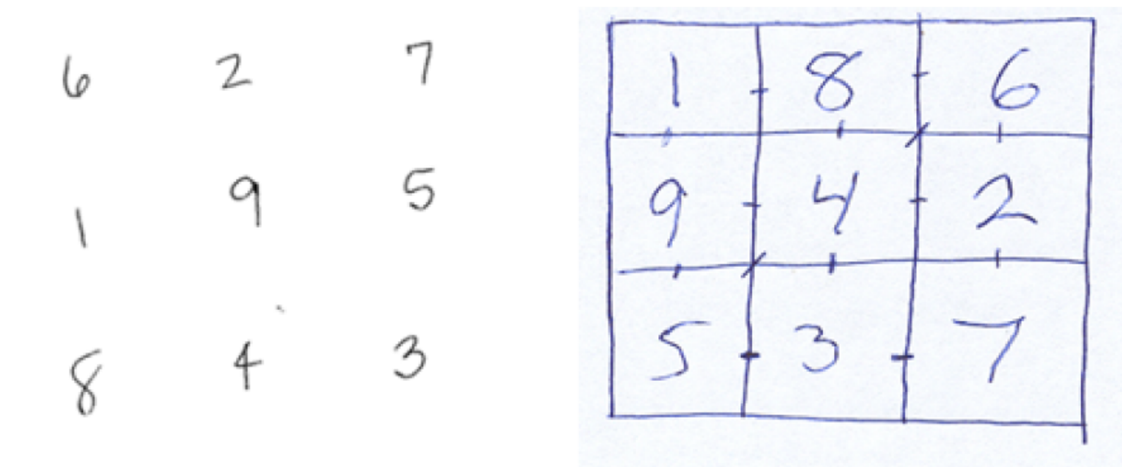
5. **Incomplete isomorph**: Strategies that do not involve all 9 numbers.

Note that categories 1–3 are visual isomorphs of the Number Scrabble problem while 4 and 5 are not. In addition, examples of different types of isomorphs are shown in figure 7.6.1.

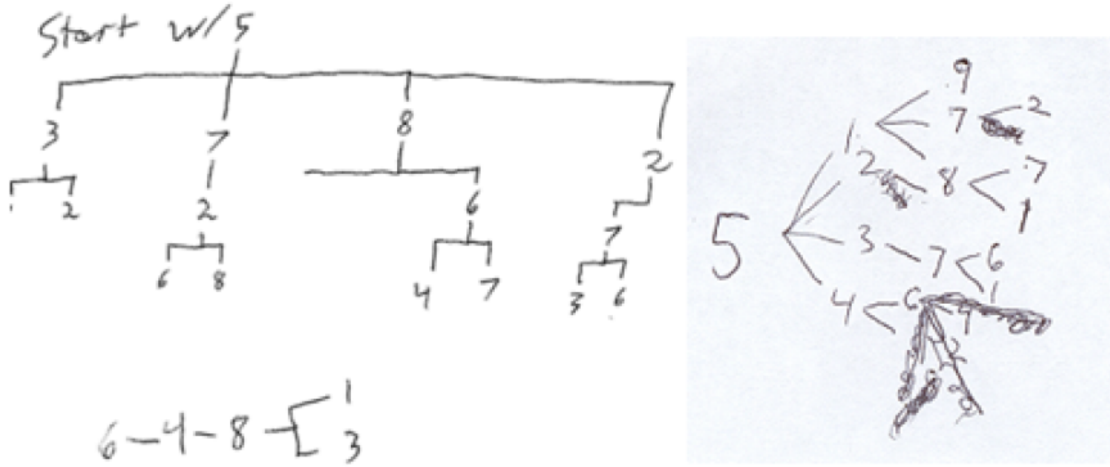
The distribution of different isomorphs developed by our subjects within each interaction constraint is shown in Figure 22. This distribution supports our hypothesis in the sense that as the interactions become increasingly constrained (from pen and paper to boundary), more participants developed visual isomorphs of the number scrabble problem. More importantly, nine out of 20 subjects under the most restrictive constraint (boundary) discovered the optimal visual isomorph (the magic square) while another six subjects developed partial magic square isomorphs. In contrast, only one out of 20 participants in either the no interaction condition or the pen and paper condition discovered any visual solution. A Pearson's chi-square test of independence finds a highly significant interaction between interaction constraint and isomorph, $\chi^2(16, N = 100) = 116.9, p < .001$. Since 15 cells have an expected count of less than five, we performed a Fisher's exact test which also yielded a probability of $p < .001$.

7.6.2 The effect of interaction constraints on Response Time and Score

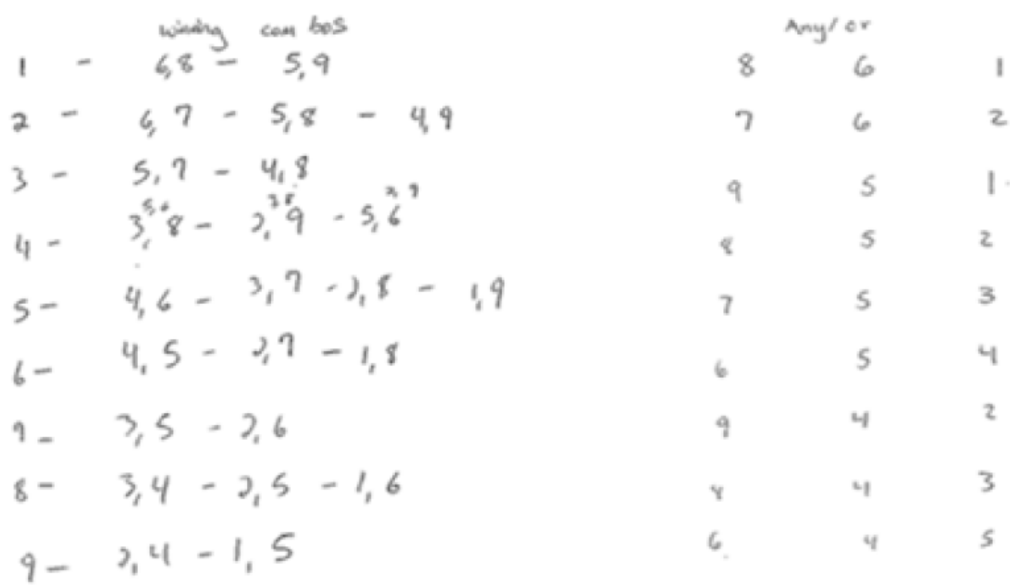
Results regarding time and score were analyzed statistically using an analysis of variance (ANOVA) followed by Tukey's HSD (Honestly Significant Difference) test for pairwise comparisons. The factor in our experiment was interaction constraint (five levels) and the dependent variables were difference in response time and improved score.



(a) Partial magic square examples



(b) Other visual isomorph examples



(c) Premuted isomorph examples

Figure 21: Isomorph examples

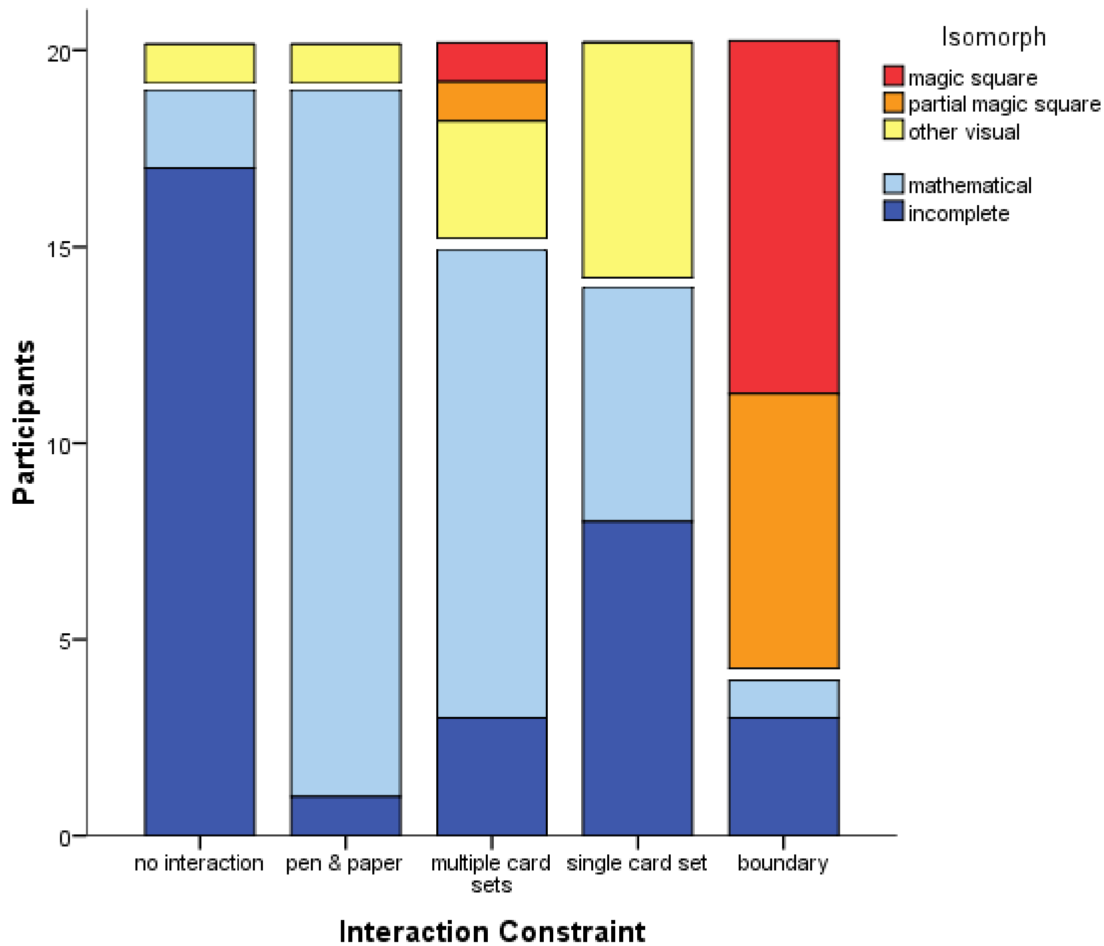


Figure 22: Distribution of isomorphs developed under five different interaction constraints. The gaps divide visual isomorphs (1,2 and 3) from non-visual isomorphs (4 and 5).

Difference in response time is derived from the time it took to decide which card to choose next at each move during a game. Response time per game is defined as the average time it took the participants to choose the next card during each game, $T = \sum ResponseTime/n$, with n being the number of cards chosen following the opponent's move during a specific game. Since both the pre-test and post-test sessions comprise six games, difference in response time is thus defined as $IT = \sum_{i=1}^{6} T(i, posttest) - \sum_{i=1}^{6} T(i, pretest)$. In a similar vein, improved score is derived from whether the subjects tied or lost to the computer during each game, with tying counted as 1 point and losing as 0 points. Thus improved score is defined as $IS = \sum_{i=1}^{6} S(i, posttest) - \sum_{i=1}^{6} S(i, pretest)$.

7.6.2.1 Response time

We expected participants to choose the next card faster during the post-test as the interaction constraints increased, since we hypothesized that they would be more likely to derive a better visual isomorph similar to the "magic square". However, we did not observe a significant main effect of difference in response time ($F(4, 95) = 1.54, p = 0.097$). Figure 23 (top) shows the difference in response time under different interaction constraints. However, interesting yet surprising findings emerged once we considered response time during pre-test and post-test separately. Figure 23 (bottom) shows the mean response time during both pre- and post-tests under the five interaction constraints. It should be noted that participants in the no interaction condition had an unusually slow average response time in the pre-test, which makes comparisons between that condition and the others problematic. In general, however, we found that most of our participants spent more time deciding which card to choose next during the post-test, and participants under the most confined constraints took the longest time, which ran counter to our expectations. We discuss possible reasons for this in Section 7.7.4.

7.6.2.2 Score

If we consider mean scores on the pre-test and the post-test separately (Figure 24 (bottom)), it is clear that in general our participants scored higher after the strategizing session under all five interaction constraints ($F(1, 1190) = 57.7, \eta_p^2 = 0.046, p < .001$). More importantly, the subjects in the more constrained interaction groups tend to score higher than those in the less restrictive interaction groups.

For improved score (Figure 24 (top)), we observed a significant main effect of interaction constraint type ($F(4, 95) = 6.5, \eta_p^2 = 0.215, p < .001$). Post-hoc tests showed that the improved scores are significantly different between numerous pairs of interaction constraints. To elaborate, the improved score for participants assigned to interaction constraint #5 (boundary) is significantly larger than that for participants assigned to interaction constraint #1 (no interaction), $p = .001$, constraint #2 (pen and paper) with $p < .01$, and constraint #4 (one set of cards) with $p < .01$. Although the result of other pairwise comparisons were not significant, we can see a clear trend (Figure 24 (top)) that as the interaction constraints become more restrictive, the improvement of score increases except in the case of constraint #4. We further analyze this unexpected "dip" in the discussion section.

7.6.3 The effect of isomorph on improvement of score

Overall, the main effect of types of derived isomorph is significant ($F(4, 95) = 8.495, \eta_p^2 = 0.263, p < .001$) on improved score (figure 25). Post-hoc tests showed that the improved scores for participants who derived the magic square isomorph is significantly higher than for participants who derived partial magic squares at $p < .05$, and significantly higher than those of all other participants at $p < .01$. The result supports our hypothesis that the optimal solution does lead to much better performance in terms of accuracy. Although the other pairs are not significantly different on mean improved score, we can see a trend that as the isomorphs are further from the optimal magic square, the mean improved score decreases. We further performed a linear contrast between visual isomorphs (1, 2, 3) and non-visual isomorphs (4, 5) on improved score. The result shows that the mean improved score for participants using visual isomorphs is significantly larger than for those using non-visual isomorphs ($t(95) = 3.822, p < .001$).

7.7 Discussion

We start our discussion by addressing the key questions based on our hypotheses:

7.7.1 Do more confined interaction constraints yield a better chance of deriving a visual isomorph?

Yes, based on figure 22 and the chi-square analysis (section 7.6.1), we observe that as the interaction constraints are increasingly restricted, larger number of visual isomorphs are developed. In addition, the strictest interaction constraints led to the highest number of the optimal visual isomorphs discovered. Nine out of 20 participants under constraint #5(boundary) discovered the magic square isomorph during the strategizing session and seven participants out of the remaining 11 discovered a partial magic square isomorph. Based on further analysis of feedback about the interaction constraints, most participants under this condition found constraint #5 very helpful in their discovery of the visual isomorphs. Many of them left comments such as, "It helped me visualize the problem and make competitive moves." Similarly, most subjects under interaction constraints #3 (multiple sets of cards) and #4 (one set of cards) felt that being able to manipulate the cards freely was helpful. Thus both statistics and user feedbacks support the hypothesis that interaction constraints significantly affect the types of isomorphs users are able to derive by altering the way participants approach the same problem. In other words, the manipulation of the isomorphs could be embodied in the interaction.

7.7.2 Does a more advanced visual isomorph outperform a non-visual isomorph in terms of score?

Yes. We consider an isomorph as more advanced if it is more similar to the optimal visual isomorph (the magic square). Thus our results summarized in Section 7.6.3 confirm that visual isomorphs lead a greater increase in score compared to non-visual isomorphs. What's more, within the group of visual isomorphs, the optimal visual isomorph outperforms the other two significantly.

7.7.3 Does more confined interaction constraint always yield larger improvements on score?

The short answer is: not always. As seen in Figure 24, the general trend shows that as the interaction constraints become more restricted, the improved score tends to rise, with the exception of constraint #4 (one set of cards). The low improved score in this condition can be explained by considering Figure 22, which shows that none of the participants under this condition derived a magic square (red) or partial magic square (orange) isomorph. Without more efficient visual isomorphs, it made sense that the subjects did not do much better in their post-test compared to the pre-test. However, when we designed the five interaction constraints, we considered one set of cards as a highly restrictive constraint, thus we expected better scores and more derivation of the optimal isomorph. Based on the comments they left, many participants in this condition felt limited by only being able to interact with one set of cards and wished they were given paper to write down combinations of numbers they found to offload the burden of having to memorize them. After the experiment, when we present the magic square isomorph to participants, most in this condition thought they were close to discovering the optimal isomorph at some point during the experiment. But without the extra boundary to further constrain their interaction, it was hard for them to find the bridge between one set of cards and the magic square. This finding highlights the fact that more restrictive interaction constraints are not necessarily helpful unless they meaningfully encode information about the problem. The single set of cards constrained interaction, but without the boundary this constraint did not by itself tell participants anything about the nature of the problem.

7.7.4 Why is difference in response time not a good measure?

Unexpectedly, we did not observe a significant result of isomorph type in terms of both post-test response time and difference in response time. In fact, response times in the post-test were generally longer than in the pre-test, and participants who discovered the optimal isomorph tend to take an especially long time responding during the post-test. We contacted them afterwards about why they made decisions more slowly during the post-test and found out that instead of playing defensively using the magic square, they spent more time thinking about how to beat the computer. Thus we can infer that the bar this particular group of participants set was higher than just "not to lose." Overall, it may have been the case that participants in the post-test took a longer time because they were consulting their cheat sheets or otherwise thinking harder about their strategy, as we encouraged them to do in the strategizing session.

Another reason we did not observe a significant result of different types of isomorphs on difference in response time is that the search time for each of the visual isomorphs our subjects derived to decide the next card might vary drastically. For example, searching through a partial magic square should yield a faster decision than searching through a 9x9 matrix, while searching through a 9x9 matrix leads to a faster decision than going through all possible combinations of three numbers adding to 15. Overall, since there are many other factors involved in the difference in response time (such as search time and self-expectation of performance), we did not observe a strong causal relationship between types of isomorph and difference in response time.

7.8 A note on the variety of visual isomorphs

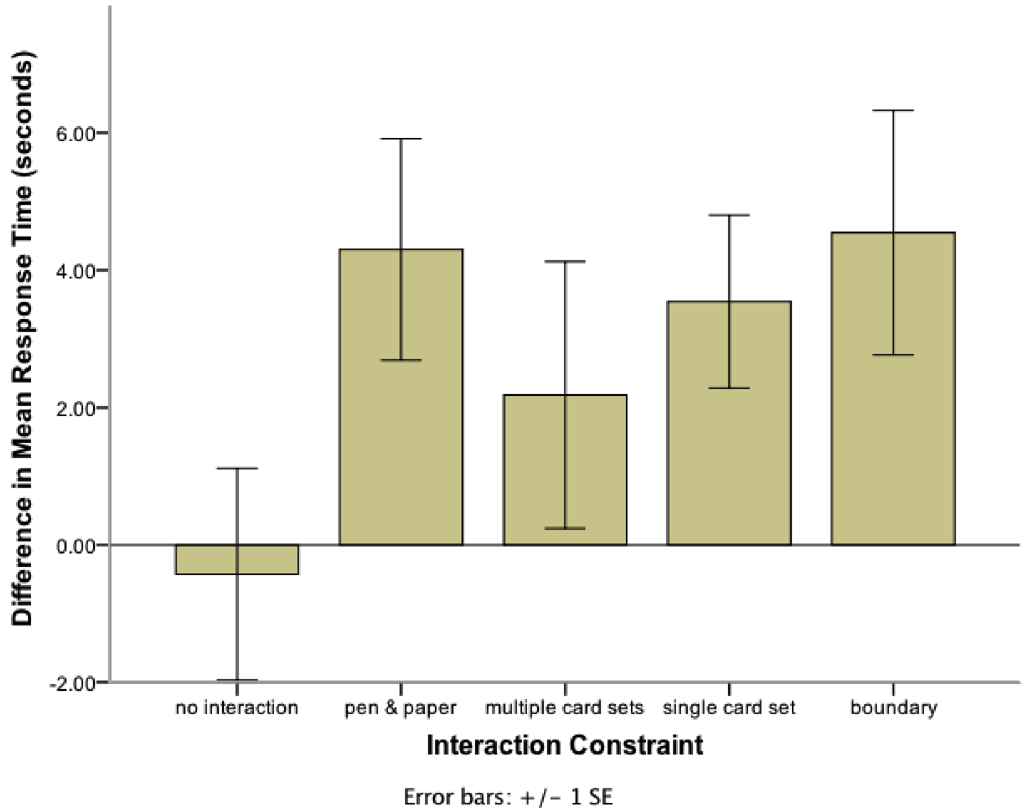
In section 7.6.1, we roughly categorized all the isomorphs our subjects developed during the study into five categories including three visual and two non-visual isomorph types. In this section we mainly focus on the visual isomorphs discovered by the participants. It is interesting to see that eight participants across interaction constraint #3 (multiple set of cards) and #5 (boundary) developed a partial magic square isomorph, and that 11 participants discovered other forms of visual isomorph across interaction constraints #1, 2, 3 and 4. Within the partial magic square isomorph, there are many variations. Figure 7.6.1(a) illustrates a few of them, and we can see that the variations are mainly caused by ordering. There are even more variations under the "Other visual isomorph" category. One type of variation was a decision tree, such as the examples in Figure7.6.1(b); additionally, a few participants built a 9x9 matrix (Figure 26). To see how this isomorph can be used in playing the Number Scrabble game, refer to Appendix A).

In Figure 22 we can see a strong contrast between the types of visual isomorphs the participants came up with. Most participants under interaction constraint #5 (boundary) developed magic square-like visual isomorphs during the strategizing session, while there are a relatively larger number of participants under both constraints #3 and #4 who discovered more creative visual isomorphs (such as different forms of decision trees and node-link diagrams). Thus, there seems to be a trade off between interaction constraint and the creativity of the resulting visual isomorph.

7.9 Summary

In this chapter, we demonstrated that constraining user interactions indeed affects problem-solving through exploring the relationship between interaction constraints, visual isomorphs, and problem-solving performance. With better isomorphs yielding higher performance, our results demonstrate that we can improve the effectiveness of problem solving activities by embodying information in user interaction.

The demonstration of the intimate relationship between user interaction constraints on problem solving has led us closer to our ultimate goal, which is to extract a user's high-level thinking process. Next we will evaluate whether such thinking process is reflected in the user interactions in a visual analytics environment.



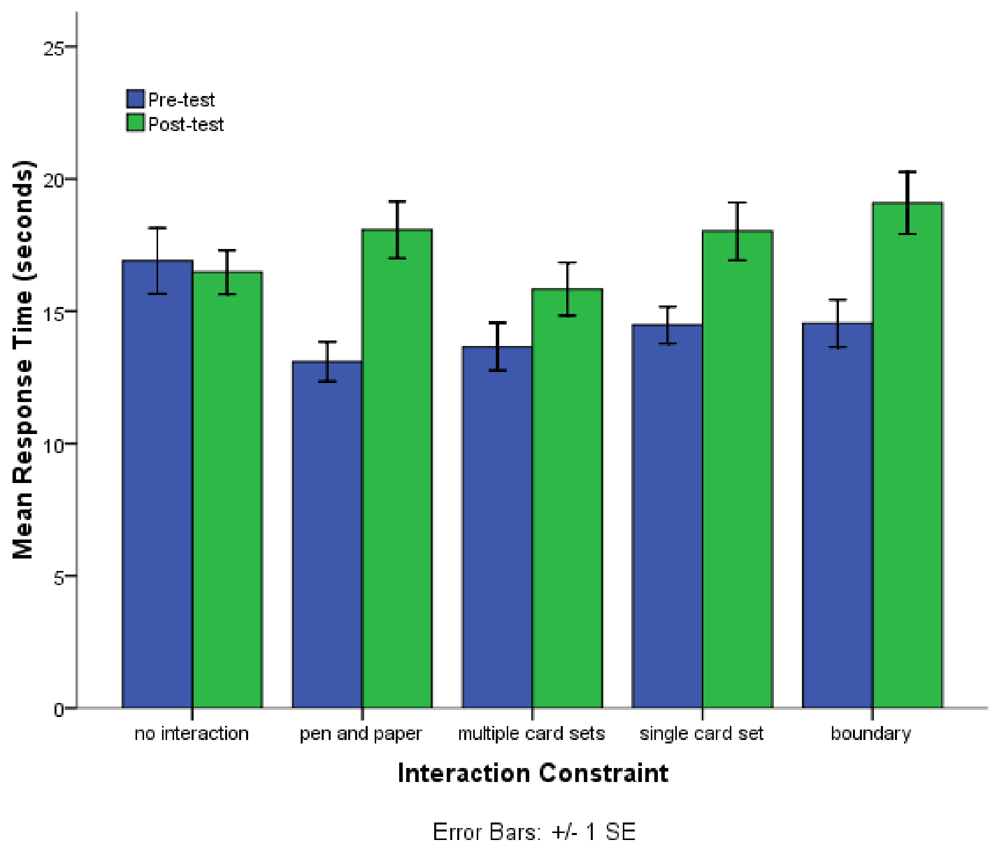
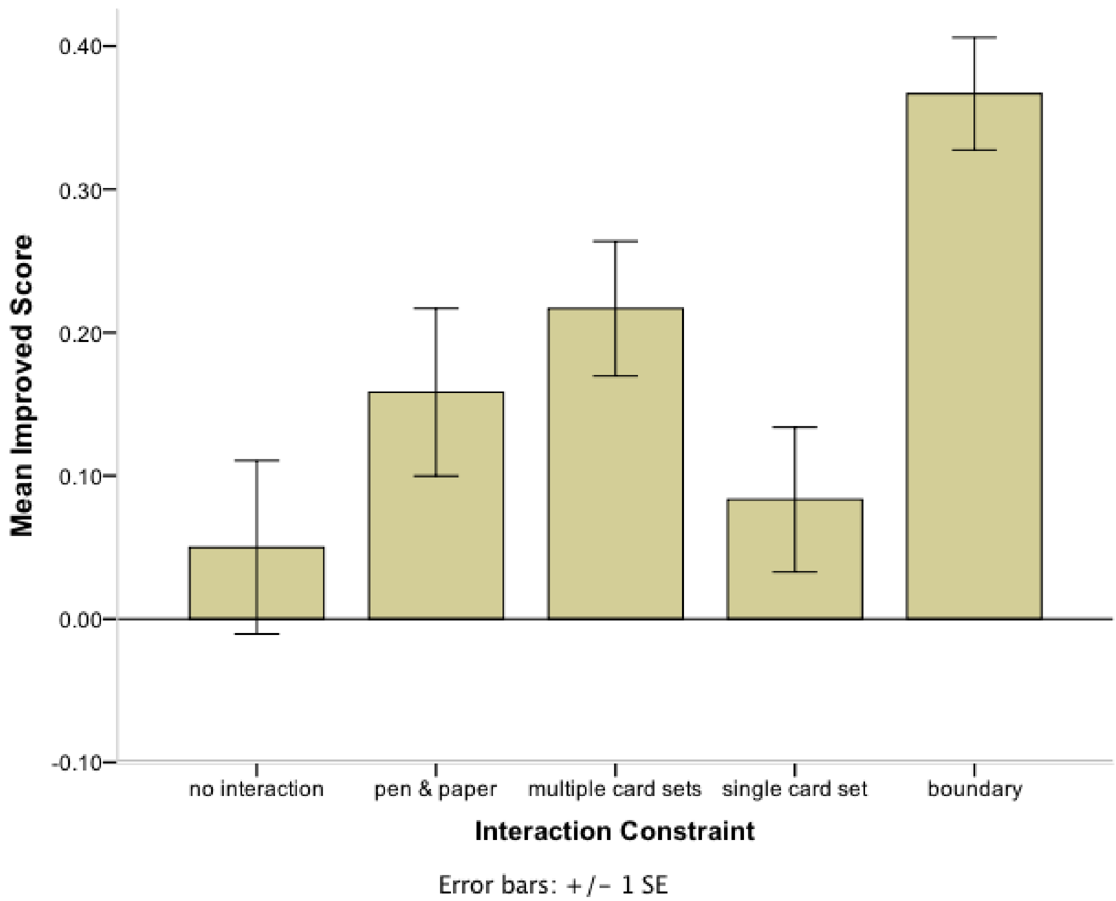


Figure 23: (top) Difference in mean response time; (bottom) mean response time(pre vs. post test)



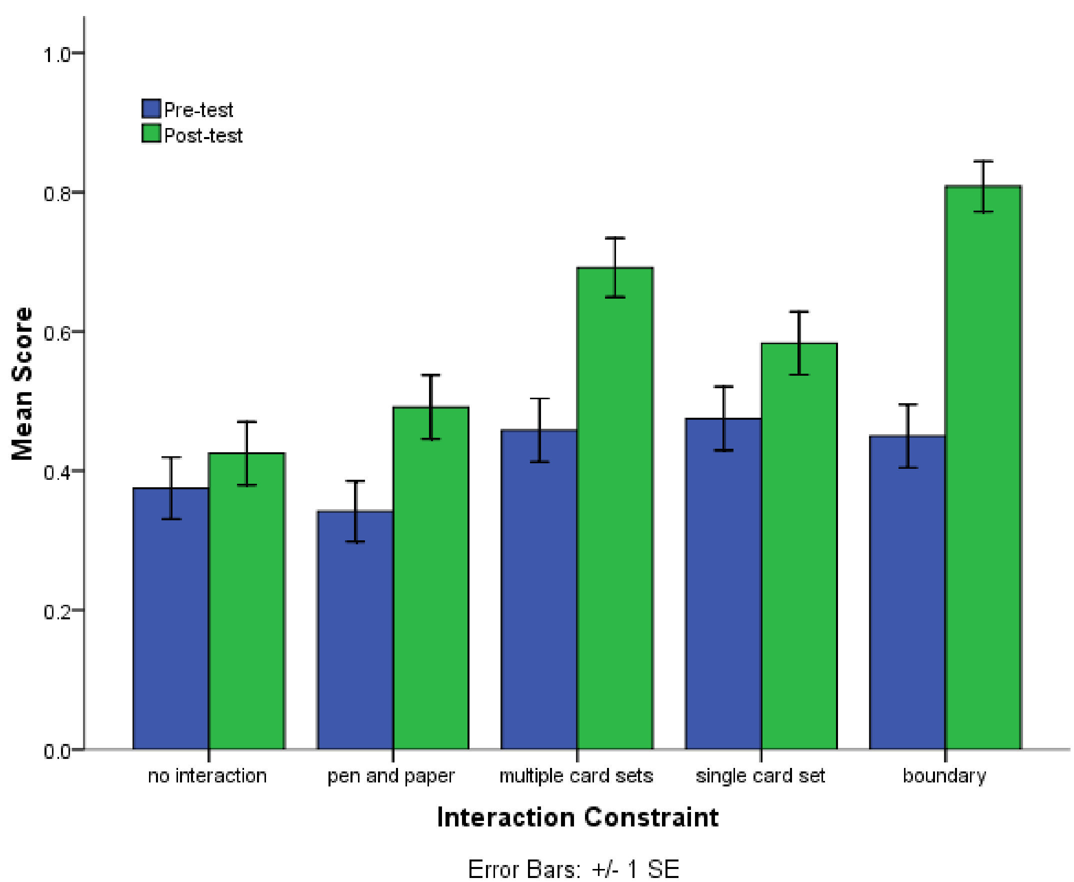


Figure 24: (top) Mean improved score; (bottom) mean score (pre vs. post test)

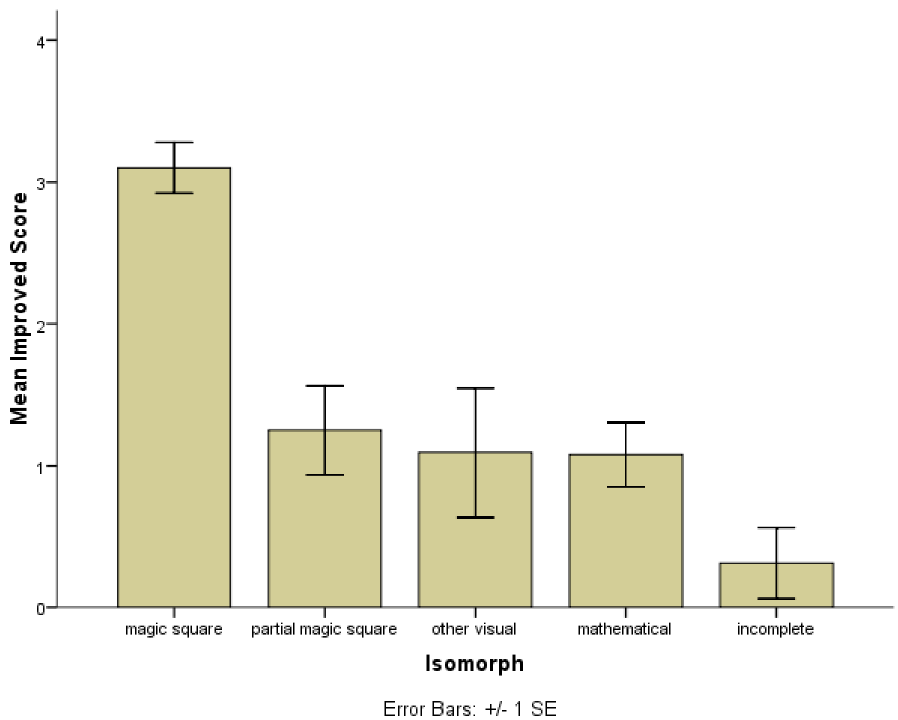


Figure 25: Mean improved score vs. Isomorph

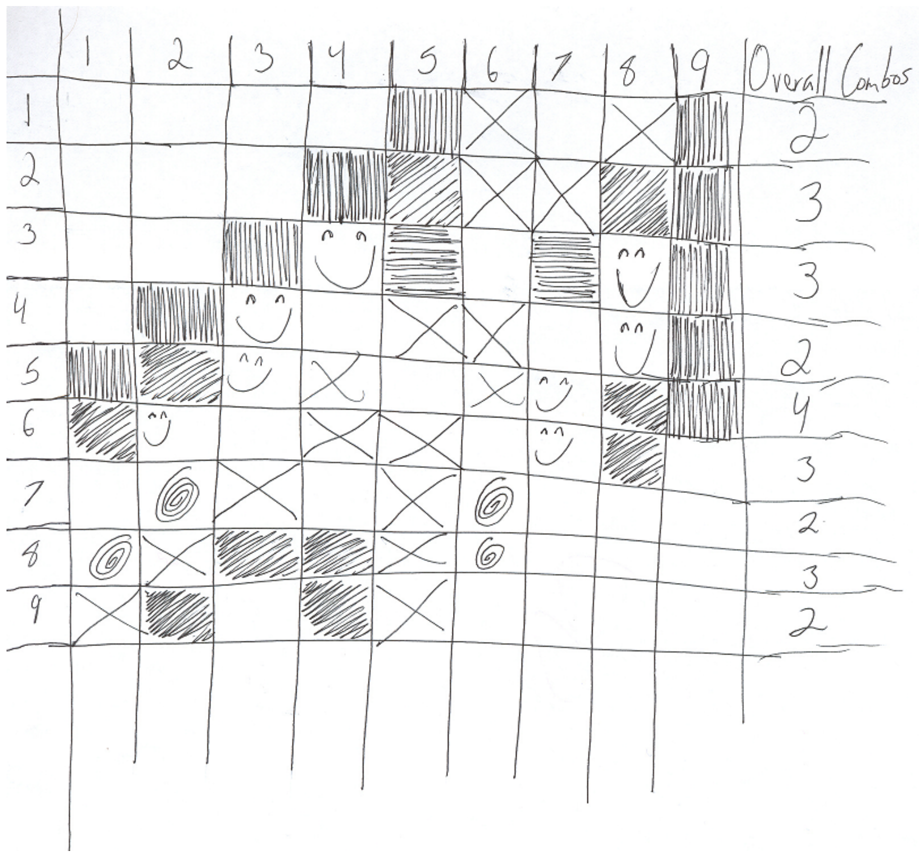


Figure 26: A matrix-like visual isomorph

CHAPTER 8: CONNECTING LOW-LEVEL USER INTERACTIONS WITH HIGH-LEVEL ANALYTICAL PROCESS

In this chapter we analyze whether we can extract user's analysis processes through analyzing interaction logs in the setting of financial visualization. We demonstrate that we were able to identify several of the strategies, methods, and findings of an analysis process using a financial visual analytical tool through the examination of an analyst's interaction log. In our study, we recorded the interactions and think-alouds of 10 financial analysts in a fraud detection task. By examining their interaction logs, we are able to quantitatively show that 60% of strategies, 60% of methods, and 79% of findings could be recovered through the use of two visual analytic log analysis tools.

8.1 Problem statement

In the short number of years since the establishment of the visual analytics research agenda, visual analytical tools have already made an impact in the intelligence and analysis communities. However, until recently, most of the research in visual analytics has focused on the techniques and methods for refining these tools, with the emphasis on empowering the analysts to make discoveries faster and more accurately. While this emphasis is relevant and necessary, we propose that it is not always the end product that matters. Instead, we argue that the process in which an analyst takes to arrive at the conclusion is just as important as the discoveries themselves. By understanding *how* an analyst performs a successful investigation, we will finally be able to start bridging the gap between the art of analysis and the science of analytics.

Unfortunately, understanding an analyst's reasoning process is not a trivial task, especially since most researchers rarely have access to analysts performing analytical tasks using classified or highly confidential material. While there has been a recent increase of activity in the visual analytics community to help analysts document and communicate their reasoning process during an investigation, there is still no clear method for capturing the reasoning processes with minimal cognitive effort from the analyst. This raises the question we look to address in this chapter: how much can an analyst's strategy, methods, and findings using a visual analytical tool be recovered?

It is our hypothesis that when interacting with a well-designed visual analytical tool, a large amount of an analyst's reasoning process is embedded within his interactions with the tool itself. Therefore, through careful examination of the analyst's interaction logs, we propose that we should be able to retrieve a great deal of the analyst's reasoning process. To validate our hypothesis, we designed a study to quantitatively measure if an analyst's strategies, methods, and findings can be recovered through human examination of his interaction logs. Our study consists of four stages: user observation, transcribing, coding, and grading. In the user observation stage, we invited 10 financial analysts to use a financial visual analytical tool called WireVis [22] to identify potentially fraudulent wire transactions within a synthetic dataset in think-aloud sessions. The analysts' interactions were logged into file and at the same time their think-alouds captured on video and audio. This information was transcribed by the authors later into files that collectively were considered to be representative of the analysts' reasoning processes and used as the "ground truth" for the study.

Four coders who are students familiar with the WireVis tool examined each analyst's interaction log using two log analysis tools (Operation and Strategic Analysis tools) that we developed [63]. Through visual inspection and analysis of each analyst's interaction log, the four coders were asked to annotate what they believed the analysts' strategies, methods, and findings were. We then compared the coders' inferences with the ground truth, and the result became the basis of our claim on the types and amount of an analyst's reasoning process that were recoverable through the examination of interaction logs.

The result of our study has been most encouraging. Aside from a few specific, low-level types of findings, the four coders (who are not trained in financial fraud detection) were able to correctly retrieve 60% of the analysts' strategies, 60% of the methods, and 79% of the findings. This result indicates that some of an analyst's strategies, methods, and findings in using a visual analytical tool are indeed recoverable through human examination of an interaction log. It is relevant to note that the extracted reasoning process is solely based on the analyst's activities within a visual analytical tool and does not include the overall intelligence analysis that often involves multiple tasks and tools such as searching through websites, phone discussions, the use of additional software, etc. However, our findings represent an important aspect of the intelligence analysis, and provide an example for visual analytics as a community to uncover a new path towards better understanding and capturing of an analyst's reasoning processes.

8.2 Related Work

We roughly categorize the current research in visualization and visual analytics for capturing the reasoning process of an analyst into two groups: capturing the user's interactions and interactive construction of the reasoning process using a visual tool.

8.2.1 Capturing User Interactions

Capturing user interactions for the purpose of understanding the user's behavior is very common both in academics and industry. Commercially, there are many off-the-shelf applications that range from capturing a user's desktop activities such as usability software to interactions on a website (which is a common feature in most web servers).

In the field of visualization, one of the most notable systems for capturing and analyzing user activities is the GlassBox system by Greitzer at the Pacific Northwest National Laboratory [45]. The primary goal of the GlassBox is to capture, archive, and retrieve user interactions [29]. However, it has also been shown to be an effective tool for capturing specific types of interactions for the purpose of intelligence analysis [30]. While GlassBox and most usability software are effective tools for capturing user activities, they focus primarily on low level events (such as copy, paste, a mouse click, window activation, etc), whereas the events captured in our system are at a higher level that corresponds directly to the data (such as what transaction the user clicked on). For more information on the differences in these two approaches, see the work by Jeong et al. [63] or work by Heer et al. [53].

More recently, Jankun-Kelly et al. [60] proposed a comprehensive model for capturing user interactions within a visualization tool. Their work is unique in that they focus on capturing the effects of the interactions on the parameters of a visualization. Although it is unclear how this framework supports higher level event capturing, the direction is interesting and could lead to a more uniform way of capturing user interactions.

The systems and approaches above are all proven to be innovative and effective. However, their objectives differ from our goal in that none of these systems fully addressed our question of how much reasoning process can be recovered through the examination of interaction logs. It is with this question in mind that we expand on this area of research to capturing user interactions and look to extract reasoning processes embedded in them.

8.2.2 Interactive Construction of the Reasoning Process

An alternative approach to retrieving reasoning through interactions is for the analyst to create a representation of the reasoning process (usually in the form of a node-link diagram) while solving a complex task. There are a few recent systems in this domain, most notably the Aruvi framework by Shrinivasan and van Wijk [107], which contains three main views, data view, navigation view, and knowledge view. Data view is the visual analytical tool itself, navigation view is a panel for visually tracking the user's history, and lastly the knowledge view allows the user to interactively record his reasoning process through the creation of a node-link diagram.

Similar to the Aruvi framework, the Scalable Reasoning System (SRS) by Pike et al. [92] allows its users to record their reasoning processes through the creation of node-link diagrams. However, unlike the Aruvi framework, the SRS focuses on the collaborative aspects of organizing the reasoning processes among multiple users and sharing their results across the web.

Most recently, Heer et al. [53] created a tool for visualizing users' histories within the commercial visualization tool Tableau [82]. Although the emphasis of this work is not on constructing or visualizing the reasoning process, the functionalities within the tool that allows for a user to edit and modify his interaction history could be used towards communicating his reasoning process effectively.

While there has not been a formal comparison between interactively constructing the reasoning process as mentioned above and our method of analyzing interaction logs, we hypothesize that the cognitive load of having to perform analytical tasks while maintaining and updating a representation of the reasoning process could be tiring [44]. We believe that the systems mentioned above will have better representations of the user's reasoning process. However, we argue that a transparent, post-analysis approach offers an alternative that can achieve comparable results without the efforts from the analysts. Most likely the best solution is somewhere in between, and we look forward to analyzing the pros and cons of the two approaches.

8.3 WireVis Interactions

We conducted our study with a particular visual analytical tool for investigating financial fraud called WireVis that logged all user interactions. We also developed two additional tools for visualizing user interactions within WireVis to help us explore the analyst's activities and reasoning process [63]. We first describe all of these tools

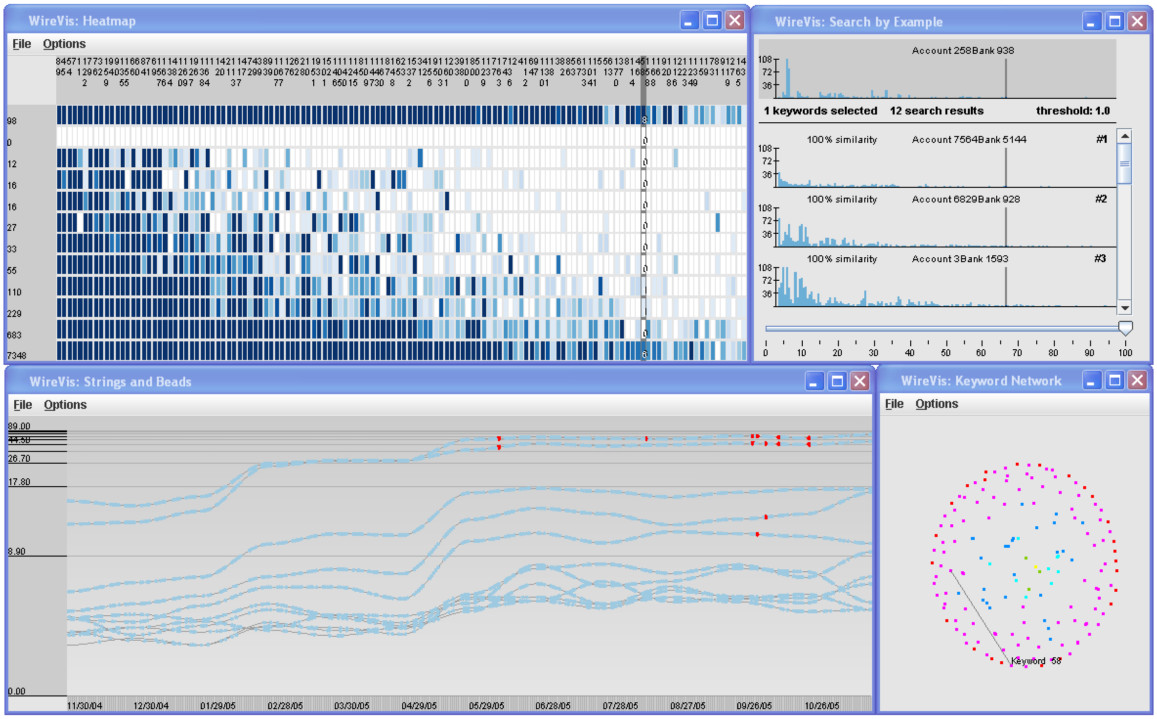


Figure 27: Overview of WireVis. It consists of four views including a heatmap view (top left), a time-series view (bottom left), a search by example view (top right), and a keyword relation view (bottom right).

before presenting the details of the user study in the next section.

WireVis is a hierarchical, interactive visual analytical tool with multiple coordinated views [22]. This visual analytical tool was developed jointly with wire analysts at Bank of America for discovering suspicious wire transactions. It is currently installed at Bank of America's wire monitoring group, WireWatch, for beta testing. Although it has not been officially deployed, WireVis has already shown capabilities in revealing aspects of wire activities that analysts were not previously capable of analyzing. Through a multiview approach, WireVis depicts the relationships among accounts, time and transaction keywords within wire transactions (see Figure 27).

8.3.1 Synthetic Data with Embedded Threat Scenarios

To preserve the privacy of Bank of America and their individual account holders, we created a synthetic dataset for the purpose of this study. Although none of the transactions in the dataset are real, we captured as many characteristics and statistics from real financial transactions as we could and modeled the synthetic data as closely to the real one as possible. The approach we took is loosely based on the methods described by Whiting et al. in developing the 2006 VAST contest [13].

For the purpose of the user experiment, it is important that the dataset is simple enough that users are able to look for suspicious transactions within the time frame of a study, but is complex enough that interesting and complicated patterns can be found. The synthetic dataset therefore contains 300 financial transactions involving approximately 180 accounts. Twenty-nine keywords are used to characterize these transactions, with some of them representing geographical locales (such as Mexico, Canada), and some representing goods and services (such as Minerals, Electronics, Insurance, Transportation, etc.) Each record of a transaction consists of the transferred amount, the sender and receivers names, date, and one or more keywords relating to the transaction. Four different types of known threat scenarios were identified. Two cases of each of the four types were created and embedded into the synthetic dataset based on the approach proposed by Whiting et al. [14]:

Incompatible Keywords in a Transaction: Transactions with two or more keywords that do not belong together. For example, a transaction containing the keywords “car parts” and “baby food”.

Accounts with Dual Roles: An account that has had transactions of different incompatible keywords is questionable. For example, an account that transacts on “gems” at one time and “pharmaceuticals” at another.

Keywords with Large Amounts: Transactions of certain keywords are expected to have corresponding dollar amounts. For example, a transactions from a local store on “arts and crafts” should not be in the millions.

Change in Amounts Over Time: An account with an established temporal and amounts pattern receiving a large sum outside of its norm should be examined further. For example, an account with a steady deposit of paychecks of fixed amounts on regular intervals receiving a transaction of a large amount.

8.3.2 Operation Analysis Tool

Our operation analysis tool is designed to support the analysis of a participant’s operational interactions in relation to his annotations (Figure 28). The tool is implemented in openGL, and is fully zoomable and pannable and supports selections of interaction elements for detailed inspection. The x-axis of the main view represents time, with a striped background indicating the length of a fixed time duration (defaulted to 60 seconds per strip). The y-axis is divided into 5 sections, with each section supporting one aspect of the participant’s investigation process. Figure 28 (A)-(E) show the 5 perspectives, which are the participant’s annotations (A), the participant’s interactions with the three views in WireVis (B), the depths of a participant’s investigation (C), the areas of the participant’s investigation (D), and the time range of the investigation (E). The sliders in (F) allow the user to scale time,

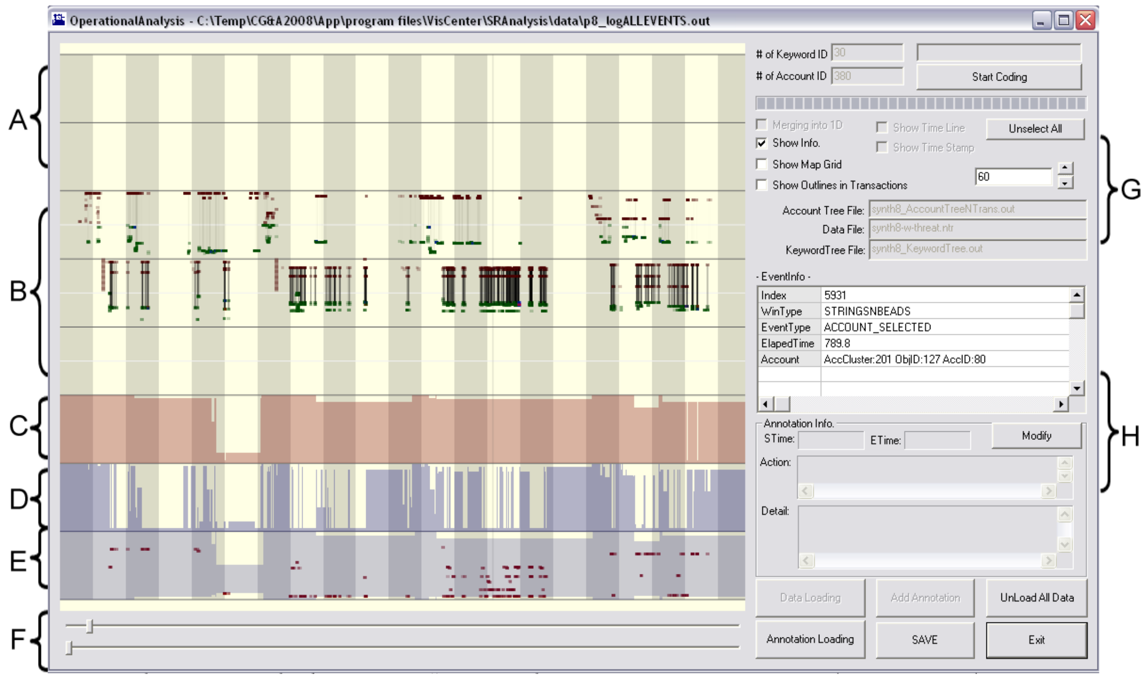


Figure 28: Overview of the Operation Analysis Tool. (A) shows the potential area for adding annotations. (B) shows the participant's interactions with the three views in WireVis (the three rows from top row to bottom correspond to heatmap, time-series, and search by example views respectively). (C) represents the depths of a participant's investigation. (D) shows the areas of the participant's investigation, and (E) the time range. Sliders in (F) control the time scale, while checkboxes in (G) change various visualization parameters. (H) shows the detail information of a participant's selected interaction element.

while checkboxes in (G) control various visualization parameters. The detail view in (H) depicts detailed information of a specific user-interaction element.

Annotation View The results of our participants' think-aloud during the experiment are recorded into separate files. These annotations to the participant's investigation are shown in this view. As can be seen, our participants often exhibit a hierarchical structure in their reasoning process, with the highest level of reasoning depicting strategies they employ such as "seek all keywords related to the keyword food" The lower levels depict specific operations to execute those strategies, ranging from "search for keywords other than food relating to account 154" to "identify the

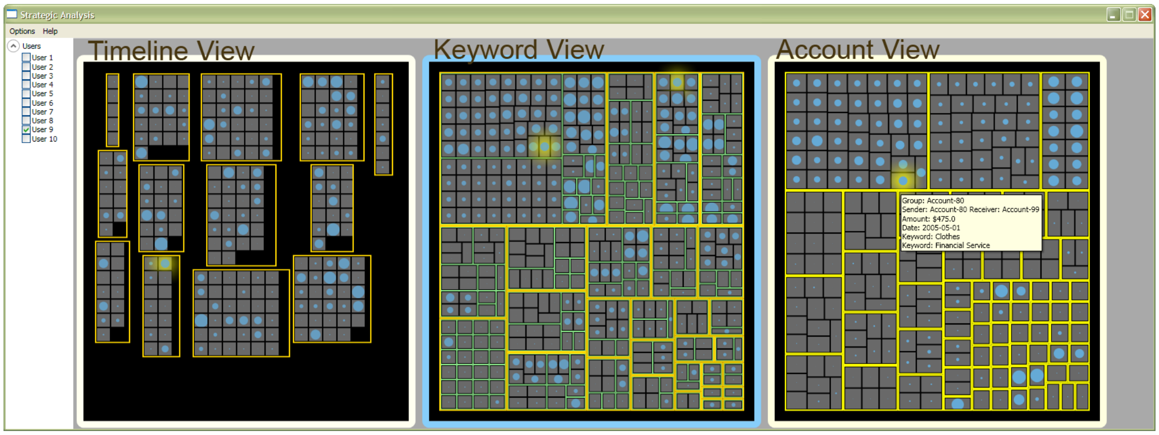


Figure 29: Strategy Analysis Tool. The left view shows transactions grouped by time, middle view shows grouping by keywords, and the right view shows grouping by accounts. The patterns in the account view indicate that the primary strategy employed by this participant was to examine two specific accounts (located on the top of the Account View).

receiver of a transaction (account 64) of account 154." The hierarchical nature of the participants' reasoning are represented in the annotation view, with the higher level annotations shown above the lower ones as interactive floating text boxes [7]. The time range of each annotation is drawn as nested boxes using different colors. The user can select any particular annotation, and its corresponding time range is highlighted across all the other views (Figure 28).

WireVis Interaction View WireVis uses multiple coordinated views to visualize different relationships within the data. In the WireVis Interaction view, we look to display the participant's usage pattern of the WireVis tool. The three rows in this view correspond to the three main interactive views in WireVis: Heatmap, Strings and Beads, and Search by Example. In each view, we can choose two different attributes of the participant's selection. In Figure 28 (B), the two attributes are keywords (shown as red dots) and accounts (shown as green dots). On first glance, it is easy to see which views in WireVis the participant interacts with over time. On closer inspection, the distribution of the red (keywords) and green dots (account) also reveal high-level patterns in the investigation. Scattered red dots could indicate an exploration of keywords, whereas concentrated green dots (e.g., if the green dots are aligned horizontally) could reveal the participant's interest in a specific account. When both red and green dots appear together and are connected by a line, it denotes that the participant is investigating the relationship between the two (such as a cell in the Heatmap view in WireVis).

Depth View On top of visualizing a participant's direct interactions with WireVis, it is also important to see some semantic information regarding the participant's investigation process. In this view, we visualize the "depth" of a participant's investigation by displaying the number of visible transactions in WireVis. For example, when the participant is looking at the overview in WireVis, our Depth view will be completely filled, indicating that all transactions are visible. As the participant zooms in to investigate specific keywords or accounts, the Depth view will show a drop in visible transactions (Figure 28 (C)).

The Depth view also indicates when a participant requests detailed information for a specific account or transaction (such as double-clicking on a bead in the Strings and Beads view). These interactions show up as a vertical line, which is easily distinguishable from a participant's operations for zooming in or focusing on a specific area in the data.

Areas View While the Depth view shows the number of visible transactions in WireVis, it is also relevant to indicate interactions that highlight areas that the user has shown interests. These interactions are commonly used in WireVis through the "mousingover" operation. As the participant mouses-over keywords, accounts, or transactions in WireVis, the system displays information about the highlighted data without requiring the user to change the zoom level or focus. Using the mouse-over operation in WireVis is common, and often indicates an exploration process in which the participant is looking for suspicious areas for further investigation. In the Areas view, a high variation in a short amount of time could indicate such an exploration process, while a more leveled section suggests that the participant is investigating specific activities (Figure 28 (D)).

Time Range View Time is an important aspect in discovering financial fraud, and WireVis provides views to explore the temporal dimension. In the Time Range view, we look to capture the participant's time-based investigation. The y-axis of the Time Range view denotes the dates represented in the data from more recent to least. A fully colored section indicates that the participant's investigation spans the entire time range, whereas a change would denote that the participant has zoomed in to a specific time period (Figure 28) (E). The dots in the Time Range view indicate selections of transactions of a specific date. In WireVis, this is done by either mousingover or double-clicking on a bead in the Strings and Beads view. A high concentration of the appearance of these dots often suggests that the participant has found some specific transactions and is looking to find out the details of these transactions.

8.3.3 Strategy Analysis Tool

As opposed to operation, strategy is a long term plan of action designed to achieve a particular goal. As shown in the Annotation view of the operation analysis tool (section 8.3.2), most of our participants exhibit both strategic and operational reasoning when investigating fraud. So besides addressing the question "what do the participants actually do" using our operation analysis approach, we also look to investigate the high level strategies that the participants employ while approaching the tasks. Through the use of our strategy analysis tool, we can identify each participant's areas of interest as well as comparing different participants' strategies.

We adopt treemap as the basis of our strategy analysis tool. The treemap visualization allows us to investigate similarities between our participants' strategies without considering the flow or speed in which our participants execute their strategies. Our participants had varied preconceived knowledge about the keywords and their meanings, and therefore approached the investigation tasks differently. Many of them identified the same embedded fraud scenarios, but none of them shared the same path in discovering these activities. Using our modified treemap visualization, we can identify the participants' strategies without regard to the paths they have chosen.

The initial layout of the strategy analysis tool shows three different treemap views classifying the transaction data based on three attributes: time, keywords, and accounts (Figure 29). The three views are coordinated such that highlighting a transaction in one view also highlights the same transaction in the other two views. We choose transactions to represent the lowest level of the treemaps because they represent the lowest granularity of the data. A colored circle is displayed on each cell, and the size depicts the amount of time the participant's investigation has included that transaction. When comparing two participants or two groups of participants, the color of the circle indicates which of the two participants spent more time on the transaction (Figure 6).

Timeline View Transactions in this view are classified based by their date. Each grouping contains transactions of the same month. As shown in Figure 3, the transactions in our synthetic data set span a 13 month period. Note that the participant depicted in this view did not perform his investigation based on the transaction dates as the circles appear fairly evenly through all 13 months.

Keyword View This view applies two different classification criteria. On the top level, transactions are grouped based on keywords (shown as yellow cells in Figure 29). Each cell is then further subdivided by individual accounts (shown as green cells).

Since a transaction often contains multiple keywords, the same transaction could appear in more than one keyword cells. Similarly, every transaction contains two accounts, a sender and a receiver, so a transaction will always appear at least twice, once for each account. Due to these two reasons, the total number of transaction cells in this view are greater than those in the Time view. However, we find this layout more intuitive for understanding a user's strategy involving keywords. For example, in Figure 3, it is easy to see that the participant focused on a few specific keywords, but even more specifically on a few accounts relating to those keywords.

Account View The Account View orders the transactions based on their corresponding sending accounts. As shown in Figure 29, this view makes clear that this participant's strategy in discovering financial fraud using WireVis is almost entirely based on the detailed investigation of one or two accounts. Time and keywords appear to be secondary considerations during his investigation.

8.4 Evaluation

We conducted a user study to determine how much of an analyst's reasoning process can be recovered using just the captured user interactions. We evaluated this recovery in a quantitative fashion by comparing the process that was inferred by a set of coders against the ground truth determined from videos of the exploration process.

Four stages are designed as user observation, transcribing, coding, and grading. The comprehensive information of each stage is provided in the following subsections.

8.4.1 User Observation

In order to understand the user's reasoning process through his interactions, we first conducted a qualitative, observational study of users analyzing data with WireVis. We recruited 10 financial analysts with an average of 9.9 years (and a median of 8 years) of financial analysis experience who all worked in large financial firms in our area. All of the participants were either currently working as a financial analyst or had professional financial analyst experience. Eight of the users were professionally trained to analyze data for the purpose of fraud detection. Of the 10 analysts, six analysts were male and four were female.

To preserve the privacy of Bank of America and their individual account holders, we created a synthetic dataset for the purpose of this study. Although none of the transactions in the dataset are real, we captured as many characteristics and statistics from real financial transactions as we could and modeled the synthetic data as closely to the real one as possible. The dataset was designed to be simple enough that users were able to look for suspicious transactions within the time frame of a study. This dataset contained 300 financial transactions, with 29 keywords. Some keywords were the names of countries, such as Mexico, and others were goods or services, such as Software or Raw Minerals. We also developed four threat scenarios and injected a total of nine cases we deemed suspicious into the dataset. The threat scenarios included transactions in which keywords should not appear together, accounts with dual roles, keywords with unusually high transaction amounts, and accounts with suspicious transactional patterns appearing over time. More details of the synthetic dataset and sample threat scenarios can be found in [63].

At the beginning of the study session, each participant was asked to fill out a demographic form and was then trained on the use of WireVis for approximately 10 minutes. The participant was also provided a one-page overview of the functionality of WireVis and encouraged to ask questions. Following the training, the user was asked to spend 20 minutes using WireVis to look through the dataset to find suspicious activities. We asked the participant to think-aloud to reveal his strategies. We specifically encouraged the participant to describe the steps he was taking, as well as the information used to locate the suspicious activities. Once the user drilled down to a specific transaction, he was asked to write it down on a Discovery Sheet for the purpose of recording and reporting his findings. Once the user documented a specific transaction, he was encouraged to continue looking for others until the time limit was reached. After the exploration process, a post-session interview was conducted for the participant to describe his strategies and additional findings.

Several methods were used to capture each participant's session as thoroughly as possible. Commercial usability software was used to capture the screen. A separate microphone was used to record the user's audio during the session. Lastly, functions built into the WireVis system captured the user's interaction with the tool itself as information relevant only to the WireVis system. Instead of recording every mouse movement or keystroke, WireVis captures events that generate a visual change in the system. For example, a mouse movement that results in highlighting a keyword in the Heatmap view will generate a time-stamped event noting that the user has highlighted a specific keyword.

8.4.2 Transcribing

The video and the think-aloud of each participant were used to create a detailed textual timeline of what each participant did during their session, along with the participant's self-reported reasoning and thinking process. While the created textual timeline is an interpretation and might not perfectly reflect the (internal) reasoning process of the participant, it was created based on the facts recovered from video and audio with conscious efforts in minimizing human bias. We therefore consider the resulting transcript to represent the "ground truth" of what each participant did during their analysis with WireVis.

During the transcribing stage, different strategies, methods, and findings in investigating fraudulent activities were identified to serve the grading process later. Specifically, we identified the following in the transcript:

- A *"Finding"* represents a decision that an analyst made after a discovery.
- *"Strategy"* is used to describe the means that the analyst employed in order to arrive at the finding.
- Also, the link between "finding" and "strategy" is captured by *"method"* which focuses on what steps the analyst adopted to implement the strategy for discovering the finding.

In a typical investigation, an analyst's *strategy* might be to search for a specific suspicious keyword combination based on his domain knowledge. For example, the analyst might determine accounts and transactions involving both the keywords Mexico and Pharmaceutical to be potentially suspicious. Using this strategy, the *methods* employed by this analyst could then be comprised of a series of actions such as highlighting or filtering those keywords, and drilling down to specific accounts and transactions. At the end of the investigation, the analyst would record his *findings* based on the encountered account numbers and transaction IDs along with their decision about whether the particular finding is suspicious or not.

8.4.3 Coding of the interaction logs through visual examination

We asked several people familiar with WireVis to view each participants' interactions and determine their reasoning. Specifically, we recruited four "coders" from our university, all of whom were familiar with WireVis (three male, one female). They then used the two interaction log analysis tools (Operation and Strategic Analysis tools) to view participant interactions, and created an outline of what occurred.

We first gave all coders comprehensive training on how to use the Operation Analysis Tool and Strategic Analysis Tool to examine the interaction logs of each analyst's investigations. We also provided a guideline of hierarchical coding procedures, asking coders to, in free-text format, provide hierarchical annotations within the visual analytical tools. The hierarchies are reflected as different levels of decision points and strategies extracted by the coders. We asked coders to identify and label findings, strategies, and methods for each analyst. In addition, coders were encouraged to annotate on the transitions if they could discover relationships between each decision point such as one strategy leads to multiple findings or one finding transforms to a new strategy.

All findings from the coders were recorded as annotations and linked to corresponding interaction events and time range. Each coder went through the 10 analysts'interaction logs one by one using the visual analytical tools, spending an average of 13.15 minutes reconstructing each analyst's reasoning process. Thus, at the end of the coding phase, we collected 10 sets of annotations from each coder, resulting in 40 sets of annotations overall.

8.4.4 Grading

We then compared the annotations the coders produced to the "ground truth" to determine how much of the reasoning process was able to be reconstructed by the coders. The comparisons are graded according to a set of pre-determined criteria by one of the researchers, which we describe below.

The categories we used in the grading were in accordance with both transcribing and coding: finding, strategy and method. Generally speaking, "strategy" and "finding" do not necessarily have a one-to-one mapping relationship since some strategies may lead to multiple or null findings. But one "finding" always comes with a "method" in the sense that a method is always needed to make a decision.

For each finding, strategy, and method, we graded according to the following criteria: "Correctly Identified", "Incorrectly Identified", "False Detections" and "Never Identified". This combination was chosen because the four measurements covered all possible scenarios and yet were explicitly distinguishable. "Incorrectly Identified" indicated that a coder noticed some meaningful interactions but incorrectly interpreted them, while "False Detections" captured the scenarios in which a coder thought that certain action took place but in fact there was none. "Never Identified" involved actions that took place, but were not noticed or annotated by the coders.

Figure 30 illustrates the overall criteria used for grading. We determined that a "finding" was correct as long as the coders correctly identified there was a decision made during the analyst's investigation. But we did not ask them to determine what the outcome of that decision was (whether the certain transaction is suspicious, not suspicious or inconclusive). Additionally, if only a part of the coder's annotation was correct, for example if he determined a "strategy" was looking for five incompatible keywords but only identified four keywords correctly, we graded that annotation as "Incorrectly Identified". This purpose for such a strict grading criteria is to minimize potential bias in the grading process.

		Ground Truth		C1	C2	C3	C4
P1	Finding	6	Correctly Identified	5	3	4	5
			Incorrectly Identified	0	1	0	1
			False Detections	0	0	2	0
			Never Identified	1	2	2	0
	Strategy	3	Correctly Identified	3	3	1	0
			Incorrectly Identified	0	0	1	3
			False Detections	0	0	1	1
			Never Identified	0	0	1	0
	Method	6	Correctly Identified	6	4	3	3
			Incorrectly Identified	0	0	1	2
			False Detections	0	0	0	0
			Never Identified	0	2	2	1
	Time Spent (min)			14.7	33.39	10.27	35.9

Figure 30: Grading results of participant 1. A participant's analysis process is separated into findings, strategies, and methods. This figure shows the results of four coders' annotations and how they match the participant's analysis according to the four grading criteria: correctly identified, incorrectly identified, false detections, and never identified.

8.5 Results

Both the quantitative and the observational results we obtained from grading are rich and informative. In this section, we first demonstrate quantitatively the amount of reasoning that can be extracted from analyzing interaction logs. We then describe some of the trends and limitations of the coding process using our interaction log analysis tools.

8.5.1 How much reasoning can we infer?

Figure 31 shows the average accuracy of each coder's reconstructed reasoning processes of all participants. The results are separated into three categories as described in section 8.4.2: findings, strategies and methods. The results indicate that it is indeed possible to infer reasoning from user interaction logs. In fact, on average, 79%

of the findings made during the original investigation process could be recovered by analyzing the captured user interactions. Similarly, 60% of the methods and 60% of the strategies could be extracted as well with reasonable deviation between the coders.

An interesting observation is that all coders performed better in extracting findings than strategies or methods. We will discuss a possible explanation for this phenomenon in section 8.6.

8.5.1.1 Across Participants

A different perspective from which to examine the results is to look for variations in accuracy across the 10 participants. Figure 32 shows the average accuracy of the coders in recovering the reasoning processes behind the 10 participants. This

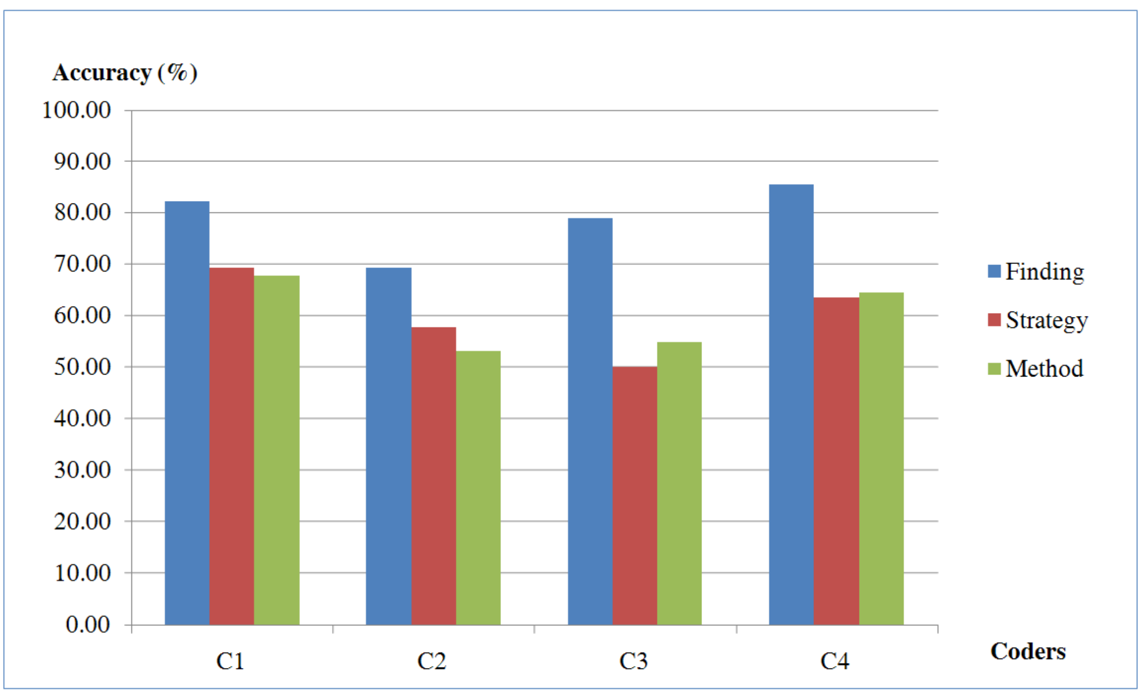


Figure 31: The average accuracy of the four coders correctly identifying “findings”, “strategies” and “methods” of all ten participants.

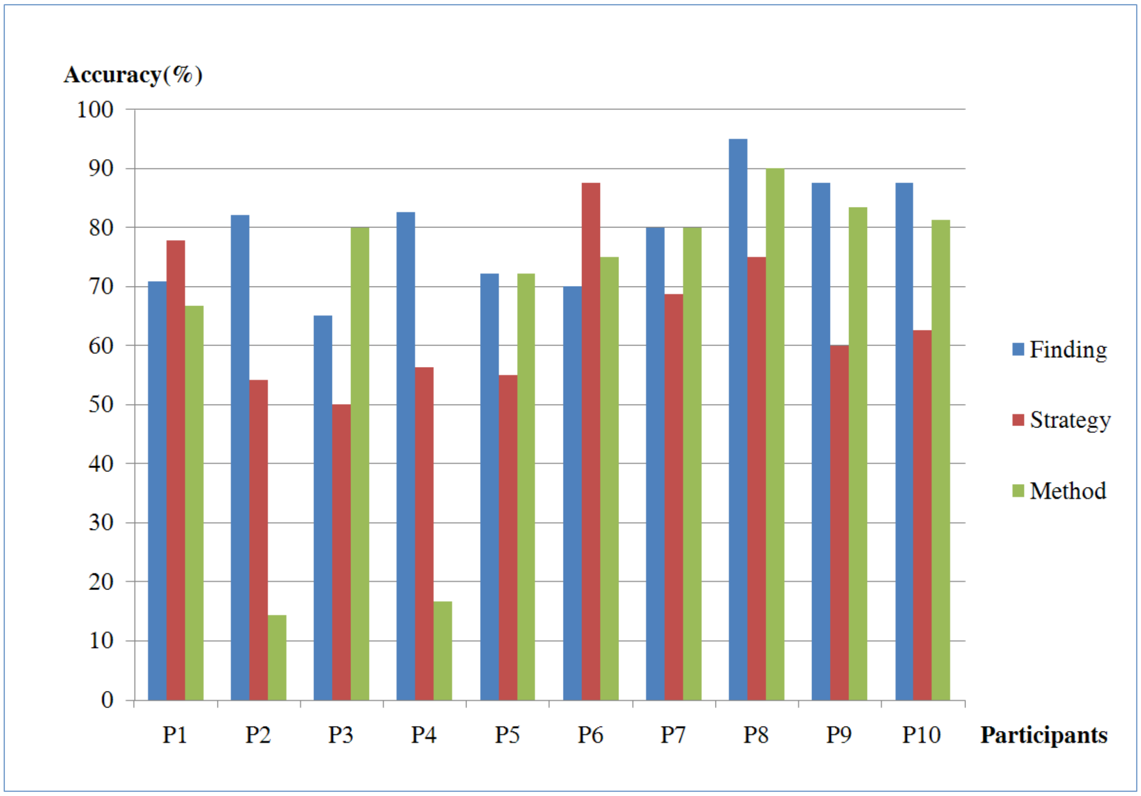


Figure 32: The average accuracy of correctly identifying “findings”, “strategies”, and “methods” based on the 10 participants.

result indicates that there is a noticeable difference between accuracies in extracting reasoning processes for different participants. This finding leads to the conclusion that there are some analysis processes that are more difficult to follow than others. Although there is no definitive answer to why this is, our own investigation suggests that there are two plausible contributors. The first is the difference in experience in financial fraud detection between our participants and our coders. Since our coders have no training in fraud detection, it is natural that some of the strategies and methods in investigative processes are lost to them.

Another cause of this variation is manifested in the acute drop in the accuracy when extracting “methods” from P2 and P4’s analysis as shown in Figure 32. As

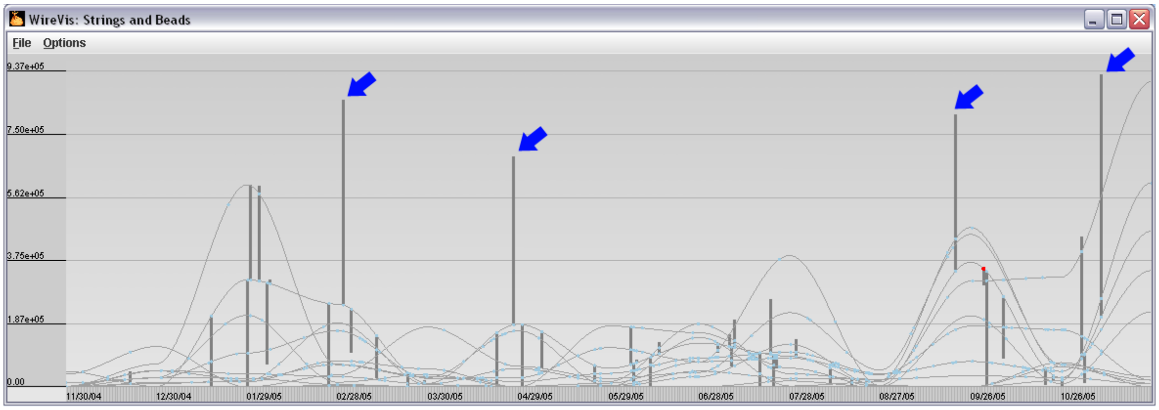


Figure 33: The time-series view in WireVis showing spikes that indicate sudden increases in the amounts or frequencies of wire transactions.

the figure suggests, the coders were baffled by the methods of these two participants. Upon investigation in the video of the participant's analysis process, we discovered that participants 2 and 4 focused their analysis on the irregularities in the time-series view in WireVis. Specifically, they closely examined "spikes" in the view (Figure 33) which indicate sudden increases in amounts or frequencies of wire transactions. Our coders had no way of seeing these visual patterns, so they were not able to identify the methods behind the participants' analyses.

8.5.1.2 Considering False Detections

Since the purpose of this study is to figure out *how much* of the reasoning process can be extracted from interaction logs, we have reported the accuracy based purely on the number of "correctly identified" elements. However, it is relevant to make note of the number of times that our coders made detections that turn out to be inaccurate. Under our grading scheme, the number of annotations made by a coder often exceeds the number of elements in the transcription due to the false detections. For example, the grading result of participant 1 in Figure 30 shows that the number of "findings"

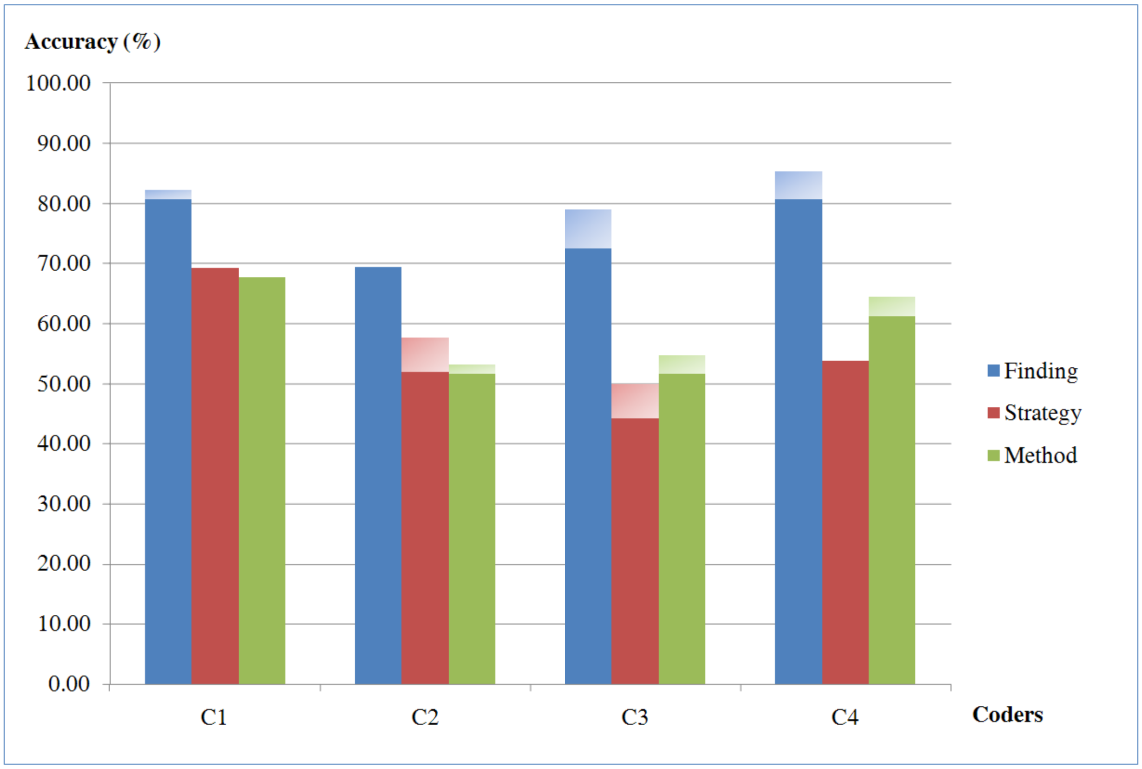


Figure 34: The accuracy of the coder's annotations in matching up to the 'findings", "strategies", and "methods" of the analyses. The semi-transparent areas indicate the decrease in accuracy compared to Figure 31. The difference between the two figures is that Figure 31 indicates the amount of reasoning that can be recovered, where as this figure shows how accurate the coders' annotations are.

in the ground truth is 6, however, coder 3 made a total of 8 annotations. He correctly identified 4 of the 6 elements, missed on identifying 2 of the 6 elements, and falsely detected 2 times when there were no corresponding elements in the ground truth.

With the "false detections" in mind, we re-examine the accuracy of the coders based not on how much of the reasoning process can be recovered, but on the accuracy of their annotations. Figure 34 shows the result of the coders' accuracies that include the coders' false detections. Not surprisingly, the accuracy of the coders all decrease slightly. The accuracy in extracting findings drop by 3% from 79% to 76%, strategies by 5% from 60% to 55%, and finally methods by 2% from 60% to 58%.

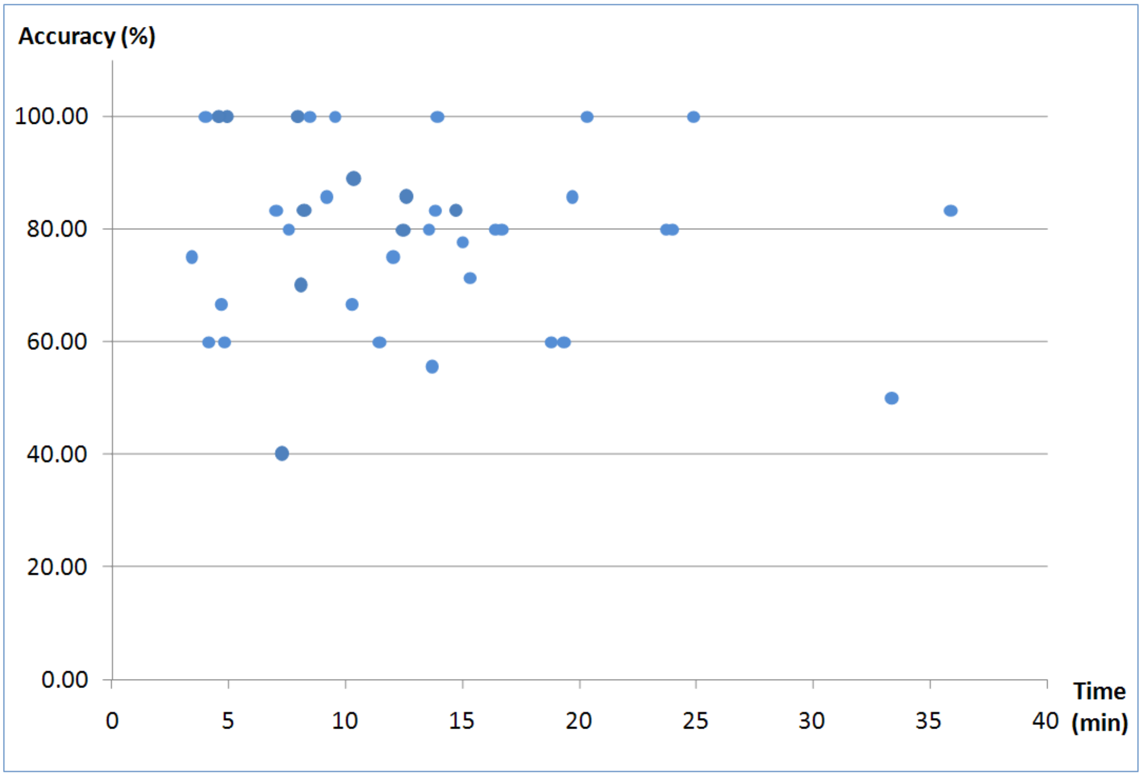


Figure 35: The accuracy of the coders in recovering "findings" of the participants and the amount of time spent.

8.5.2 Amount of time spent by coders

One important aspect in extracting reasoning process is the amount of time necessary for analyzing the interaction logs. In this section, we discuss the effect of time spent by a coder in analyzing an individual interaction log, as well as the learning effect that the coders exhibit after gaining proficiency in extracting the participants' reasoning processes.

8.5.2.1 Capturing time spent by a coder

Built into our Operation and Strategy Analysis tools is the ability to track the amount of time that a coder spends using the tools. The coders were made aware of this feature and were told not to take breaks during an analysis. Since the coders directly annotated their discoveries into the Operation Analysis tool, the system was able to record the amount of time spent by each coder when analyzing an interaction log.

Furthermore, the system tracked when the coder started and stopped the annotations. The purpose of this feature was to separate the time spent in analyzing the interaction log from the time spent in annotating. On average, the coders spend 23.9 minutes analyzing one interaction log, of which 10.75 minutes were spent on annotation and the remaining 13.15 minutes on investigation.

8.5.2.2 Time spent vs accuracy

We examine the relationship between the time spent by a coder and accuracy. Overall, there is no correlation between the two. Figure 35 plots the relationship between the coders' time spent in analysis (not including time spent for annotation) and their accuracies in extracting "findings". With the exception of the two outliers in the far right, it appears that the coders are consistently successful when spending anywhere from 5 to 15 minutes. This suggests that spending more time in the analysis does not always yield better results. The two outliers represent the analysis of coders 2 and 4 in their first investigation (participant 1). As we will show in the following section, all coders become more proficient in their analysis as they gain experience.

8.5.2.3 Increase in accuracy

As shown in Figure 32, the accuracy of the coders increase as they gain experience in investigating interaction logs as all four coders began with examining participant 1's interactions and end with participant 10's. Based on analyses using

Pearson's correlation coefficient, we find that the number of participants a coder has examined is positively correlated to the coder's accuracy. This correlation is statistically significant when extracting "findings" ($r(40) = .37, p < .05$) and "methods" ($r(40) = .52, p < .01$). Only in extracting "strategies" is the correlation weaker ($r(40) = .21, p = .182(preferred)$). While the sample size is relatively small, these statistics nonetheless imply a subtle but potentially important discovery: with more experience in analyzing interaction logs, a coder could become more proficient in extracting an analyst's reasoning process.

8.6 Discussion

The study described in this paper is complex and intricate. On top of involving real financial analysts, the transcription process, the coding, and the grading were all performed with great care and consideration. Although many of the nuances encountered during the study do not affect the results and therefore have not been described in this paper, there are some findings that might be of interest to the community. First of all, during our informal debriefing of the coders, the coders discussed the strategies that they employed in analyzing the analysts' interaction logs. It turned out that our coders often began their investigation by looking for "gaps" in the timeline of the operational view (Figure 28), which are the byproducts of the analysts taking time to write down their findings in the Discovery Sheet (section 8.4). Based on the gaps, the coders looked for the analysts' findings, and then worked backwards to discover the strategies and methods used to derive the findings.

While this strategy may seem specific to this study and non-generalizable, we argue that in a real life scenario, analysts either directly make annotations in the visualization to make note of a finding, or they write down their finding on a piece of paper for future reference. Either way, there will exist a visible marker that suggests a relevant discovery by the analyst. Therefore, while we did not anticipate this strategy by the coders, we find their quick adoption of this method to identify the analysts' findings to be effective and relevant.

A second interesting trend pointed out by our coders concerns the usefulness of our visual tools for depicting the operational and strategic aspects of the analysis (section 8.3). According to the coders during the debriefing, all of them used the Operational Analysis tool first to gain an understanding of the overall impression of an analyst's interactions. However, the Strategic Analysis tool is often utilized to examine a specific sequence of interactions when the interactions appear random and jumbled. By presenting the results of the interactions from three perspectives (accounts, keywords, and time) in the Strategic tool, the coder could often identify the focus and intent behind the series of interactions. This finding not only validates our design of the tools, but also reconfirms the importance of visualizing both the strategic and operational aspects of an analysis process. In fact, most of the coders began their investigation by identifying the "findings" through looking for gaps in the interactions, followed by looking for "strategies" through examining the overall visual patterns in both the Strategic and Operational Analysis tools without focusing on individual user interactions. Finally, "methods" were extracted through the use of the Operational Analysis tool where specific interactions were examined in detail.

One last relevant aspect of our study is the measurement of "incorrectly identified" elements in the grading process. In all of our results shown in section 8.5, we do not take into account elements that have been graded as "incorrectly identified." As mentioned in section 8.4.4, any annotation by a coder that does not perfectly match the transcription is considered to be incorrectly identified. This includes scenarios in which a coder identifies the analyst's strategy to be examining 4 keywords when in fact the analyst was examining 5, or when a coder determines that the finding of the analyst is a transaction between accounts A and B instead of accounts A and C. If we were to give half a point to these incorrectly identified elements, the overall accuracy of extracting strategies increases drastically from 60% to 71%, methods from 60% to 73%, and findings from 79% to 82%.

8.7 Summary

The path to perfectly capture an analysts reasoning process is still elusive. However, in this chapter, we have demonstrated that it is indeed possible to extract a great deal of the reasoning process through the visual examination of the analysts interactions with a financial visual analytical tool. Our results indicate that with careful design in capturing user interactions and the use of both operational and strategic tools to visually analyze an analysts interaction logs, we can understand some of the strategies, methods, and findings of an expert's analytical process.

CHAPTER 9: CONCLUSION

This dissertation identifies a bi-directional relationship between a user's analysis process and an interactive visualization system. Most of the current research fall into only one direction in this loop, which is to incorporate the analysis process into the interactive visualization. This dissertation recasts the importance of doing so, as well as proposing considerations on how to incorporate the analysis process effectively. As noted in later chapters, extracting part and eventually most of the analysis process while using an interactive visualization is a promising research direction and started gaining attention recently [86].

9.1 Review of Dissertation Contributions

The first problem addressed by this dissertation is how to incorporate target users' analysis processes into the design of interactive visualization systems, specifically in the domain of text analytics. In order to develop an understanding of the general analysis processes, we consult the Sensemaking model [94] and several visualization tasks taxonomies in Information Visualization. In order to study how to interact with domain users to develop an understanding of the domain analysis processes, we refer to ethnographical methods well-studied in the HCI community. Several general and domain-specific tasks are presented through our own interactions with domain experts and through surveying literature in the domain of (visual) text analysis.

We then developed an interactive visualization system, ParallelTopics, which incorporates a subset of the tasks to facilitate the analysis of text corpora. Specifically, ParallelTopics help users answer questions regarding a text corpus such as: What are the major topics? What topic is each document in the corpus about? Which document addresses two or more topics at once? How do the topics evolve over time? ParallelTopics leverages state-of-the-art topic models and performs further processing to visually represent both topical and temporal patterns. ParallelTopics has been evaluated by domain users and was considered great help to analyzing topical and temporal trends in documents such as scientific publications.

Through the development and user evaluation of ParallelTopics, we discovered a more pressing need from domain user to summarize and explore textual information on a large-scale. Therefore we presented a general framework, I-Si, for handling large-scale textual corpora. I-Si takes advantage of parallel computing methods to speed up the processing of large-scale text collection and supports interactive visual exploration and analysis. I-Si has been applied to analyze large text corpora such as tweet collections. Through analyzing the tweet collections, we were able to understand the evolvement of a large-scale, spontaneously organized movements such as the Occupy Wall Street, and were able to further identify early indicators preceding the actual event.

Having demonstrated the relationship from analysis process to an interactive visualization through theories and examples, the second problem this thesis addresses is how to extract some of the analysis process from evidences recorded during the use of interactive visualization systems. Grounded in the operational model of visualization [118], chapter 6 analyzes various types of evidences that are available for tracking a users's analysis process during the use of an interactive visualization system. The chapter further analyzes the advantages and shortcomings of capturing each kind of evidence. The chapter concludes that capturing user interactions within a visualization system is the best method in terms of cost/benefit ratio.

To theoretically demonstrate the effect of user interactions on analysis process or problem-solving process in general, chapter 7 presents an experiment, which studies the effect of constraining interactions on solving a mathematical puzzle. In this experiment, 5 different constraints were enforced so that participants were only able to interact with the problem under one of these constraints. The results have shown that constraints on interactions do affect problem solving, and that certain interaction constraint can lead to a higher chance of developing the best strategy.

Lastly, following the theoretical analysis, chapter 8 further studies how much analysis process can actually be recovered through analyzing user interactions within an visual analytics system. The analysis processes recovered by non-domain experts were compared against ground truth gathered from the domain experts. Our results show that more than 60% of high-level strategies, and 79% of findings could be recovered through analyzing interaction logs.

Taken together, the design considerations, interaction visualization systems, and user experiments described in this thesis demonstrate effect ways to incorporate target users' analysis processes into the design of an interactive visualization system, and that a large part of the analysis processes can in turn be captured through analyzing interaction logs while using the visualization system.

9.2 Limitations and Future Work

There is a great deal of future work that can be done to further the two research problems centered by this thesis. Theories and principles can be developed to guide how to gather and incorporate users' analysis processes into the design of interactive visualizations for multiple domains. In turn, by helping initiate research into provenance tracking through user interaction, this thesis opens the door to new lines of inquiry; we hope it serves as a prelude to a continuing stream of research.

9.2.1 Moving from Requirement to Design

Some visualization models [83, 25, 20] have focused on the importance of gathering requirements through interacting with domain users for visualization design. But few have detailed the steps of transforming needs and requirements to the first design. Although researchers have proposed methods to select appropriate visual encodings given data types [105, 81], there is still a lack of general guidelines on how to transform task requirements into visual encodings and interaction design. Recently, Wang et al. [119] proposed a framework in which task requirements are transformed first into "actionable knowledge" items, and then these items are implemented into a visual analytics system. The design framework is proposed in the context of organizational environments. We hope to see similar guidelines in other domains and ultimately general guidelines for the field of visual analytics.

9.2.2 Extending Current Research on Visual Text Analytics

One exemplar domain chapter 3, 4, and 5 focused on is visual text analysis. Visual text analysis refers to the area of using visualization to facilitate the analysis of textual information, often large-scale and unstructured. Visual text analytics systems often leverage natural language processing techniques such as named entity recognition [17], sentiment analysis [42], latent semantic analysis [75], topic models [15, 14], etc.to present different characteristics of a large document corpus. Although visual sentiment analysis [100, 49, 70] and visualization of entities [123, 113] have been advanced to aid analysis of opinions and relationship between entities (including people, organization, location, and time), few interactive visualization system has utilized a combination of theses techniques to discover and even predict meaningful "events" hidden in the text corpora. Detecting events within text streams especially in the context of social media would be extremely informative to multiple stakeholders, including emergency responders who want to track spontaneously organized gatherings, corporations to provide customized services or advertisement to potential customers, etc. With the recently development of topic models [15, 14], the detection of events is made possible on a topical level as opposed to pre-determined keywords. Therefore the resulting events are more comprehensive and semantically meaningful. In the future, we plan to focus on the temporal event detection especially in the realm of social media.

9.2.3 Extracting Analysis Process

Chapter 6 analyzed different types of evidences available to recover a user's analysis process while using an interactive visualization system. Given time and resource constraint throughout my thesis work, we only focused on the interactions logs. But we believe that to get a more comprehensive understanding of the high-level analysis process, interaction logs alone is not enough to capture the complex process. More semantic-level evidence such as annotations should be considered in the future. In addition, how much analysis process can be recovered is dependent upon how a visualization system is designed. In particular, factors such as how much interactivity is supported by a visualization system, information change caused by interactions, and the semantics of user interactions affect the amount of information one can capture within an interactive visualization system. Therefore, to extract analysis process in a systematic manner, a framework is needed to to not only inform what and how to capture for extracting analysis process, but also inform the design of an interactive visualization system from the recovery perspective.

9.3 Closing Remarks

Interactive visualization integrates new computational and theory-based tools with innovative interactive techniques and visual representations to enable human-information discourse. It is through the interactive manipulation of a visual interface, the dialogue between a user and a visualization could be established and maintained, it is also through the interactive manipulation, a user is able to explore and analyze the underlying data within a visual interface. Therefore, deepening the understanding of how interaction and visual representations could facilitate the human-information discourse yields significant impact on designing future interactive visualization systems and improving the design of existing ones. This dissertation hopes to push us closers towards systems that better incorporate and reflect high-level human reasoning process.

REFERENCES

[1] Apache cassandra project. http://cassandra.apache.org/.

[2] Google map api. http://code.google.com/apis/maps/documentation/geocoding/.

[3] Google plda. http://code.google.com/p/plda/.

[4] Mongodb project. http://www.mongodb.org/.

[5] Pacific northwest national laboratory. http://in-spire.pnl.gov/.

[6] Twitter. http://www.twitter.com.

[7] Vast challenge committee 2011. http://hcil.cs.umd.edu/localphp/hcil/vast11/.

[8] Cbs news. http://www.cbsnews.com/8301-501465_162-20117291-501465.html, Oct 7, 2011.

[9] Occupy wall street to turn manhattan into tahrir square. http://newyork.ibtimes.com/articles/215511/20110917/occupy-wall-street-new-york-saturday-protest.htm, Sep 17, 2011.

[10] Amar, R., Eagan, J., and Stasko, J. Low-level components of analytic activity in information visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 15–.

[11] Bavoil, L., Callahan, S., Crossno, P., Freire, J., Scheidegger, C., Silva, C., and Vo, H. Vistrails: enabling interactive multiple-view visualizations. *Visualization, 2005. VIS 05. IEEE* (Oct. 2005), 135–142.

[12] Becker, R. A., and Cleveland, W. S. Brushing Scatterplots. *Technometrics 29*, 2 (1987), 127–142.

[13] Blascovich, J., Mendes, W. B., Hunter, S. B., and Salomon, K. Social "facilitation" as challenge and threat. *Journal of Personality and Social Psychology 31* (1999), 422–429.

[14] Blei, D., and Lafferty, J. *Text Mining: Theory and Applications.* Taylor and Francis, 2009, ch. Topic Models.

[15] Blei, D. M. Introduction to probabilistic topic models. *Communications of the ACM* (2011).

[16] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3* (Mar. 2003), 993–1022.

[17] Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. Nyu: Description of the mene named entity system as used in muc-7. In *In Proceedings of the Seventh Message Understanding Conference (MUC-7* (1998).

[18] Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems (NIPS)* (2009).

[19] Card, S. K., and Mackinlay, J. The structure of the information visualization design space. In *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)* (Washington, DC, USA, 1997), IEEE Computer Society, pp. 92–.

[20] Card, S. K., Mackinlay, J., and Shneiderman, B. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[21] Carroll, J. M., Thomas, J. C., and Malhotra, A. Presentation and representation in design problem-solving. *British Journal of Psychology 71*, 1 (1980), 143–154.

[22] Chang, R., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Ziemkiewicz, C., Kern, D., and Sudjianto, A. Wirevis: Visualization of categorical, time-varying data from financial transactions. *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on* (30 2007-Nov. 1 2007), 155–162.

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

[24] Chi, E. H. A taxonomy of visualization techniques using the data state reference model. In *Proceedings of the IEEE Symposium on Information Vizualization 2000* (Washington, DC, USA, 2000), INFOVIS '00, IEEE Computer Society, pp. 69–.

[25] Chi, E. H.-h., and Riedl, J. An operator interaction framework for visualization systems. In *Proceedings of the 1998 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1998), IEEE Computer Society, pp. 63–70.

[26] Cleveland, W. S., and McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association 79*, 387 (1984), 531–554.

[27] Constantine, L. L., and Lockwood, L. A. D. *Software for use: a practical guide to the models and methods of usage-centered design*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1999.

[28] Convertino, G., Chen, J., Yost, B., Ryu, Y.-S., and North, C. Exploring context switching and cognition in dual-view coordinated visualizations. *Coordinated and Multiple Views in Exploratory Visualization, 2003. Proceedings. International Conference on* (July 2003), 55–62.

[29] Cowley, P., Haack, J., Littlefield, R., and Hampson, E. Glass box: capturing, archiving, and retrieving workstation activities. In *CARPE '06: Proceedings of the 3rd ACM workshop on Continuous archival and retrival of personal experences* (2006), pp. 13–18.

[30] Cowley, P., Nowell, L., and Scholtz, J. Glass box: An instrumented infrastructure for supporting human interaction with information. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005. HICSS '05* (January 2005), pp. 296c–296c.

[31] Davis, A. M. *201 Principles of Software Development*. McGraw-Hill, Mar. 1995.

[32] Dean, J., and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Commun. ACM 51* (Jan. 2008), 107–113.

[33] Diakopoulos, N., Naaman, M., and Kivran-Swaine, F. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (oct. 2010), pp. 115 –122.

[34] Dionisio, J. D. N. Course slides: Hci principles. http://myweb.lmu.edu/dondi/share/ixd/principles.pdf.

[35] Dix, A., Finlay, J. E., Abowd, G. D., and Beale, R. *Human-Computer Interaction (3rd Edition)*, 3 ed. Prentice Hall, Dec. 2003.

[36] Dork, M., Gruen, D., Williamson, C., and Carpendale, S. A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (nov.-dec. 2010), 1129 –1138.

[37] Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R., and Chang, R. Recovering reasoning process from user interactions. *IEEE Computer Graphics and Applications Speicial Issue on Visual Analytics Evaluation* (2009). To Appear.

[38] Dou, W., Wang, X., Chang, R., and Ribarsky, W. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (oct. 2011), pp. 231 –240.

[39] Duchowski, A. T. *Eye Tracking Methodology: Theory and Practice*. Springer, January 2003.

[40] Eccles, R., Kapler, T., Harper, R., and Wright, W. Stories in geotime. pp. 19–26.

[41] Feldman, R. Text Analytics: Theory and Practice. ACM Thirteenth Conference on Information and Knowledge Management (CIKM) Workshops, 2004.

[42] Gamon, M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics* (Stroudsburg, PA, USA, 2004), COLING '04, Association for Computational Linguistics.

[43] Gotz, D., and Zhou, M. Characterizing users visual analytic activity for insight provenance. *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on* (Oct. 2008), 123–130.

[44] Green, T. M., Ribarsky, W., and Fisher, B. Visual analytics for complex concepts using a human cognition model. *Visual Analytics Science and Technology, 2008. VAST 2008. IEEE Symposium on* (November 2008), 91 – 98.

[45] Greitzer, F. Methodology, metrics and measures for testing and evaluation of intelligence analysis tools. PNWD-3550, Battelle-Pacific Northwest Division, Richland, WA, 2005.

[46] Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101* (Apr. 2004), 5228–5235.

[47] Groth, D. P., and Streefkerk, K. Provenance and annotation for visual exploration systems. *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1500–1510.

[48] Hansen, W. J. User engineering principles for interactive systems. In *Proceedings of the November 16-18, 1971, fall joint computer conference* (New York, NY, USA, 1971), AFIPS '71 (Fall), ACM, pp. 523–532.

[49] Hao, M. C., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L.-E., and Hsu, M. Visual Sentiment Analysis on Twitter Data Streams (Poster Paper). In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST 2011)* (2011).

[50] Havre, S., Hetzler, B., and Nowell, L. Themeriver: visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on* (2000), pp. 115 –123.

[51] Hayes, J. R., and Simon, H. A. Psychological differences among problem isomorphs. In *Cognitive Theory*, J. Castellan, D. B. Pisoni, and G. Potts, Eds., vol. 2. Erlbaum, 1977.

[52] Heer, J., Mackinlay, J., Stolte, C., and Agrawala, M. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (Nov.-Dec. 2008), 1189–1196.

[53] Heer, J., Mackinlay, J. D., Stolte, C., and Agrawala, M. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1189 – 1196.

[54] Heer, J., Viégas, F. B., and Wattenberg, M. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2007), ACM, pp. 1029–1038.

[55] Hilbert, D. M., and Redmiles, D. F. Extracting usability information from user interface events. *ACM Comput. Surv. 32*, 4 (2000), 384–421.

[56] Huang, W. Using eye tracking to investigate graph layout effects. pp. 97–100.

[57] Hundhausen, C., Douglas, S., and Stasko, J. A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing 13*, 3 (2002), 259–290.

[58] Hutchins, E. Material anchors for conceptual blends. *Journal of Pragmatics 37*, 10 (October 2005), 1555–1577.

[59] Inselberg, A. The plane with parallel coordinates. *The Visual Computer 1* (1985), 69–91. 10.1007/BF01898350.

[60] Jankun-Kelly, T., Ma, K.-L., and Gertz, M. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics 13*, 2 (March/April 2007), 357–369.

[61] Jankun-Kelly, T. J., Ma, K.-L., and Gertz, M. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics 13*, 2 (2007), 357–369.

[62] Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (New York, NY, USA, 2007), WebKDD/SNA-KDD '07, ACM, pp. 56–65.

[63] Jeong, D. H., Dou, W., Lipford, H., Stukes, F., Chang, R., and Ribarsky, W. Evaluating the relationship between user interaction and financial visual analysis. *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on* (Oct. 2008), 83–90.

[64] Kindlmann, G. Lack of reproducibility hinders visualization science. *IEEE Visualization Compendium, IEEE CS Press* (2006).

[65] Kirsh, D. The intelligent use of space. *Artificial Intelligence 73*, 1-2 (1995), 31–68.

[66] Kirsh, D. Interaction, external representations and sense making. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Austin, TX, 2009), N. A. T. . H. van Rijn, Ed., Cognitive Science Society, pp. 1103–1108.

[67] Kirsh, D. Problem solving and situated cognition. In *The Cambridge Handbook of Situated Cognition* (Cambridge, 2009), Cambridge University Press, pp. 264–306.

[68] Kirsh, D. Projection, problem space and anchoring. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Austin, TX, 2009), N. A. T. . H. van Rijn, Ed., Cognitive Science Society, pp. 2310–2315.

[69] Kirsh, D., and Maglio, P. On distinguishing epistemic from pragmatic actions. *Cognitive Science 18*, 4 (1994), 513–549.

[70] Kisilevich, S., Rohrdantz, C., Veronica, M., and Keim, D. A. What do you think about this photo? A novel approach to opinion and sentiment analysis of photo comments. *International Journal of Data Mining, Modelling and Management (IJDMMM)* (2011).

[71] Kotovsky, K., and Kushmerick, N. Processing constraints and problem difficulty: A model. In *Proc. Conference of the Cognitive Science Society* (1991), K. J. Hammond and D. Gentner, Eds.

[72] Kreuseler, M., Nocke, T., and Schumann, H. A history mechanism for visual data mining. *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on* (0-0 2004), 49–56.

[73] Kristen Christian. Bank transfer day. http://www.facebook.com/Nov.Fifth/, Nov 5, 2011.

[74] Lam, H. A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1149–1156.

[75] Landauer, T. K., and Dumais, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*, 2 (1997), 211 – 240.

[76] Larkin, J. H., and Simon, H. A. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science 11* (1987), 65–99.

[77] Lipford, H. R., Stukes, F., Dou, W., Hawkins, M. E., and Chang, R. Helping users recall their reasoning process. *Visual Analytics Science and Technology, 2008. VAST 2008. IEEE Conference on* (2010), 187–194.

[78] Liu, Z., Nersessian, N., and Stasko, J. Distributed cognition as a theoretical framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1173–1180.

[79] Liu, Z., Nersessian, N. J., and Stasko, J. T. Distributed cognition as a theoretical framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1173–1180.

[80] MacEachren, A., Jaiswal, A., Robinson, A., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (oct. 2011), pp. 181 –190.

[81] Mackinlay, J. Automating the design of graphical presentations of relational information. *ACM Trans. Graph. 5* (April 1986), 110–141.

[82] Mackinlay, J., Hanrahan, P., and Stolte, C. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1137–1144.

[83] Munzner, T. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics 15* (November 2009), 921–928.

[84] Newman, D., Smyth, P., and Steyvers, M. Scalable parallel topic models. *Journal of Intelligence Community Research and Development* (2006).

[85] Nielsen, J. Evaluating the thinking-aloud technique for use by computer scientists. 69–82.

[86] North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B., and Fink, G. Analytic provenance: process+interaction+insight. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (New York, NY, USA, 2011), CHI EA '11, ACM, pp. 33–36.

[87] Oelke, D., Hao, M., Rohrdantz, C., Keim, D., Dayal, U., Haug, L.-E., and Janetzko, H. Visual opinion analysis of customer feedback data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (oct. 2009), pp. 187 –194.

[88] Oelke, D., Spretke, D., Stoffel, A., and Keim, D. Visual readability analysis: How to make your writings easier to read. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (oct. 2010), pp. 123 –130.

[89] Palmer, S. E. Fundamental aspects of cognitive representation. In *Cognition and Categorization*, E. Rosch and B. B. Lloyd, Eds. Lawrence Erlbaum Associates, 1978, pp. 262–304.

[90] Pike, W., Bruce, J., Baddeley, B., Best, D., Franklin, L., May, R., Rice, D., Riensche, R., and Younkin, K. The scalable reasoning system: Lightweight visualization for distributed analytics. pp. 131–138.

[91] Pike, W., May, R., and Turner, A. Supporting knowledge transfer through decomposable reasoning artifacts. In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007* (January 2007), pp. 204c–204c.

[92] Pike, W., May, R., and Turner, A. Supporting knowledge transfer through decomposable reasoning artifacts. In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007* (January 2007), pp. 204c–204c.

[93] Pike, W. A., Stasko, J., Chang, R., and O'Connell, T. A. The science of interaction. *Information Visualization 8* (2009).

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

[95] Qu, Y., and Furnas, G. W. Sources of structure in sensemaking. In *CHI '05 extended abstracts on Human factors in computing systems* (New York, NY, USA, 2005), CHI EA '05, ACM, pp. 1989–1992.

[96] Ramage, D., Dumais, S., and Liebling, D. Characterizing microblogs with topic models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (2010), AAAI.

[97] Ramage, D., and Rosen, E. Stanford Topic Modeling Toolbox, http://nlp.stanford.edu/software/tmt/tmt-0.3/.

[98] Ribarsky, W., Fisher, B., and Pottenger, W. M. Science of analytical reasoning. *Information Visualization 8*, 4 (2009), 254–262.

[99] Robinson, A. C., and Weaver, C. Re-visualization: Interactive visualization of the process of visual analysis. In *GIScience Workshop on Visual Analytics & Spatial Decision Support* (2006).

[100] Rohrdantz, C., Hao, M. C., Dayal, U., Haug, L.-E., and Keim, D. A. Feature-based Visual Sentiment Analysis of Text Document Streams. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Visual Interfaces for Text Analysis 3*, 2 (2012).

[101] Rushkoff, D. Think occupy wall st. is a phase? you don't get it. http://www.cnn.com/2011/10/05/opinion/rushkoff-occupy-wall-street/index.html, Oct 5, 2011.

[102] Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems* (New York, NY, USA, 1993), CHI '93, ACM, pp. 269–276.

[103] Shapiro, M. A. *Measuring Psychological Responses To Media Messages.* Lawrence Erlbaum Associates, 1994, ch. Think-Aloud and Thought-List Procedures in Investigating Mental Processes, pp. 1–14.

[104] Sharp, H., Rogers, Y., and Preece, J. *Interaction Design: Beyond Human-Computer Interaction*, 2 ed. Wiley, Mar. 2007.

[105] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. *Visual Languages, IEEE Symposium on 0* (1996), 336.

[106] Shrinivasan, Y. B., and van Wijk, J. J. Supporting the analytical reasoning process in information visualization. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), ACM, pp. 1237–1246.

[107] Shrinivasan, Y. B., and van Wijk, J. J. Supporting the analytical reasoning process in information visualization. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), ACM, pp. 1237–1246.

[108] Silva, C., Freire, J., and Callahan, S. Provenance for visualizations: Reproducibility and beyond. *Computing in Science & Engineering 9*, 5 (Sept.-Oct. 2007), 82–89.

[109] Simon, H. A. *The Sciences of the Artificial.* The MIT Press, 1996.

[110] Site Analytics. Twitter. http://siteanalytics.compete.com/twitter.com/.

[111] Sizov, S. Geofolk: latent spatial semantics in web 2.0 social media. In *Proceedings of the third ACM international conference on Web search and data mining* (New York, NY, USA, 2010), WSDM '10, ACM, pp. 281–290.

[112] Slocum, T. A., McMaster, R. B., Kessler, F. C., and Howard, H. H. *Thematic Cartography and Geovisualization.* Prentice Hall, 2008.

[113] Stasko, J., Gorg, C., Liu, Z., and Singhal, K. Jigsaw: Supporting investigative analysis through interactive visualization. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on* (30 2007-nov. 1 2007), pp. 131 –138.

[114] Stenning, K., and Oberlander, J. A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science 19*, 1 (1995), 97–140.

[115] Thomas, J. J., and Cook, K. A. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[116] Trafton, J. G., Kirschenbaum, S. S., Tsu, T. L., Miyamoto, R. T., Ballas, J. A., and Raymond, P. D. Turning pictures into numbers: extracting and generating information from complex visualizations. *Int. J. Hum.-Comput. Stud. 53*, 5 (2000), 827–850.

[117] Tweedie, L. Characterizing interactive externalizations. In *Proceedings CHI* (1997), ACM Press, pp. 375–382.

[118] van Wijk, J. The value of visualization. *Visualization, 2005. VIS 05. IEEE* (Oct. 2005), 79–86.

[119] Wang, X., Dou, W., Butkiewicz, T., Bier, E. A., and Ribarsky, W. A two-stage framework for designing visual analytics system in organizational environment. In *Visual Analytics Science and Technology (VAST), 2011, IEEE Symposium on* (2011).

[120] Wang, X., Miller, E., Smarick, K., Ribarsky, W., and Chang, R. Investigative visual analysis of global terrorism. *Computer Graphics Forum 27*, 3 (2008), 919–926.

[121] Wang, X., Miller, E., Smarick, K., Ribarsky, W., and Chang, R. Investigative visual analysis of global terrorism. In *Computer Graphics Forum* (2008), Computer Graphics Forum, pp. 919–926.

[122] Wanner, F., Rohrdantz, C., F. Mansmann, A. S., D. Oelke, M. K., Keim, D. A., Luo, D., Yang, J., and Atkinson, M. Large-scale Comparative Sentiment Analysis of News Articles (InfoVis 2009). Presented at the poster session at IEEE Information Visualization Conference 2009 (IEEE InfoVis 2009), Atlantic City, USA, 2009.

[123] Wanner, F., Schaefer, M., Leitner-Fischer, F., F. Zintgraf, M. A., and Keim, D. A. DYNEVI - DYnamic News Entity VIsualization. In *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010)* (June 2010), J. Kohlhammer and D. A. Keim, Eds., Eurographics, pp. 69–74.

[124] Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L., and Zhang, Q. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2010), KDD '10, ACM, pp. 153–162.

[125] Wijk, J. J. v. Bridging the gaps. *IEEE Comput. Graph. Appl. 26* (November 2006), 6–9.

[126] Wright, W., Schroh, D., Proulx, P., Skaburskis, A., and Cort, B. The sandbox for analysis: concepts and methods. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), ACM, pp. 801–810.

[127] Yi, J. S., Kang, Y. a., Stasko, J., and Jacko, J. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1224–1231.

[128] Zanbaka, C. A., Ulinski, A. C., Goolkasian, P., and Hodges, L. F. Social responses to virtual humans: implications for future interface design. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2007), ACM, pp. 1561–1570.

[129] Zhou, M. X., and Feiner, S. K. Visual task characterization for automated visual discourse synthesis. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1998), CHI '98, ACM Press/Addison-Wesley Publishing Co., pp. 392–399.