DESIGNING DISTRIBUTED COLLABORATIVE VISUAL ANALYTICS SYSTEMS

by

Syavash Nobarany
Bachelor of Science, University of Tehran, 2008

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
School
of
Interactive Arts and Technology

SIMON FRASER UNIVERSITY
Fall 2010

APPROVAL

Name: **Syavash Nobarany**

Degree: **Master of Science**

Title of Thesis: **Designing Distributed Collaborative Visual Analytics Systems**

Examining Committee:

Chair: ______________________________

Dr. Marek Hatala
Associate Professor of School of Interactive Arts and Technology

____

Dr. Brian Fisher
Senior Supervisor
Associate Professor of School of Interactive Arts and Technology

____

Dr. John Dill
Supervisor
Professor Emeritus of School of Interactive Arts and Technology

____

Steve DiPaola
External Examiner
Associate Professor of School of Interactive Arts and Technology

Date Defended/Approved: September 3, 2010

SFU SIMON FRASER UNIVERSITY
LIBRARY

Declaration of Partial Copyright Licence

STATEMENT OF ETHICS APPROVAL

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

(a) Human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

(b) Advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

(c) as a co-investigator, collaborator or research assistant in a research project approved in advance,

or

(d) as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Simon Fraser University
Burnaby, BC, Canada

ABSTRACT

Analysts should be able to collaboratively work on enormous amount of available information and share their findings and understandings to effectively and efficiently make sense of the situation under investigation.

The general question this thesis addresses is “How can a distributed collaborative analytics system support efficient and effective distributed (in time and space) collaboration among analysts?” and we focus on answering “How can a collaborative analytics system support efficient and effective reuse of the reasoning artefacts such as arguments, causal maps, etc.?”

Through deepening our understanding of the individual and collaborative sensemaking processes that analysts go through, we identified design guidelines for enhancing, facilitating collaborative processes, fostering sharing and reuse, and improving collaboration efficiency. The design guidelines informed the design of a collaborative analytics system called AnalyticStream. We validate the proposed guidelines through the evaluation of the system.

Keywords: analyst, reuse, visual analytics, distributed cognition, activity theory, distributed reasoning.

ACKNOWLEDGEMENTS

During the last two years of my research and studies, I was fortunate to benefit from great mentors, supporters and colleagues. I thank my supervisor, Brian Fisher for his patience and continuous support during these two years. His deep knowledge helped me navigate several unfamiliar areas of science and his calm caring and thoughtful manner created a suitable environment for my education and research. I also thank my committee members, John Dill and Steve DiPaola for their insightful comments and suggestions.

I thank my wife, Mona, for creating a happy and peaceful environment and collaborating with me in every stage of my research.

I thank my family for their continued tireless support and encouragement that enabled me to focus on my studies and research.

I thank my friends and colleagues at SFU Visual Analytics Research and SCIENCE lab and UBC MAGIC lab that helped me during various stages of my research with their comments.

I thank all who contributed to my research during the last two years directly or indirectly.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1: INTRODUCTION

One of the primary goals of science of visual analytics is to support dealing with complex and dynamic situations and problems that require several analysts or even several teams of analysts to involve in a richly collaborative analytical process. In a collaborative analysis process, analysts engage in intertwined processes of sensemaking and sensegiving to address a problem more accurately and more comprehensively. However, the lack of support for collaboration in analysis and reasoning support systems inhibits analysts from efficient cooperation and engaging in collaborative sensemaking processes (Heer, F. B. Viégas, & Wattenberg, 2009; J. J. Thomas & Cook, 2005). Recently there have been efforts for clarifying the deficiencies and identifying the requirements for designing collaborative visual analytics tools (Heer & Agrawala, 2007; Viegas & Wattenberg, 2006). These studies were followed by implementation of collaborative information visualization systems, and user studies that revealed several patterns of social visualization and the value of focusing on social visual analytics (Heer et al., 2009; A. B. Viégas, Wattenberg, Van Ham, Kriss, & Mckeon, 2007).

Following those efforts, we investigated the collaborative analytical reasoning processes, as a rather less emphasized area in visual analytics research. We focus on how analysts can benefit the most from each other's analysis outcomes and processes to perform a more comprehensive and deeper analysis. This leads us to several questions such as “How can analysts reuse each other’s pieces of reasoning? How can the visual analytics tool support the process of sharing and reusing? How can analysts decompose the analysis process into sensible pieces that are sharable and reusable? What incentives may be used to engage users in the collaborative analysis process and satisfy them? How can we encourage users to share their pieces and reuse each other’s pieces of analysis?”

1.1 Thesis Problem and Approach

Human reasoning and analysis is such a valuable resource that should be reused as much as possible, however, reusing analysis products and processes is so challenging that the reuse occurs very rarely if we consider the amount of reasoning that every person does. The absence of reusing analysis processes and products is a part of a more general problem that is the lack of infrastructure for supporting collaboration in analysis and reasoning tools. This problem inhibits analysts from efficient cooperation and engaging in collaborative sensemaking processes.

The analysis process involves seeking *relevant* resources followed by extraction, marshalling and summarizing the relevant pieces of information. Although *relevant* pieces can be different for various purposes, many of the problems and situations share their requirements and their perspectives for finding relevant pieces. This result in repetition and redundancy of the activities aimed at preparing resources for integration and reuse. This redundancy particularly increases when the problems and situations under investigation are related and intertwined at different levels, which is the case in many domains.

We believe that designing a platform for explicit support of sharing and reuse of pieces of analysis can be beneficial in many ways. Reusable pieces of analysis can facilitate analytical communication process that can support construction and clarification and sharing of meaning, while increasing the performance of sensemaking and evaluation of the previous efforts.

Considering the tendency of analysts and researchers for working independently (Bos et al., 2007), a loosely coupled asynchronous collaboration process can be more favourable; however it amplifies the difficulty of knowledge transfer, especially at early stages of analysis, in which the analyst has not yet come up with a well-formed representation of his ideas. This challenge requires the design to avoid imposing any structure or language to the analysts. In the current research, we focused on asynchronous collaboration through reusable pieces of analysis. We aim to address the issues associated with their reusability and understand the design considerations for facilitating and fostering sharing and reuse of reasoning artefacts.

We started with making sense of the design space of distributed collaborative visual analytics systems and identifying guidelines to devise the design of a web-based collaborative visual analytics system called AnalyticStream to validate and expand our understanding of the design space. We adopted a mixed methods approach and conducted a case study to better understand various aspects of the design and the application of theories. The findings of this study are discussed in relation to relevant theories and the preliminary exploration of the design space to refine and prioritize the suggested guidelines.

1.2 Thesis Contributions

This thesis contributes a new understanding of the design space of distributed collaborative analytics systems and a web-based system for facilitating distributed collaborative analysis process:

- We explored the design space of collaborative visual analytics systems though analyzing various related concepts and models and applying two different theoretical perspectives, namely Distributed Cognition Theory and Activity Theory, to provide a comprehensive view of the design space.
- We designed AnalyticStream, a web-based distributed collaborative analytics system, to demonstrate how some of the theoretically grounded suggestions and guidelines can be implemented.
- Finally, we conducted a mixed-methods case study evaluating AnalyticStream, to expand, validate and refine our understanding of requirements for designing distributed collaborative visual analytics systems.

1.3 Thesis Outline

Chapter 2 covers the most important related visual analytics systems and design space explorations, to describe the research area and research trend that has led to this work. Chapter 3 draws from the related theoretical concepts and models as foundations of understanding and designing collaborative visual analytics systems to help us understand the underlying processes that should be supported and concepts that should be understood to approach the design of collaborative visual analytics systems. Chapter 4 explores the design space of distributed visual analytics systems through the two theoretical lenses of Distributed Cognition Theory and Activity Theory and using the concepts described in chapter 2. Therefore, chapter 4 provides a comprehensive understanding of the design space, which can greatly help system designers to think about their specific design space and identify requirements and considerations for their specific design situation. Chapter 5 describes various aspects and features of a web-based distributed visual analytics system called AnalyticStream, to demonstrate how the theoretically based suggestions can be applied. Chapter 6 presents the methodology and results of the case study that was conducted to evaluate AnalyticStream and discusses the findings and limitations of the study. Finally, Chapter 7 summarizes the contributions, discusses the implications of the reported findings of this thesis, and describes the future works that can follow-up this research.

2: RELATED WORKS

Considering that this thesis has two main components, exploring the design space of collaborative analytics systems, and the design and evaluation of an asynchronous distributed visual analytics system. We divide the related works into two categories. The first category includes the research efforts aiming at applying theories and frameworks for understanding the design spaces of collaborative visual analytics systems. The second category consists of research projects that have built related visual analytics systems. In the following sections, the two categories will be discussed briefly.

2.1 Related Visual Analytics Systems

Analysts often need to keep track of many facts, assumptions and hypotheses, and use any analysis method such as analysis of competing hypotheses (ACH), inference networks, evidence marshalling, etc. The amount of information that they need to keep track of and process is more than what human internal cognitive resources can handle. This requires some sort of externalization (Norman, 1994) of the analysis outcomes and process, which is traditionally performed using paper notes, concept maps and other reasoning and argumentation diagrams (Kirschner, Buckingham-Shum, & Carr, 2003). Visual analytics tools extend these artefacts by providing interactive visualizations of the information space under investigation.

Most of the visual analytics tools such as Improvise (C. Weaver, 2004) and Jigsaw (John Stasko, Görg, & Spence, 2008) focus on individual analysis and do not offer any explicit support for collaboration. However, due to the increasing demand for supporting teams of analysts (J. J. Thomas & Cook, 2005; Viegas & Wattenberg, 2006), there is a significant trend toward enhancing the support for collaborative analysis (Heer, van Ham, Carpendale, Chris Weaver, & Isenberg, 2008). Recently some of the information visualization systems such as Tibco Spotfire, Tableau, and special-purpose information visualization systems such as Sunfall for astrophysicists (Aragon, Poon, Aldering, R. C. Thomas, & Quimby, 2009) support sharing and annotating visualizations to enable collaborative processes.

Information visualization systems such as Sense.us (Heer, F. B. Viégas, & Wattenberg, 2007) and IBM's ManyEyes (A. B. Viégas et al., 2007) can be considered as pioneers of collaborative visualization. They revealed some of the patterns of social visualization and the value of focusing on social visual analytics, which shed some light on many interesting research areas that involve understanding collaborative and social processes. Other related web-based information visualization systems are Swivel.com, that supports sharing and conversations on visualizations, and Wikimapia that enables collaborative annotation of maps. Although these systems provide many possibilities for analysts' communication and cooperation, most of them do not provide specific mechanisms for facilitating the reuse of analysis products.

Another class of visual analytics tools help users to manage reasoning artefacts, entities and notes and allow sharing of them. These systems date back to idea sharing systems such as NoteCards (Halasz, Moran, & Trigg, 1987; Trigg, Suchman, & Halasz, 1986). More recently, Entity workspace (Bier, Ishak, & Chi, 2006) is designed to support evidence filing and marshalling using an entity graph. Analyst's Notebook supports capturing, reviewing, and sharing reasoning artefacts. X-media project (Dadzie, Lanfranchi, & Petrelli, 2009) enables collective creation of a visual semantic information space. Users can interactively explore the ontology using knowledge lenses and graphs during an analysis. They can review, reuse and share the ontology during an analysis. Oculus Sand Box supports an extensive set of analytical methods and helps analysts to manage entities and create stories based on visualizations and share their analysis results (Wright, Schroh, Proulx, Skaburskis, & Cort, 2006). PNNL's Scalable Reasoning System (SRS) (W.A. Pike et al., 2008; William Pike, Richard May, & Turner, 2007) which takes a similar approach to ours and focuses on the ability to disassemble knowledge representations. Based on hermeneutic principles, they provide a comprehensive theoretical view of the analytic discourse that is supported through sharing of knowledge representations that carry the history of how they are shaped based on the integration of other artefacts. Following these studies, we designed AnalyticStream as a distributed collaborative evidence filing and marshalling system with an emphasis on facilitating sharing and reuse of reasoning artefacts.

2.2 Related Design Space Explorations

Recently there have been efforts to clarify the problem and requirements of designing collaborative visual analytics tools. Viegas and Wattenberg published "call for action" (Viegas & Wattenberg, 2006) to invite researchers to consider visualization as a mean of communication more seriously. They focus on the critical role of synchronous and asynchronous conversations, and social activities surrounding visual data analysis. Illuminating the path (J. J. Thomas & Cook, 2005) introduced the challenge of human scalability and calls for developing "techniques that gracefully scale from a single user to a collaborative" and emphasizes supporting collaborative sensemaking in details (chapter 2). Heer and Agrawala discussed the importance of collaborative visual analytics systems and collected several design considerations for collaborative information visualization systems (Heer & Agrawala, 2007). These studies provide a valuable set of design guidelines and highlight several important aspects of designing collaborative visual analytics systems. However, due to the extensiveness of the design space, they usually could not thoroughly cover it and their picture of the design space is not comprehensive.

We believe that providing system designers with appropriate theoretical lenses that can present a comprehensive continuous view of the design space while magnifying important aspects of it can facilitate the process of making sense of the design space under investigation. This approach requires more effort on the system designer's part by demanding careful investigation of the design space, but can provide the system designer with a deeper and more comprehensive understanding of various aspects of the design space. This approach has been employed in some of the related research areas such as computer-supported collaborative work (CSCW) in (Halverson, 2002) which analyzes the Distributed Cognition Theory and Activity Theory as frameworks for exploring and understanding design considerations for CSCW systems. Nardi and Kaptelinin presented Activity Theory as a framework for understanding and designing human-computer interactions and interfaces (Victor Kaptelinin & Bonnie A. Nardi, 2009). Moreover, they identified design guidelines to facilitate applying their understanding of Activity Theory's implications for human computer interaction (V. Kaptelinin, B. A Nardi, & Macaulay, 1999). Hollan, Hutchins and Kirsh suggested Distributed Cognition Theory as a candidate foundation for designing and understanding human-computer interactions, systems and workspaces (Hollan, Hutchins, & Kirsh, 2000). They proposed and integrated research framework which shows how Distributed Cognition Theory, ethnography and lab experiments can fit together to ultimately guide the design of tools and workplaces. More recently, Liu, Nersessian and Stasko proposed Distributed Cognition Theory as a possible framework for analyzing, designing and evaluating information visualization system (Liu, Nersessian, & J. T Stasko, 2008). Following these studies, in current research, we attempt to explore the design space of distributed visual analytics systems considering the theoretical lenses of Distributed Cognition Theory and Activity Theory.

3: FOUNDATIONS OF COLLABORATIVE SENSEMAKING PROCESS

In this section, we discuss the foundations of designing distributed collaborative analysis support systems and we explain how various social and collaborative aspects of analysis and decision-making processes can contribute to the process. Collaborative systems should support the social distribution of cognitive processes. This means that the system should facilitate the integration of analysts' efforts through supporting the coordination, sharing and reuse of reasoning processes and products. Integration is not achieved by a summation of different perspectives; it requires sharing of perspectives and dialogues and interaction for recognizing conflicts in analysis results or sources of information (Richard J. Boland, Maheshwari, Te'eni, Schwartz, & Tenkasi, 1992). Additionally distributed analytics systems should address situations that are too complex for one person to understand entirely; in these situations, agents have limited models of the problem as a whole, and may never develop a proper comprehensive understanding of the problem (Brehmer, 1991). These issues and several others that will be discussed in upcoming sections define a complex design space. We try to make sense of a complex design space using various related concepts and models in this chapter and by analyzing the design space through theoretical lenses of Distributed Cognition Theory and the activity theory in chapter 4.

We start with reviewing sensemaking models, and continue with reviewing the literature related to understanding and designing for collaboration, and finally we discuss the various aspects of reuse process from cognitive and collaborative perspectives.

In the following section, we review the two most widely accepted sensemaking models to elaborate on how a collaborative system may support the social distribution of sensemaking processes.

3.1 Sensemaking Models

A number of candidate definitions for sensemaking are offered throughout the literature. For example, (Klein, Moon, & Hoffman, 2006) has defined sensemaking as a motivated continuous effort to understand connections among entities in order to anticipate their trajectories and act effectively. Wieck (2005) reviewed various definitions of sensemaking and broadly describes it as the process of making meaning from information. Recent models of sensemaking process aim to inform the design of reasoning environments and facilitate finding ways for supporting the sensemaking processes. Often-cited sensemaking models are offered by (Russell, Stefik, Pirolli, & Card, 1993) as “learning loop complex” and by (Pirolli & Card, 2005) as “sensemaking loop for intelligence analysis”.

3.1.1 Learning Loop Complex

Russel's learning loop complex theory (Russell et al., 1993) emphasizes the role of representations and describes sensemaking as the process of searching for a representation and encoding data in that representation to answer questions. The learning loop complex (Figure 1) starts with creating representations that capture relevant aspects of the data. This process is referred to as generation loop. Then, the sensemaker discovers new information and encodes them in the selected representation from the generation loop. Some of the new information do not exactly match the representations and procedures (residues in the diagram), and lead to the representational shift loop in which representations are expanded, merged, refined, etc. to better represent those information.

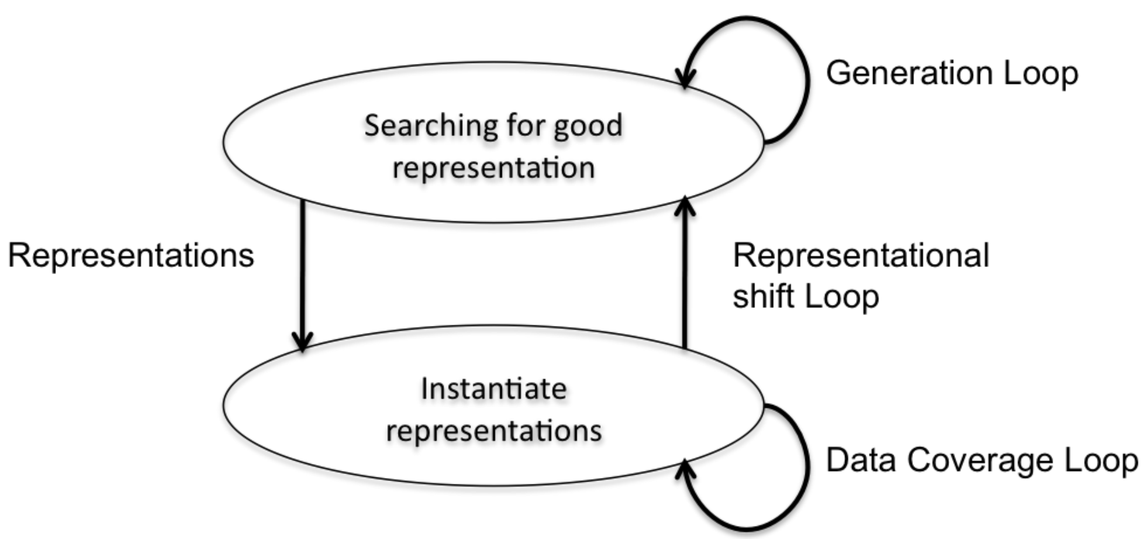

Figure 1. Learning Loop Complex theory of sensemaking

Finally, the representations will be used for a specific task to guide the sensemaker in his analysis process by pointing at relevant features of the information and the questions that should be asked for performing the task in an effective and efficient way.

3.1.2 Sensemaking Loop for Intelligence Analysis

Card and Pirolli studied intelligence analysts using cognitive task analysis and think aloud protocol analysis to characterize the activities that they go through during the sensemaking process. Figure 2 shows the transformation of information from raw evidences in external data sources to final products or reports. The shoebox is the subset of the external data that is relevant for processing. The evidence file contains snippets extracted from items in the shoebox. These snippets are marshalled or re-represented as schemas to be used for hypothesizing and drawing conclusions, which result in final products that are presented to other analysts and decision makers.

The transformations from less structured and less meaningful information to more organized and meaningful information are described as bottom-up processes. This processes start with searching and filtering external data and extracting relevant pieces of information and are followed by schematizing them using information visualization, evidence marshalling and other analytical methods. The last phases are building cases that support or reject ypotheses, and finally telling stories that are aimed at transferring the meaning and results of the whole process to audiences. They also identified feedback loops from the products of transformation to the less structured information supporting them, which are essentially re-evaluation, and searching for various types of supporting information. The lower part of the diagram is identified as foraging loop in which the focus is on collecting and extracting relevant information. The upper part of the diagram is considered as sensemaking loop, which focuses on developing and adding structure to the collected information.

The sensemaking loop for intelligence analysis describes many of the activities that are not in the core of sensemaking, but serve it or are served by it and interact with it.

In collaborative settings, almost all of the activities mentioned in the two

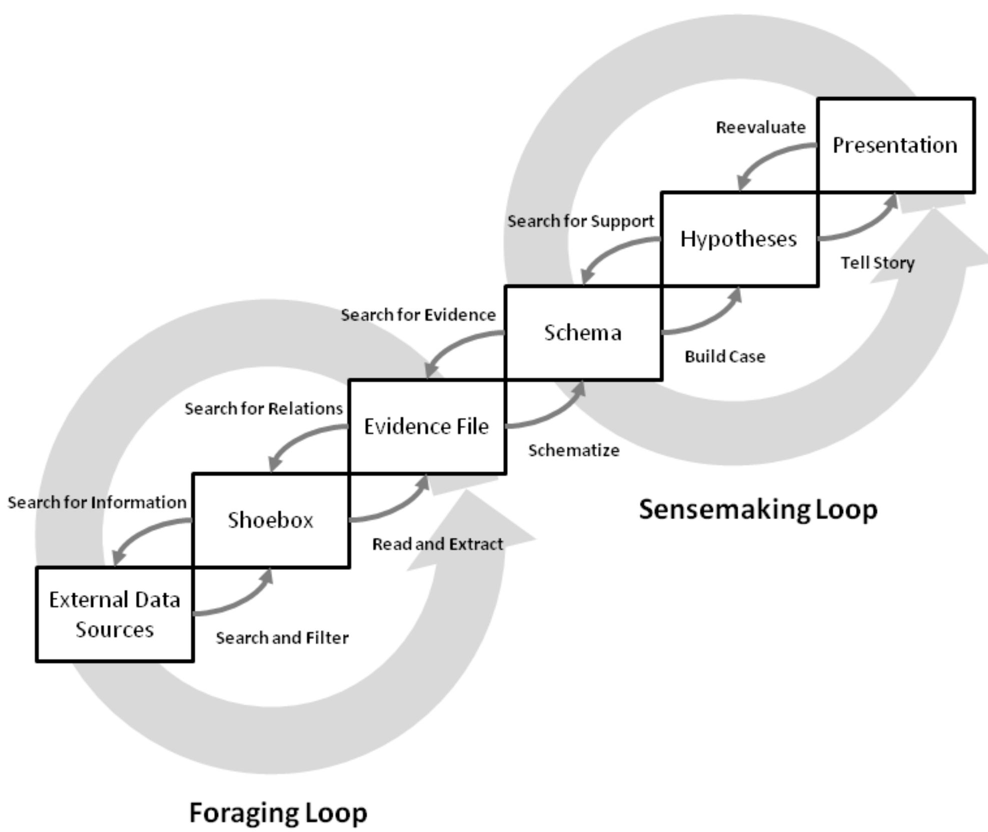

Figure 2. Notional model of sensemaking loop for intelligence analysis

models can be conducted by several analysts and the collaborative system should support the division of labour and the integration of efforts to enable an efficient process of collaborative sensemaking. In a distributed collaborative analytics system, analysts usually prefer to work independently while trying to contribute to their shared goals. Therefore, the distributed analytics system should minimize its assumptions on the specifications of the collaboration that analysts go through. Analysts may want to work on common information spaces and common activities, or they may want to focus on different activities that are related. In either scenario, the system should support efficient integration of their efforts with providing awareness and planning mechanisms.

In the next section, we briefly explore the design space of asynchronous collaborative systems to understand the related concepts in the computer-supported collaborative work literature that can be used as the basis for designing an asynchronous distributed collaborative visual analytics system.

3.2 Asynchronous Collaboration

Computer-supported collaborative work technologies are often characterized by CSCW matrix (Johansen, 1988) separating synchronous from asynchronous communication (time dispersion) and distributed collaborators from non - distributed collaborators (space dispersion). Each of these subspaces imposes different challenges on communication and coordination mechanisms that participants can use in their collaborative process. Although the focus of this research is on asynchronous distributed subspace, we believe that real-world collaborative systems should be able to support all modes of collaboration or provide mechanisms for integration with other systems that support other modes of communication. For example, analysts may login to the system at the same time, and they should be able to take advantage of this opportunity and engage in synchronous collaborative processes or they may happen to be in the same place at the same time or different times and their platform for collaboration should enable them to take advantage of these opportunities. This issue showed up during the case study, reported in chapter 5. An asynchronous collaborative environment should deal with the challenge of establishing the common ground, as the analysts are not able to closely engage in communication to ask for clarification, whenever required. The notion of least collaborative effort (Clark & Wilkes-Gibbs, 1986) emphasizes the economy of grounding and implies that the participants in a collaborative task try to minimize the collaborative effort. However, in an asynchronous setting, it is easy to under-estimate or over-estimate the amount of effort that is required to convey meaning and due to the lack of immediate request for clarification, the communication tends to be inefficient. The principle of least collaborative effort implies that collaborators tend to underestimate the effort required for communication of meaning. This can lead to creation of reasoning artefacts that are hard to comprehend. Artefacts that are hard to understand are less probable to be reused as their cost of reuse is high and collaborators may prefer to develop the artefact from scratch or they may simply ignore it if it does not seem to be crucial for their task.

3.2.1 Designing an online community

Online community is defined in various ways throughout computer-supported collaborative work literature and several candidate definitions and aspects of it are reviewed in (Preece & Maloney-Krichmar, 2005). Preece & Maloney-Kricmar's preferred definition considers "the *people* who come together for a particular *purpose*, and who are guided by *policies* (including norms and rules) and supported by *software*" as an online community. In this thesis, we focus on supporting distributed communities of analysts, which involves asynchronous and different-place mode of collaboration. Another aspect of our approach is de-emphasizing the meaning of the term *purpose*, allowing a dynamic and flexible framework for it in the definition of online community. We believe that despite the possible differences in the purpose of engaging in the community, the ad-hoc interactions that occur as people share their resources in the system lead to temporary shared purposes that create a dynamic online community. These features are rather common in communities of practice in which a group of like-minded professionals engage in loosely coupled collaborative processes to promote their understanding and support each other (Wenger, 1999).

Our research is aimed at providing a digital infrastructure for more efficient large-scale analytical collaborative systems such as open community contribution systems and virtual communities of practice, based on Bos et al.'s taxonomy of collaboratories (Bos et al., 2007). A *collaboratory* is defined as "an organizational entity that spans distance, supports rich and recurring human interaction oriented to a common research area, and fosters contact between researchers who are both known and unknown to each other, and provides access to data sources, artefacts, and tools required to accomplish research tasks”. A taxonomy of collaboratories is suggested by (Bos et al., 2007), to describe and differentiate various types of scientific collaborative settings. The definition of a collaboratory greatly describes some of the aspects of the purpose of this research. Although the term collaboratory is defined to convey the various aspects of scientific collaboration, other types of analysts in different contexts (e.g. journalists, intelligence analysts, etc.) can greatly benefit from engaging in various types of collaborative structures, similar to various types of collaboratories. These collaborative structures can be greatly supported by web-based systems.

A distributed collaborative system should support a spectrum of collaborative processes, starting from the point in which users do not know each other, to the point that users explicitly and consciously interact with each other to achieve shared goals. The communication and awareness mechanisms that the system provides for various stages of collaboration, influences the productivity of each stage as well as the transition phases between the stages.

The social interactions in a collaborative system define and maintain the community and collaborative systems can play the role of a social catalyst (Karahalios & Donath, 2004) and facilitate interactions and relationships that had not been easy to initiate or maintain. Some of the important factors in determining the effect of the collaborative system on the users are, granularity of contributions, visibility of contributions, roles that users may take and authority and role structure (e.g. hierarchical vs. flat, or informal vs. formal), privacy of activities and contributions, profile information, group making and group management mechanisms, multiplicity of communication mechanisms, etc. Considering that focusing on all of these factors is overwhelming and not practical in the scope of this thesis, we decided to eliminate group making and group management and role structure from our design space explorations to focus on granularity, visibility and privacy of contributions while having an eye on some other influential factors such as awareness and communications. This decision was made to facilitate focusing on addressing the main goal of this project, namely facilitating the reuse process, as granularity, privacy, and visibility of contributions can greatly affect the discoverability of the contributions, and the motivations for sharing and reuse of them.

3.2.2 Incentives

Various incentive mechanisms are used to ensure sustainability of online communities. Sustainability, in this context, is defined as being able to self-organize and retain members and keep them participating and contributing to the system so that the community as a whole makes progress toward its goals (Ellis, Halverson, & Erickson, 2005). Although there are various reasons for implementing incentive mechanisms such as encouraging people to join the community, to stay in the community or to be active in particular ways, we focused on encouraging analysts to share their reasoning artefacts and reuse products of each other's analysis, whenever practical. Sharing and reuse are two inherently different types of behaviour and different factors can play role in fostering them, which are discussed further in the following sections.

3.2.2.1 Motivations for Sharing Reasoning Artefacts

Sharing resources in an online community can be considered as an altruistic behaviour (Antoniadis & Le Grand, 2007) that requires spending time and energy, however it could be motivated by self-interest in implicit rewards such as self-esteem or reputation caused by what other people might think or by fear of being treated the same way (Axelrod, 2006). Moreover, online community designers can provide mechanisms that increase the benefits that participants achieve for their contributions. These benefits may enable the participants to take actions that they otherwise could not, or they may be valuable because the community appreciates them. The fist type is known as instrumental incentives, and the second type is known as symbolic incentives (Ellis et al., 2005). However, as we decided to eliminate the complexity of authority, roles and management structure from our design space explorations, using instrumental incentives was not practical anymore. We implemented a wiki-like open community management structure in which every analyst could independently create and update resources or participate in both planning and reasoning activities.

Therefore, we implemented two symbolic incentive mechanisms through enhancing visibility of participants' behaviours (also known as social translucence (Erickson, Halverson, Kellogg, Laff, & Wolf, 2002)) and enabling peer oversight (Cosley, Frankowski, Kiesler, Terveen, & Riedl, 2005).

3.2.2.2 Motivations for Reusing Reasoning Artefacts

The inherent motivation for reusing resources is the reduction of the amount of work required for accomplishing a task. However, the reuse process imposes several costs and poses various risks, which degrade its benefits. Among the costs of reuse are the costs of discovering a useful piece, cost of understanding it and finally cost of integrating it. Moreover, reuse is an act of trust and if the required trust-based platform is not available, the cost of validating information will be added, or not validating the information poses risks to the analysis outcome. Section 3.3 describes the reuse process in more detail to provide insights into designing mechanisms for lowering the various costs associated with the reuse process.

3.2.3 Privacy Levels in Collaborative Systems

The concept of privacy has been defined in several ways in social sciences literature. Parent (1983) reviewed several of the definitions proposed over years and after criticizing all of them, he proposed a new definition that describes privacy as “the condition of a person's not having undocumented personal information about himself known by others” or in other words, ”the absence of undocumented personal knowledge about a person”. However, this definition does not capture nuances of this concept. The definition suggests that if personal information is documented somewhere, it does not matter if others know it or not. This issue is challenged by recent privacy researchers such as Danah Boyd. She introduced the concept of privacy and security through obscurity (Boyd, 2008), which implies that one can put his personal information in a public or shared space, while maintaining some level of privacy. She illustrates her point by an example: Imagine that you are screaming to be heard in a loud environment, if someone comes closer, he could overhear you. Nevertheless, you do not care because, you probably cannot see the person, it is not abnormal to be overheard, and what you were saying probably does not really matter to him. The obscurity provided by the environment, creates some level of privacy and security. She used this concept to analyze privacy settings in Facebook and underscore the difference between something being public versus something being announced to others.

The concept of privacy by obscurity suggests that privacy is not just about existence of information in a public place, but also it is strongly interrelated with the availability of information to others and their awareness, and the amount of attention that it attracts. The design of collaborative systems should carefully consider the interactions between these concepts and variables to provide clear mechanisms that enhance discoverability and availability of information and foster sharing of the potentially useful information while not violating collaborators' privacy.

3.3 Reusable Reasoning

Analysts develop their analysis by building upon other analysts' findings and analysis outcomes; however, reusing other analysts' analysis outcomes requires considerable effort, and the process of summarization and making sense of one's contributions repeats over and over again. Moreover, sharing the outcomes of the analysis process is usually postponed to the end of the process or specific milestones, many times the analysis results do not reach to presentable states and many of the intermediary results that are not directly related to the main goal of the analysis are not recorded appropriately.

Several benefits are associated with the sharing of ideas and outcomes in the early stages of analysis including early identification of flaws in the analysis process (e.g. misconception, cognitive biases, overlooked evidence, etc), refining the ideas, accelerating the data collection process and avoiding the missing of the important sources of information, etc.

In this section, we draw from the related literature in software engineering and cognitive science as well as computer-supported collaborative work to elucidate the various aspects of reuse process.

3.3.1 Reuse Process Model

Various reuse process models have been proposed to decompose the reuse process and make it easier to approach. Basili suggests a four-step process (Basili, 1990):

1. Identifying the candidate reusable pieces from an old project
2. Understanding them
3. Modifying them to our needs
4. Integrating them into the new process of project

However, this model does not cover the creation of reusable pieces. Sumner and Dawe identify a cycle of reuse by observing the reuse of scientific research outcomes among researchers (Sumner & Dawe, 2001). The reuse cycle starts with creating a reusable resource and sharing it, and continues with discovering the resource by other users or the same user at another time, understanding it, integrating the resource in the current task and ultimately sharing the outcome as a new resource that can be a starting point for another cycle of reuse. Frakes and Terry reviewed the metrics and models that have been suggested for software reuse (Frakes & Terry, 1996). Some of them are cost/productivity models that determine the cost effectiveness of reuse based on its costs comparing with development costs, and reuse maturity model that determines the maturity of reuse process in organizations based on various factors. Moreover, they identify the reuse failure modes:

- No attempt to reuse
- Part does not exist
- Part is not available
- Part is not found
- Part is not understood
- Part is not valid
- Part cannot be integrated

These modes can be associated with complications at various stages of the reuse cycle suggested by Sumner and Dawe. We used the reuse cycle model as the primary way of looking at the reuse process in this thesis. This model is a more general model that matches well with other more specific reuse process models. This model as a model that is grounded in observation of researchers' reuses process can be well applicable to other analytical processes that other types of analysts go through.

This model can clearly show that each of the steps should be possible and easy to complete, to enable the whole cycle to work. In order to design a system that fosters and facilitates the reuse process, we should consider the following:

- Providing a shared space as a platform for sharing of artefacts
- Encouraging users to share their artefacts
- Providing easy to use discovery mechanisms such as browsing and searching
- Encouraging users to create artefacts that are easy to understand
- Encouraging users to create artefacts that are easy to integrate
- Providing mechanisms for evaluation and validation of artefacts
- Encouraging users to reuse artefacts
- Providing mechanisms for integrating artefacts

Collaborative visual analytics systems should provide incentives and mechanisms that can satisfy these considerations. In chapter 5, we will explain how various features of AnalyticStream aim at satisfying these requirements.

3.3.2 Reusable Pieces of Analysis

Reuse process has been extensively studied in software engineering literature as a way of accelerating development process by taking advantage of algorithms, code and logic developed in the previous system development processes and avoiding redundant efforts. Currently the most significant architectural paradigm that addresses this problem is Service Oriented Architecture (SOA). SOA is an architectural style for building distributed systems based on interacting loosely coupled, coarse-grained components (Rotem-Gal-Oz, Bruno, & Dahan, 2007). To understand how SOA can inform our design decisions, we would define the mapping between concepts in collaborative analytics and software engineering: Pieces of analysis can be considered as the *components* that are supposed to be integrated by the analyst to shape the analyst's representation of the situation or problem at hand. The whole representation can be considered as a *system*, being built based on the distributed components. Inspired by the principles of SOA, our design process tries to address the requirements of analysts' collaborative analysis and sensemanking.

3.3.3 Levels of Reuse

The reuse process can occur at different levels and our design process aims at facilitating it at least in two levels: the first one is the level of reasoning artefacts such as analysis outcomes, evidences, partial causal networks, etc. This level of reuse helps users to extend or deepen their analysis based on other analysts' expertise through understanding and integrating the other analyts' pieces of analysis. For example, a researcher may cite article A without having actually read it only by reading article B that has discussed a relevant part of article A. Although this practice is typically discouraged in academic world, analysts in various fields usually have hard time verifying every piece of information that is collected or produced by other analysts and tend to trust them unless they have specific knowledge to the contrary. An example of a slightly higher-level reuse can be citing Greenberg and Buxton's paper (Greenberg & Buxton, 2008) in CHI papers with an unconventional evaluation. This type of reuse involves generalization based on several artefacts and building an abstract concept (i.e. "A CHI paper with an unconventional evaluation").

Reusing the analysis process and methods is a high-level reuse that we are envisioning. CZSaw (Kadivar et al., 2009) is one of the pioneers of the movement toward this goal by the automatically recording of analysts' interactions in the form of scripts and allowing storing, replaying and modifying them.

The reuse of the reasoning processes can help analysts take advantage of each other's expertise to better understand how to use the well-known analysis techniques. In addition, analysts can share their customized crafted techniques and processes that can be effective in specific domains and situations. Through these processes, *research patterns* or *reasoning patterns* can be recognized and reused in a large scale. Reasoning patterns are repeating structures of reasoning artefacts that can serve well for a common purpose. Reasoning patterns can be defined in various levels. For example, analogy, recursion, and divide and conquer can be considered as high-level reasoning patterns while counterbalancing to avoid order effects in experiments is a more specific pattern.

In addition to facilitating learning by example, we hypothesize that through supporting high-level reuse processes, the analysis techniques can be augmented, fine tuned and mixed with each other to create more effective analysis processes.

3.3.4 Types of Reuse

When an analyst reuses a piece of analysis, at least five scenarios may arise:

1. The analyst reuses the piece as it was designed. Analyst wants to use it as an evidence outside of his field of expertise/focus and he does not want to change it, however the analyst may
    a. Accept that the creators of that piece update it whenever needed. (Dynamic Copy) or
    b. Want the piece to remain as it was at the time of reuse. (Static Copy)
2. The analyst reuses the piece after slightly changing it , so that it better represents the purpose of the analyst (Shallow Appropriation)
3. The analyst reuses the piece after significantly changing it in a way that it does not represent the same information anymore (Deep Appropriation). In this type of reuse, the analyst reuses the contents of

the piece to create a new piece, which is related in some way to the original piece.

4. The analyst reuses the piece by creating another piece based on that or decomposing it to several pieces, so that it becomes easier to reuse (Refactoring)
5. The analyst reuses the piece and as it is relevant to his/her current analysis process, he/she may want to update the piece as he progresses and the creators of the piece may or may not accept his/her updates (Shared Copy)

System designers should design mechanisms for supporting these scenarios and decide if they want to explicitly support some of them in the user interface.

3.3.5 Refactoring

We believe that, collaborative analytics systems should not disturb analysts' original flow of analysis, to make sure that the productivity and creativity of the analysts will not be negatively affected. Paying attention to reusability while producing a reasoning artefact can facilitate future reuse and the collaborative system should foster it. However, inspired by the agile software development principles (Cockburn, 2001), we believe that the original disassembly and advance planning for reuse of the artefacts are usually inadequate and need further efforts before reusing.

Refactoring is a software engineering concept, which is defined as the process of redesigning the abstractions in a program. We can recontextualize this definition by replacing “program” with “reasoning artefact”. Refactoring per se is a prominent analysis task, which may comprise disassembling, restructuring and designing or redesigning abstractions of a reasoning artefact. Knowledge refactoring processes often happen when an analyst restructure or interpret another analyst’s reasoning to better support or integrate them with his contributions. In addition, analysts can create new representations of pieces of analysis. In addition to the crucial role of refactoring in enabling and fostering reuse, it makes the analysis process easier to understand and helps the analysts to detect the possible flaws of the analysis process.

In sum, the refactoring and reuse process can happen at various levels from reusing the pieces of analysis to finding common structures and reusing the higher-level abstractions of them such as analysis patterns. Different levels of refactoring and reuse processes can be used for increasing the coverage and depth of analysis, learning analytical methods and augmenting them, identifying new analysis patterns, and improving the quality and comprehensibility of the reasoning artefacts.

3.3.6 Case based Reasoning

Cognitive scientists and artificial intelligence researchers have been studying case-based reasoning or reuse of problem solving and reasoning for decades. Case-based reasoning or reasoning from experience assumes a memory model for representing, indexing, and organizing past cases and a process model for retrieving and modifying old cases and assimilating new ones (Slade, 1991).

Case-based reasoning assumes that memory as the container of experiences is predominantly episodic and richly indexed. It also assumes that memories and experiences are connected to each other in various ways and the organization and structure of memory changes over time. Based on this paradigm, people interpret new situations in terms of prior experience and use their experiences in prior situations to guide them in the new ones. Finally, based on principles of case-based reasoning learning occurs when an expectation from a previous case fails to predict a new situation.

Schank proposed a theory of reminding and several knowledge structures for representing episodic information. These structures include scene that defines a setting and the sequence of action that take place in that setting, scriptlet as a sequence of actions in a scene, memory organization packets that organize scenes, and thematic organization points that capture similarities between situations that occur in different domains (Schank, 1983, 1999). Based on this theory we process events as they happen and automatically look for particular episodes in memory that are closely related to the current input, we are processing. Schank’s ideas can be extended, in the light of Distributed Cognition Theory, to be used as a systematic reuse process for distributed cognitive systems such as teams of analysts who collaborate with each other to make sense of shared information spaces.

In the next chapter we depict the design space of distributed collaborative visual analytics systems through the theoretical lenses of Distributed Cognition Theory and Activity Theory. We attempt to extend our understanding about individuals' cognition such as the reminding process mentioned above, to collaborative structures like teams of analysts.

4: EXPLORING THE DESIGN SPACE OF DISTRIBUTED VISUAL ANALYTICS ENVIRONMENTS

Designing a distributed collaborative visual analytics system requires a comprehensive understanding of an enormously vast design space expanded through several areas of cognitive science, social science, and computer science.

In order to understand this complex design space, in this chapter, we use Distributed Cognition Theory and Activity Theory to analyze and elucidate various aspects of designing asynchronous collaborative visual analytics systems. Integrating the perspectives resulted from these theoretical lenses, provides collaborative visual analytics system designers and researchers with a rather comprehensive view of the design space, and sheds light on some of the less explored areas of the design space. Based on this integrated perspective, we present suggestions and design guidelines for the development of asynchronous collaborative visual analytics systems aimed at facilitating collaborative sensemaking processes.

4.1 Making Sense of the Design Space from Multiple Perspectives

As was stated by Simon (Simon, 1996), "...solving a problem simply means representing it so as to make the solution transparent". In order to gain a better understanding of the design space of distributed visual analytics systems and the type of artefacts that influence analysts' interaction with information spaces, we can investigate and represent the design space from multiple perspectives. In this chapter, we investigate the collaborative analysis process and the design space of distributed analytics systems through theoretical lenses of Distributed Cognition Theory and Activity Theory. Both of the theories have been influential as theoretical frameworks for human-computer interaction research (Hollan et al., 2000; Victor Kaptelinin & Bonnie A. Nardi, 2009) and the Distributed Cognition Theory has also been suggested as a theoretical framework for information visualization(Liu et al., 2008).

We start with the Activity Theory (Engeström, Miettinen, & Punamäki, 1999) that has more rhetorical power than the Distributed Cognition Theory, as it names the constructs that should be considered in exploring the design space. We use Activity Theory to analyze the collaborative analysis process as an *activity* aimed at making sense of an information space as its *object* (according to the terminology of Engeström's version of Activity Theory (Engestrom, 1987)). Activity Theory's perspective helps us to, identify the original influential artefacts of collaborative analysis process such as analytical methods, and identify the opportunities for augmenting them or designing new artefacts (e.g. user interface elements) as mediators to improve or facilitate the interactions between analysts and the problem space. In addition, the requirements and possibilities for supporting the division of labour between analysts are discussed to understand how the design of a distributed analytics system may support the decomposition of the analysis process and enable implicit collaboration (e.g. through using shared reasoning artefacts) as well as explicit collaboration during the analysis process.

Then, we apply the Distributed Cognition Theory (Hutchins, 1996) to investigate how analysts as human actors, and both reasoning artefacts and user interface elements as cognitive artefacts take part in the cognitive processes to form a distributed cognitive system. We furthermore assess how analytical skills and reasoning artefacts at different levels such as evidences, causal relationships and hypotheses can be created and transferred through the system. Additionally, we investigate how elements of the user interface can take part in the cognitive processes. Ultimately we try to answer how distributed analytics system can be designed to support the coordination and cooperation of the distributed actors and artefacts in collaborative analytical processes. We will elaborate on the cognitive processes that can be distributed among analysts and cognitive artefacts to envision how this distribution can be supported by a distributed analytics systems.

Ultimately, we draw design guidelines and suggestions based on our understanding of the design space based on each of the two above mentioned theories and by integrating them into our design guidelines whenever practical.

4.2 Activity Theory Perspective

Activity Theory was initiated by Vygotsky (Vygotsky, 1978), further developed by his disciple Leont’ev (Leont'ev, 1978), and later by Engeström who has provided a framework for understanding a phenomenon in its context at multiple levels. Activity theory considers an activity system as the unit of analysis (Engeström et al., 1999). Engeström depicted a graphical model of activity system (Figure 3) to reveal the salient components in an activity system. In this model, *subject* refers to the individual or group that the analysis is performed from their point of view. The *object* refers to the problem space upon which the subject is acting. The transformation of object to a desired *outcome* motivates the activity. The transformation is performed with the help of external and internal *mediating instruments and artefacts*. The *community* comprises the groups who share the same general object. The *division of labour* refers to both the horizontal division of tasks between members of the community and vertical division of power and status. Finally, the *rules* refer to the explicit and implicit regulations or norms that constrain interactions within the activity system (Engeström et al., 1999).

Activity triangle helps to expand our field of view in analysing the design space through careful consideration of each of the components and their effects

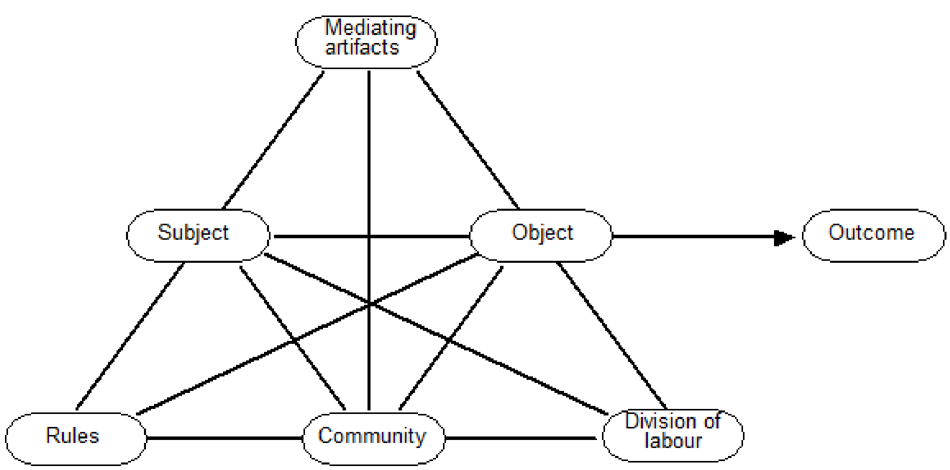

Figure 3 Activity triangle for analyzing an activity system

on the activity. Looking from the lens of Activity Theory to the design space of distributed visual analytics systems, analysis process can be considered as an activity, in which analysts try to understand an information space through applying various analytical methods and with the help of various representations of information. The activity system of analysis process can be described by determining its salient components.

4.2.1 Subject, Object & Outcome

Subjects comprise users of the system, and object of the analysis process is the information space that the analysts need to investigate and understand. One of the main categories of outcomes of the analysis process is analytical outcomes that is the documented tangible results of the analysis process that can be shared with other analysts and decision makers (J. J. Thomas & Cook, 2005). These analysis products are prepared to address a requester's need, which may have various types such as assessment of a situation, forecast of possibilities, or development of options, and comparative assessment of their implications and effectiveness. Each of these types can take different forms in various contexts; for example, assessment of a situation can be interpreted to literature survey in scientific research context, or making sense of a set of interrelated events occurred in the past in the intelligence analysis context. Forecast of possibilities can be interpreted as estimating threats, vulnerabilities, and opportunities in intelligence analysis context, or forecast of human behaviour, natural phenomena, future technologies, trends, and etc. in scientific research context. The development of options can be interpreted as determining and comparing possible ways of defending against a particular threat in intelligence analysis context, or as a comparative analysis of several design alternatives in a design situation, or several treatments for a medical problem in scientific research context.

Regardless of the type of product, the analysis outcome is the result of restructuring the information space and re-representing the knowledge structures through various reasoning processes.

In general, analysis process may have several other outcomes that are not visible through the lens of Activity theory, as it only considers the transformations of the object of activity (information space) as outcome. For example, another outcome that can be achieved through analysis process is gaining experience and expertise that is more significant for novice analysts. If we also consider collaborative analysis processes, the social aspect of the analysis process may result in other outcomes such as new levels of trust, reputation, and credibility among analysts.

4.2.2 Mediating Artefacts

From the Activity Theory perspective, artefacts mediate human thought and behaviour. In a distributed analytics system, at least three categories of artefacts are identifiable which are discussed in the following sections. Although we tried to focus solely how Activity Theory looks at the mediating artefacts, we believe our understanding of them is also influenced by the Distributed Cognition Theory.

4.2.2.1 Representations of information

Representations of information provide or facilitate a set of cognitive operations that define how humans interact with an information space (Card, Mackinlay, & Shneiderman, 1999; Zhang, 1997). They also determine which information can be perceived and how it can be structured and processed. Various aspects of visual representations and their benefits to analysis process are discussed in (J. J. Thomas & Cook, 2005). However, in a distributed analytics system, representations that support communication become prominent. For example, the narrative form is a type of knowledge representation that has been particularly effective for human communication and knowledge transfer (Haven, 2007). If we compare a well-thought narrative representation of a complex reasoning artefact to a well-designed visualization or a causal network that represents the same information, while the causal network shows the relationships clearer, the narrative has several characteristics that supports transferring the embedded knowledge with less cost (less cognitive load and in a shorter time). It is clear, in a narrative, where to start and what flow to follow, where to pay attention to.I It is easier for its author to create context and finally the purpose and meaning of the narrative is easier to extract. In addition, similar to narrative comprehension, narrative generation seems to be among the earliest powers of mind, which allows us to organize specific kinds of complex knowledge structures such as our experiences, in the form of stories for communicating them (Bruner, 1991). To achieve benefits of visual representations and the ease of knowledge transfer of narrative representation it is desired to combine the two representations in various forms (Schnotz, Bannert, & Seufert, 2002), e.g.

visualization and caption, annotated visualization, or visualization embedded in text.

4.2.2.2 Analytical methods

Analytical methods are the procedures or activities that analysts use to deal with their information space. They may be considered as representations of the analysis process that are chosen by the analysts based on their task and the goal of the analysis. For example, analysis of competing hypotheses (ACH) is an analytical method that helps judgment of issues requiring weighing of alternative explanations by making the analyst to think about the alternatives in a specific way to minimize the effect of cognitive biases of analysts (Heuer, 2005). Similar to data representations, analytical methods mediate the interactions of subject (analysts) and object (information space). Many of the visual analytics systems such as Oculus SandBox (Wright et al., 2006) support several analytical methods. However, the relative level of support for various representations may affect the outcomes of the analysis process; that is the ease of applying some of the analytical techniques, biases analysts toward using them. In many cases, analytical techniques serve different purposes, however in the cases that the application of methods overlap, the choice of analytical methods may change the quality of the process. This implies that the design of visual analytics tools should include and provide similar levels of availability and ease of use for the analytical methods that their users might use. That can be interpreted in two ways; the first one is that, if a visual analytics system is supposed to be used as a comprehensive toolbox for analysts, it should very well support as many techniques as possible; or the second interpretation that can be more practical is that, system designers should strive toward designing and developing open extensible architectures (Birsan, 2005; Wolfinger, 2008) that can easily accept new analytical methods or representations in the form of plug-ins, or service-oriented architectures (Rotem-Gal-Oz et al., 2007) that can be easily integrated with other systems to enable analysts to use various representations and analytical methods. Using these architectures is especially important for distributed analytics systems in which several analysts with different needs, conventions and points of focus and tasks try to take advantage of each other's products. The choice of analytical methods and representations needs to be matched with the analytical tasks (Tolcott & Holt, 1987), thus, various groups of analysts need to use different sets of analytical representations and methods to perform their tasks.

4.2.2.3 User interface elements

User interface elements determine how analysts access and manage information and which information is more accessible. Activity Theory points to a set of the concepts to keep in perspective during the design process, and usually it does not assist in further elaboration about the design decisions. However, the Distributed Cognition Theory helps to think about how artefacts and user interface elements affect the cognitive processes. We will take a closer look at user interface requirements from that perspective.

4.2.3 Community, Division of Labour, and Rules

We consider users of a distributed analytics system as the subjects of the activity. With this assumption, the community will include at least four groups that share the same object (information space): 1. Other users of the system that investigate information spaces that intersect with the information space under investigation of the subject, 2. Other analysts that investigate those information spaces, without using the system, 3. Analysis consumers such as decision makers that interact with those information spaces, and 4. Suppliers of raw information for the information space.

Various aspects of communication and collaboration within a collaborative visual analytics system are discussed in (Heer & Agrawala, 2007; Viegas & Wattenberg, 2006). Activity Theory on the other hand, helps to think of a broader community that can affect the design decisions. For example, if we think of other analysts that are not already using the system, the following design issues and implications become apparent. How can we convert them to users of the system? How can we enable them to use the reasoning artefacts that are being generated within the system, and facilitate this process? How can we enable users to take advantage of their outcomes and facilitate this process?

System designers should be aware of the rules and conventions of the interacting groups. For example, a distributed analytics system should support importing information from various possible suppliers of information and other analysts. In addition, the system should facilitate the conversion of reasoning artefacts to representations that are understandable by consumers of analysis process such as decision makers and other analysts.

Integration with community of analysts, data providers and decision makers, and facilitating the conversion of outside analysts with users of the system can be considered as two of the important aspects of designing distributed analytics system. In a distributed analytics system, analysts may have different points of view and different backgrounds and conventions. Considering these factors implies that the system should facilitate the cross boundary cooperation by fostering the use of common terminologies and common representations, and support the creation of boundary objects, and conversion between the representations and terminologies (Bannon & Bødker, 1997)

. For example using strategies such as Tagraph (Nobarany & Haraty, 2009), for encoding the mapping between terminologies can be beneficial for making artefacts of one group discoverable for another group.

Another aspect of distributed analysis process is the division of labour. Distributed analytics systems should support methods that enable efficient and flexible division of labour. Analysts should be able to take part and accept roles in loosely coupled collaborative processes for performing an analysis task. It requires the distributed analytics system to maintain awareness of what other collaborators have done, or are willing to do.

4.2.4 Activity Hierarchy

Based on Activity theory, interaction between human and the world is organized into functionally hierarchical levels, mainly activity, actions, and operations. Various descriptive studies have decomposed the analysis process in various domains into a set of actions such as searching, extracting information, schematizing, building cases and telling stories (Pirolli & Card, 2005). Hierarchical decomposition of analysis process from analysts' perspective can help organizing the user interface based on the position of user's current action in the hierarchy and predicting the useful user interface elements. Moreover, finding a mapping between user interactions and the hierarchy of analysis process can help to generate more meaningful interaction histories that support intuitive navigation.

4.2.5 Internalization and Externalization

Activity theory emphasizes the constant transformation between external and internal activities as the basis of human cognition. Visual analytics tools facilitate the transformation of analysts' internal understanding of the investigated areas of the information space to external representations that can reveal the bigger picture of the information space. They also facilitate the formation of external representations such as visualizations that bring forth the relevant aspects of the information. Analysts externalize the connections between pieces of information and entities. Analyst's interaction with the visual representations and discovering patterns and new relations between pieces of information, is an example of iterative process of internalization and externalization.

4.2.6 Reflection on applying Activity theory

The main benefit of using the Activity Theory is to broaden our perspective about the factors that should be considered in the design process and to help us to ask meaningful questions (V. Kaptelinin et al., 1999). Activity Theory provides some inferential power by highlighting the essential components and enabling the designers to predict the consequences of their design decisions based on their effects on various components of the activity system. Kaptelinin's Activity checklist derives practical guidelines and more detailed design considerations from Activity theory; however, in my opinion, the checklist items stray too far from the original principles of the Activity Theory, so that they may be derived from (and even seem to be influenced by) various other related theories such as the Distributed Cognition Theory.

An interesting aspect of the Activity Theory is the focus on the role of community in shaping an activity. It greatly points out how the community around a subject imposes rules and norms on the subject and how the community can be organized around an object through division of labour. However, it is not clear how we can separate the subject from the community, when dealing with one or multiple groups of subjects. Various types of relations can be seen inside a subject group, or between different groups of subjects or between subjects and the outside community that are affecting the activity. Both power and weakness of the Activity Theory lie in its broadness, which requires more effort from the designers to refine it and appropriate the concepts for the design situation under consideration.

In the next section, we use the Distributed Cognition Theory to understand how various cognitive processes can be supported in a distributed analytics system.

4.3 Distributed Cognition Theory Perspective

From the Distributed Cognition Theory perspective, a cognitive process is delimited by the functional relationships among the elements that participate in it, rather than by the spatial collocation of the element (Hutchins, 1996). A cognitive process can be distributed across people, through time, or across internal and external representations. In this section, we demonstrate how the Distributed Cognition Theory can inform and shape the design process of distributed visual analytics systems. Looking through the lens of the Distributed Cognition Theory, a distributed visual analytics system is a cognitive system consisted of several cognitive resources that take part in various cognitive processes.

4.3.1 Cognitive Resources and Actors

Analysts and their mental resources usually constitute a significant part of distributed cognitive systems, and are responsible for most of the computations and coordination of cognitive resources including themselves in the organization. They also act as part of the distributed memory of the system. However most of the data encoded in their memory, especially in the long-term memory, should be externalized and backed-up in various external forms of reasoning artefacts (e.g. notes, reports, visual causal networks, recorded hypotheses, etc.), to be easily stored, shared, searched, retrieved, and reused by all the actors in the organization.

Reasoning artefacts as sharable cognitive resources play a major role in the way analysts think about collaborating with each other. Granularity of reasoning artefacts affects the frequency of sharing and collaboration between analysts. For example if we consider a report as a reasoning artefact, it will greatly reduce the possibility of close collaboration, comparing to considering a short note as a reasoning artefact. Reasoning artefacts can be as large as a full report of a long-term analysis or as small as a relation or an inchoate idea. Another aspect of sharing a reasoning artefact is the certainty of its producer (or forager) about its validity and reliability and the distributed analytics system should foster the sharing of artefacts at different levels of reliability and convey the certainty of their producers. Fostering the externalization of uncertain hypotheses and ideas, helps analysts to reflect on their analysis process and advance or refine their ideas and hypotheses further.

Another major element of this cognitive system is the user interface that enables analysts to interact with each other and with their reasoning artefacts. Affordances of the user interface, affect how closely analysts can work together and how the information can be transferred through the system and how the reasoning artefacts can be externalized and represented by the analysts. The power of external and internal representations is greatly discussed in (Norman, 1994). Reasoning support tools should equip analysts with the suitable external representations for their tasks. The differences in availability and ease of using representations biases analysts toward using the easiest and most available ones, which may affect the quality and outcome of the analysis process.

The representation of the reasoning artefacts transforms the cognitive processes that they are involved in. For example, a node-link graph representation of a causal network affords understanding the transitive causal relations while a narrative representation better affords communicating causations with the aid of language and explanatory clarifications on how events or artefacts are related or caused by each other. Various visual and non-visual representations of reasoning artefacts help analysts to perform certain processes easier. The variety of different representations’ affordances makes it desirable to support dual or multiple coding of reasoning artefacts in collaborative visual analytics systems, where the goal of representation is both reasoning and communication of the reasoning artefacts.

4.3.2 Cognitive Processes

In a distributed cognitive system, cognitive actors collaboratively engage in cognitive processes to achieve various goals. Cognitive processes are the processes, by which sensory input is transformed, reduced, elaborated, stored, recovered, or used (Neisser, 1967). Learning, abstraction, reminding and remembering, computation, attention, reasoning and problem solving are some of the cognitive processes that may happen in a cognitive system. In a distributed visual analytics system, the basic goal is to support distributed collaborative problem solving. This goal consequently requires many other cognitive processes to be distributed. In the next sections, we discuss some of the cognitive processes that can be distributed in a distributed visual analytics system.

4.3.2.1 Memory Processes

Memory of a distributed visual analytics system is distributed among analysts and the software system. The distributed memory should be accessible by all the analysts to support the sharing of reasoning artefacts. The system should support various processes that are associated with memory such as memorizing, remembering, reminding and forgetting. Various cognitive architectures have considered different memory components such as long-term and short-term memory, implicit and explicit memory, semantic and episodic memory. Based on the requirements of the design situation, designers can take advantage of the different meories. For example, in a collaborative system it is desirable to record (memorize) all of the user interaction in the system to facilitate distributed reasoning, however the system may run out of space. Therefore, forgetting mechanisms should be designed to handle the unbounded interaction logs. The design process can be inspired by human memory architecture and processes; for example based on the Schema Theory of Cognition, forgetting occurs when certain pieces of information get integrated and lose their individual identity (Ausubel, Novak, & Hanesian, 1978). In a distributed visual analytics system, this integration may occur over several analysts' interaction histories through time, which can lead to a very efficient yet effective mechanism for keeping track of events and interactions over a long period. Another benefit of this process is that after a while, the system may become able to identify frequent chunks of interactions that are meaningful, which can facilitate navigation of history in a hierarchical organization.

For designing a distributed analytics system, supporting the storage and retrieval of events or user interaction sequences (similar to episodic memory) and reasoning artefacts (similar to semantic memory) is crucial, as it is the only way for the analysts to share, reuse and integrate the outcomes of their analysis.

The most important function of a distributed analytics system is facilitating the sharing and reuse of reasoning artefacts and analytical processes. Distributed Cognition Theory defines culture as the process of cumulating partial solutions to frequent problems. This definition implies the need for storing these partial solutions as well as reusing them whenever beneficial. The basic support for sharing of reasoning artefacts can be memorizing or storing them in a distributed cognitive system and every analyst can remember the memorized artefacts by searching for them or browsing through the system. However, this process may be ineffective when the users are not aware of the items that they have memorized. To generate a query, analysts must know enough to know what is not known and which of the many unknowns can be understood from the other analysts' outcomes in the distributed analytics system. There is no easy way for finding out about all the relevant abstractions and building blocks if you do not even know what to ask (Fischer, 1987). A significant portion of analysis process is dealing with a vague situation model, which makes it near impossible to ask the right questions.

This challenge can be addressed to some extent by looking from the distributed cognition perspective. As we mentioned earlier, sharing an item can be considered similar to the process of memorizing an item in the distributed cognitive system; therefore, reusing it requires remembering it. There are two types of memory processes for remembering: The first one is actively trying to remember something that we are already aware of having it in our memory, and the second one is being reminded about the artefacts and processes that are related to the current context, which happens all the time in our daily lives. Designing a reminding process in a distributed cognitive system may significantly facilitate the reuse process by addressing the aforementioned challenges. We can use Schank's theory of reminding (Schank, 1999) to understand the reminding process and inform the design the of distributed analytics systems. Based on this theory, when we process events as they happen, we need to find specific episodes in memory that are closely related to the current input, however we do not consciously look for them, because we are not explicitly aware of their existence. In a distributed visual analytics system, the system can assist by reminding the analyst about other analysts' reasoning artefacts that are related to the analyst's current analysis process.

4.3.2.2 Attention and Awareness

Attention is a cognitive process that serves various purposes. It helps us to select a target item from distracters in a visual scene, and enables the ongoing awareness of the environment (LaBerge, 1997). Attention can be socially distributed in a distributed analytics system where user interface elements and mechanisms can support it. User awareness in a collaborative system is defined as the way users perceive their collaborators and what they are doing, without direct communication (Dourish & Bellotti, 1992). From the point of view of Distributed Cognition Theory, awareness can be seen as a distributed attention mechanism in which users' behaviour attracts other users' attention to the extent that makes them aware of their behaviour. In a distributed visual analytics system, shared attention among several analysts should be supported. Attention can be shared at different levels: between two analysts, among a group or among all the users of the system. In other words, analysts should be able to share their attention and actively attract other analysts' attention. Analysts as elements of the distributed cognitive system should be able to become aware of the points of attention of the system, their subsystem, or other elements of the system. Awareness helps analysts to better align their efforts toward their common or related goals by participating and sharing their cognitive resources on the points of attention. The need for supporting awareness in collaborative sensemaking systems is mentioned in (Paul & Morris, 2009) as a key factor in group-sensemaking processes.

4.3.2.3 Reasoning processes

Form the Distributed Cognition Theory perspective, reasoning processes can be distributed over internal and external representations and over people and through time and a distributed visual analytics system can support each of the three types of distribution. Visual analytics systems inherently require the distribution of human reasoning over internal and external representations. Visual representations are designed to transform the reasoning task by representing the information and the task in a domain where the answer or the path to the solution is transparent. The social aspect of sensemaking processes has recently come under spotlight by systems such as (Viegas & Wattenberg, 2006; A. B. Viégas et al., 2007) and specially (Heer et al., 2009) that identified several patterns of social sensemaking. Also several visual analytics systems has started to focus on the distribution of analysis process through time by providing history recording and history navigation mechanisms (Heer, Mackinlay, Stolte, & Agrawala, 2008; Kadivar et al., 2009; Robinson, C. Weaver, & Center, 2006). It is hard to think of the distribution of reasoning processes over time, without having distributed memory processes. However, if the reasoning process is being recorded rather than just the products of the process, the probability of influencing an analysts' reasoning process will go up. In other words, if we record the internal or external analytical steps that an analyst goes through and let other analysts to replay or change the steps or simply continue them, it may significantly affect the reasoning process that other analysts go through. An example of such process is implemented in CZSaw (Kadivar et al., 2009).

Another type of reasoning process that can be supported, is reasoning about certainty and reliability of reasoning artefacts. Reuse is an act of trust and analysts are more willing to reuse the reasoning artefacts that they believe are valid and reliable. For example, researchers tend to use findings of other credible researchers or cite the papers that are published in venues that are more credible. It becomes more critical in intelligence analysis domain, where analysts should deal with an information space characterized by uncertainty and inconsistency. A distributed analytics system can use trust-based mechanisms to support distributed reasoning for evaluation of reasoning artefacts.

4.3.2.4 Learning processes

Learning is a cognitive process of knowledge and behaviour acquisition. Based on the Schema Theory, a primary process in learning is subsumption in which new material is related to relevant ideas in the existing cognitive structure. In a distributed analytics system, learning can occur at least at two levels; the first one is learning about the information space, which is apparent and is a primary goal of using a distributed analytics system, and the second is learning analytical processes. The second type of learning can be especially useful for novice analysts, but also it can accelerate the process of creation, evolution and evaluation of new analytical methods in analysts' community. Learning by example is one of the typical ways in which people learn by inferring general rules from examples. A distributed analytics system can support this method, by making the intermediate products and the analysis process of analysts available to each other. This can be achieved by keeping track of processes, and persisting of the relationships between reasoning artefacts as described in (Pike et al., 2007). Analysts should be able to understand how reasoning artefacts are generated by tracking a script-like history of analysis or manually tracking "rich products" (Pike et al., 2007) that are linked to other artefacts that can represent how and why they are produced and how valid they are. Through this process, analysts may learn about interesting patterns of reasoning and reuse them in their own analysis process. People can construct new knowledge based on their experiences through discussions by selecting and transforming information (Palincsar, 1998). A distributed analytics system can also support collaborative learning through discussions and arguments between analysts, which can lead to clarifying relations in the information space under investigation, and help them to refine their analytical reasoning processes.

4.3.3 Reflection on Applying Distributed Cognition Theory

We believe that Distributed Cognition Theory provides inferential power by pointing out the possible effects of design decisions on the user's cognitive processes. Additionally, some of the principles that are emphasized by Distributed Cognition Theory such as offloading or externalization of cognitive processes can provide designers with rough guidelines for aligning their design decisions with human cognitive processes and the ways people improve their performance using cognitive aids. Distributed Cognition Theory elucidates some of the topics mentioned by the Activity Theory such as the meaning and role of culture in interacting with the environment. While Activity theory puts more emphasis on the role of culture by considering it as a generative force involved in the production of mind, where Distributed Cognition Theory operationalizes this concept by perceiving it as a process that accumulates partial solutions to frequently encountered problems. The two definitions do not closely match, but both are aimed at highlighting the prominent role of culture in our cognitive processes. Another example is explaining the role of representations in the interactions of the user with the information space, which is highlighted by both theories. Activity theory explains them as mediators and points out the internalization and externalization processes to demonstrate their role; however, Distributed Cognition Theory deepens this understanding by analyzing the role of internal and external representations in various cognitive processes and understanding how they can transform these processes.

4.4 Summary of Suggestions

Activity theory and Distributed Cognition Theory provide an extensive view of the design space of distributed visual analytics systems. However, to avoid re-iterating the previous efforts for introducing design considerations such as (Heer & Agrawala, 2007), we only focus on the less explored areas of the design space that are emphasized by the two theories, to complement the previous studies.

4.4.1 Integration with community

4.4.1.1 Enabling the integration with data providers

Analysts need to acquire data from various sources of information and the tool designer should be aware of norms and conventions of the data providers' community to support seamless integration of the efforts of the communities. In other words, analysts should be able to use various types of data without having to deal with the various inconsistencies. Additionally data providers should be able to get feedback from analysts to align their data collection efforts with the analysts' needs. That is, the strategies for data collection and surveillance should be informed by analysts' needs or concerns and both sides should support the interoperability.

4.4.1.2 Enabling the integration of efforts with the community of analysts and decision makers

Analysts should be able to take advantage of the products of other analysts' tasks.This implies that designers should be aware of the neighbour domains' conventions and data structures and facilitate the process of cross-domain or cross-organizational collaborations.

The two guidelines (4.4.1.1 and 4.4.1.2) lead to designing multi-layer artefact structures (e.g. separation of data and representation) to support communication of reasoning artefacts in various representations and formatting. Another aspect of them is that distributed analytics systems should provide communication-minded representations to enable cross-expertise and cross-domain communication.

4.4.2 Designing open extensible architectures

The uneven support for representations and analytical methods biases analysts toward using the most available ones. Using the right representations and right analytical methods can improve the performance of analysis process and visual analytics systems should enable analysts to extend them and appropriate them for their analysis processes. Supporting seamless integration with other visual analytics tools, to extend the set of available representations and analytical methods is only possible through using open extensible architectures and designing analysis platforms rather than embedding the support for a fixed set of representations and methods.

4.4.3 Supporting the division of labour

The division of labour may affect the design of collaborative visual analytics systems in two ways. The first one is to delimit the system based on possible components of the activity under consideration. For example various sensemaking models (Pirolli & Card, 2005; Russell et al., 1993) or information visualization models (Card et al., 1999) can be used as a reference for dividing the support for activity into interoperable services. For example, based on Pirolli and Card's sensemaking model (Pirolli & Card, 2005), designers may decide to support some of the sensemaking processes such as "Search and Filter" and "Read and Extract" processes while enabling the interoperation with other systems that support other sensemaking processes.

Additionally, division of labour can be considered as a process within a distributed analytics system. Distributed visual analytics systems should help analysts align their efforts for making sense of intersecting information spaces perhaps toward shared goals and that is possible through both implicit and explicit division of labour.

4.4.3.1 Supporting awareness and shared attention mechanisms

Implicit division of labour can be facilitated through providing shared attention and awareness mechanisms. Analysts should be able to track each other's efforts to reuse each other's products and avoid redoing. Additionally, analysts should be able to attract each other's attention to their products and processes as a method for initiating or enhancing collaborative work. From the distributed cognition perspective, the software system and its users form a distributed cognitive system, in which the relationships between users and the way users contribute to each other's cognitive processes is largely determined by the communication and awareness mechanisms that the collaborative system provides.

4.4.3.2 Supporting explicit collaboration

Distributed analytics systems need to support a range of collaborative processes from independent analysis processes (completely decoupled) to loosely coupled collaborative analysis processes and perhaps closely coupled processes. That means that analysts should be able to make the transitions between modes of collaboration by inviting each other, forming teams and explicitly planning the analysis process.

4.4.4 Enabling the distribution of memory and reasoning

Various memory processes such as memorizing/storing, remembering, reminding, and forgetting of entities and reasoning artefacts can be distributed across analysts, and this distribution should be supported by distributed analytics systems. Models of human memory can be used to analyze each of these processes to find clever ways for distributing them. For example, analysts can be reminded of each other's reasoning artefacts, which can foster the reuse process. Reasoning artefacts and the reason and process behind them should be persisted in the system to help analysts reuse them whenever practical. That is analysts' interactions should be recorded and presented to the analysts for adding metadata and reflection as well as making them easier to understand for themselves and others.

4.4.5 Providing a personalizable hierarchical taxonomy of analyst interactions

Inspired by the hierarchy of activity in the Activity Theory, we believe that having a customizable hierarchical taxonomy of user interactions that matches the units of interaction from analyst's perspective can facilitate the generation of meaningful activity histories. Consequently, analysts can better make sense of their analysis process and reflect on that and perhaps extract patterns of reasoning or discover potential errors.

This hierarchy can be derived from various sensemaking models such as (Pirolli & Card, 2005; Russell et al., 1993) or information visualization reference model (Card et al., 1999); however, those models should be customized to match with the system and the type of actions that the system supports and also users should be able to customize it to meet their needs. This hierarchy can also support the division of labour by determining the type and granularity of tasks (actions) that analysts can perform to contribute to a shared objective.

5: ANALYTICSTREAM: A DISTRIBUTED ANALYTICS SYSTEM

We designed AnalyticStream, a proof of concept distributed visual analytics system that adopts some of the proposed guidelines presented in the previous sections. AnalyticStream is an evidence filing and marshalling environment in which analysts bring in their pieces of analysis from various visual analytics tools to schematize them, build cases and tell stories to communicate their analysis and disseminate their products. The main difference of AnalyticStream with the previous evidence filing and marshalling systems such as Oculus SandBox (Wright et al., 2006) and Xerox PARC's Entity Workspace (Bier et al., 2006), is in its focus on fostering and facilitating the sharing and reuse of reasoning artefacts. Human reasoning is an expansive resource and facilitating the reuse process can help dealing with it as it deserves. We can multiply the effectiveness of analysis process by ensuring that the reasoning and the expertise used for it will be transferred to other problems, whenever possible.

Analysts can use AnalyticStream to communicate their reasoning process and outcomes and take advantage of the shared pieces of analysis, produced by other analysts, and integrate them with their analysis results to validate, extend or refine them.

As explained in section 3.3.1 the reuse cycle starts with creating a reusable resource and sharing it, and continues with discovering the resource, understanding it, integrating the resource to the current task and sharing the outcome as a new resource (Sumner & Dawe, 2001). A piece of analysis is reusable if it is discoverable and understandable by other analysts. Producing reusable pieces of analysis requires analysts to attach cues for enhancing discoverability such as tags, and contextual information for enhancing comprehensibility, to the pieces of analysis. For example, a diagram with labelled axes can be easy to reuse for the analyst who has produced it, by other analysts need a description and contextual information to be able to understand it and

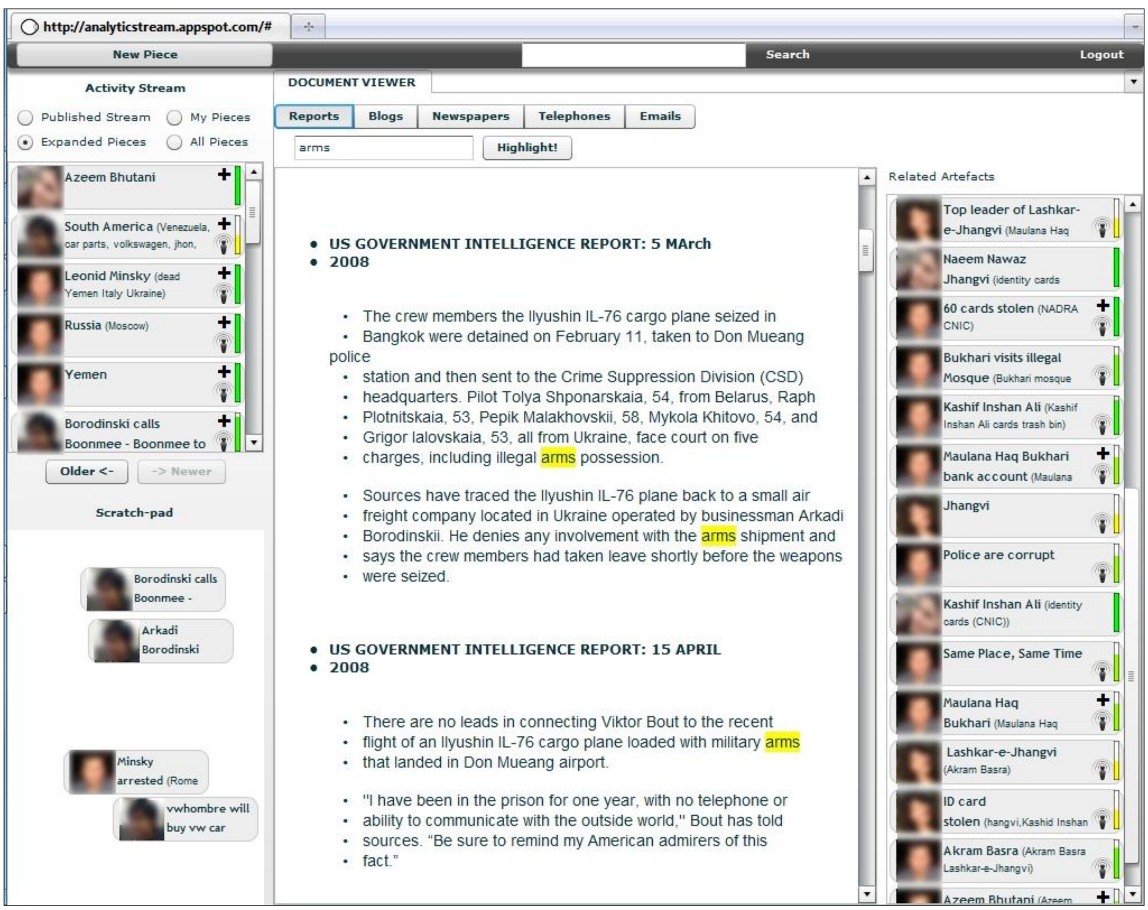

Figure 4. Initial AnalyticStream user interface: supporting asynchronous collaboration for making sense of VAST 2010 challenge dataset

reuse it.

Considering the great ability of people for organizing information in the form of narrative and understanding narratives, narrative form is selected as the main organization structure in AnalyticStream. Analysts can create stories or cases by bringing pieces of analysis from the collection of pieces collected by various analysts, marshal them to convey their line of reasoning, and create new relations and pieces to enhance the story line.

AnalyticStream is implemented as a web-based system with Flash as frontend and Java as backend, and runs on Google App-Engine runtime environment. Figure 4 shows the main user interface of AnalyticStream, being used for analyzing VAST 2010 challenge’s dataset. The main user interface includes Activity Stream panel and Scratch Pad on the left and Document viewer and StorySpaces tabbed view that has filled the rest of the screen. Each of the components of the user interface is explained in the following sections.

5.1 Pieces of Analysis

A piece of analysis is the basic unit of analysis in AnalyticStream. A piece of analysis is consisted of a title, rich text contents and possibly an image. A piece of analysis has an arbitrary semantic role (fact, assumption, entity, relation, etc.). Moreover, a piece of analysis can be expanded to a sequence of pieces of analysis that may describe or explain it. This simple data structure was aimed at providing maximum flexibility and fostering appropriation.

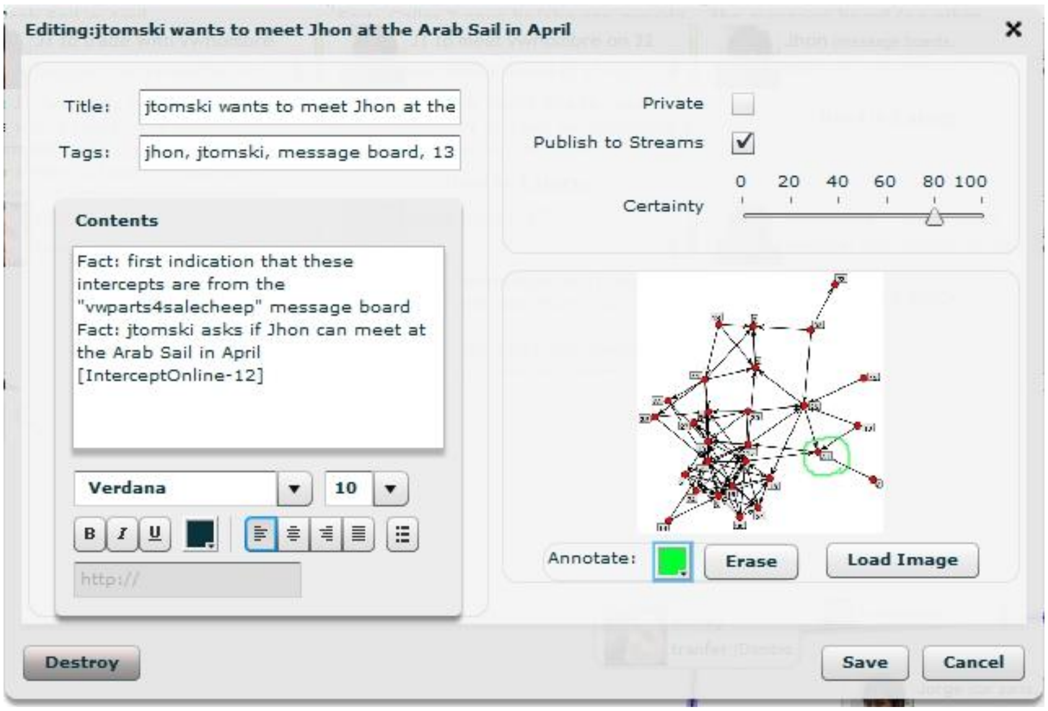

Figure 5 User interface for creating and editing a piece of analysis

Visual representation of a piece of analysis (Figure 5 and Figure 6) in AnalyticStream comprises its producer's profile picture, title, visual indicator of its producer's certainty about the piece (dual coded in length and colour of the indicator), HTML contents and possibly a separate visualization that can be annotated. Contents of a piece are in HTML language and a rich text editor was used to facilitate formatting of contents.

A piece of analysis may have one of the following three privacy states: Private, Public, or Published. Private pieces are only visible to the analyst that has produced them. Public pieces are visible to all of the analysts, but they do not show up in the Activity Stream panel (unpublicized). Finally, published (or publicized) pieces are visible to everyone and show up in the Activity Stream

Figure 6 Visual representation of a piece of analysis

panel. We consider the Published state as a privacy state, because it is announced in the stream, which is an instance of removing privacy by obscurity. For a more detailed discussion on this matter, see section 3.2.3.

5.2 Story Space

Each piece of analysis can be expanded to an expanded piece or a story and a StorySpace (Figure 7) represents the expansion of the piece. The expansion of a piece could have various meanings depending on the intention of the analyst that creates it. StorySpace uses a multiple-coordinated view approach to enable both sequential marshalling and relation-oriented interaction with the information space. A StorySpace includes a narrative view or a sequential view of the pieces, a graph view for representing the relations between the pieces inside the story and other pieces, a History panel showing the history of pieces of the story, comments panel showing the comments on

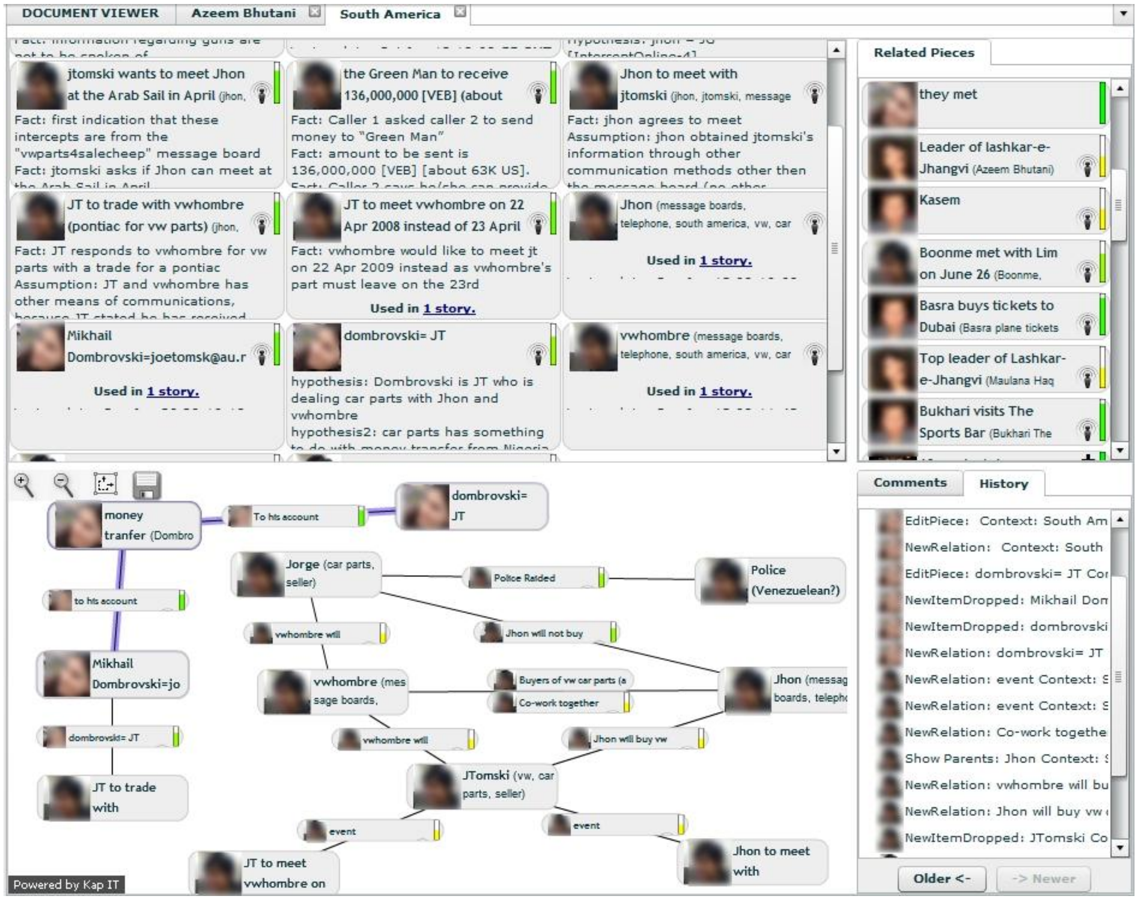

Figure 7. StorySpace represents the expansion of a piece of analysis ("South America" in this example) and includes narrative view, graph view, related pieces panel, history panel and comments panel.

each of the pieces and the related artefacts panel. Each of these components is explained in the following sections.

5.2.1 Narrative/Sequential View

Stories are one of the most successful knowledge sharing structures (Haven, 2007), and the narrative view (Figure 8) facilitates the construction of narrative-like sequential organization of pieces of analysis. Fostering the creation of a narrative view for pieces of analysis is aimed at creating a clear flow of reading for the sensemaker and facilitating the comprehension process that is a prerequisite for reusing pieces of analysis.

Analysts can drag and drop pieces of analysis to the narrative view from outside or inside of the narrative view, to build a case or a story. The simplicity of the knowledge structures enables users to use them for various purposes and we were interested in seeing what the possible applications of the narrative view are.

The visual appearance of narrative view is inspired by comic books, as the efficient use of space, easy comprehension and close coupling between text and visualization have been their shared goals. To avoid layout problems and to facilitate the navigation of a story, the size of panels is fixed and by clicking on a panel, it expands to show the rest of its contents.

We decided not to explicitly support some of the reuse mechanisms such

Figure 8. Narrative/Sequential view is tiled view of pieces of analysis, that represents an expanded piece of analysis.

as shallow appropriation to simplify the user interface as much as possible. Moreover, because all of the analysts have been analyzing a shared information space, the appropriation scenarios were of less importance. However, we analyzed the user interface to make sure that all of the scenarios are possible. The default mode of reuse was using dynamic shared instances of reasoning artefacts. That is, all of the pieces could be changed by all the analysts. However, analysts could also create new appropriated or easier to reuse pieces based on them. Moreover, analysts could define relations to determine if a piece is derived from another piece. Another mechanism that is considered for shallow appropriation is putting a piece in context, by surrounding it with other pieces that help conveying the desired meaning.

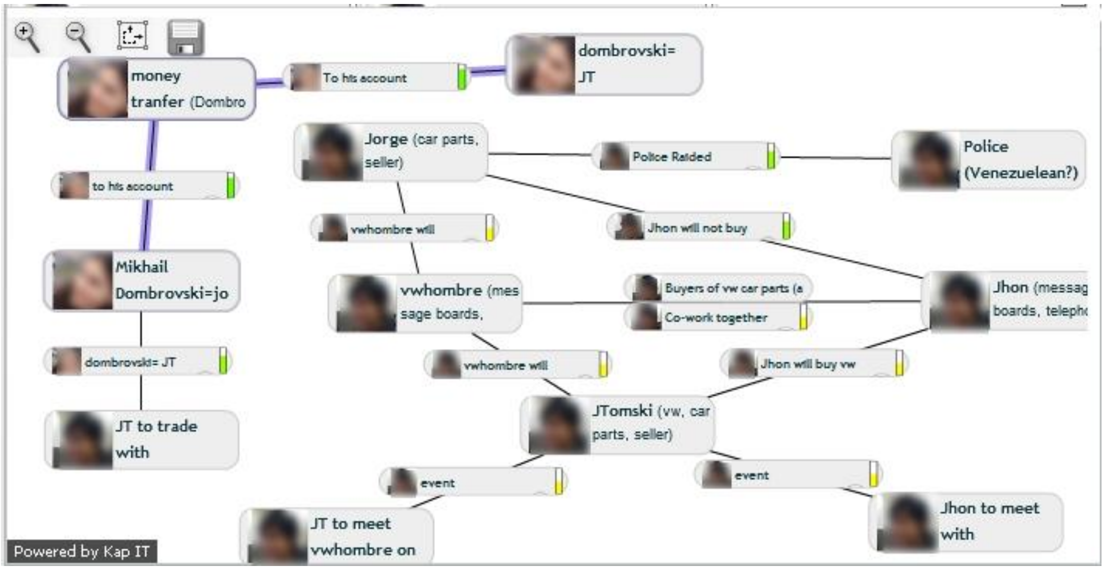

Figure 9. Graph view shows the relations between the pieces of analysis in an expanded piece with other pieces of analysis

5.2.2 Graph View

Graph view (Figure 9) was designed to enable users to interact with the visualization of relations between pieces of analysis. The combination of narrative view and graph view of analytical stories can better support collaborative sensemaking processes; although the comparative evaluation of this hypothesis is outside the scope of this thesis, we try to understand the role of each view through collecting qualitative data about their usage.

5.2.3 Related Pieces Panel

Related Pieces panel (Figure 10) automatically suggests the pieces of analysis that are related to the analyst's workspace, assuming that they might be useful for the analyst. Content similarity and collaborative filtering algorithms have been effective for similar purposes in other suggestion-oriented systems (Cosley, Frankowski, Terveen, & Riedl, 2007). These suggestions are aimed at improving the discoverability of the pieces of analysis, which is another prerequisite of the reuse process. Due to the exploratory nature of the analysis and investigation process, the discoverability is a prominent problem. Usually the situation model is vague and the analyst may not even know what kind of information can be available and helpful for his analysis, to look for them. Therefore, the suggestion-oriented workspace can alleviate this problem by

Figure 10 Based on content similarity, Related Pieces Panel suggests pieces of analysis that are related to the analyst's workspace

unobtrusive suggestions. During the course of our study, we found out that researchers at Eindhoven University of Technology have used recommending relevant reasoning artefacts to facilitate the discovery of connections between notes, visualizations and concepts (Shrinivasan, Gotzy, & Lu, 2009).

5.2.4 History of Analysis

Although we decided not to focus on providing a rich interactive history mechanism, based on our analysis in chapter 3 and 4 we knew that not providing any history mechanism might inhibit analysts to engage in collaborative processes. Providing a history mechanism is crucial for supporting the distribution of cognitive processes over time and people. Therefore, the history

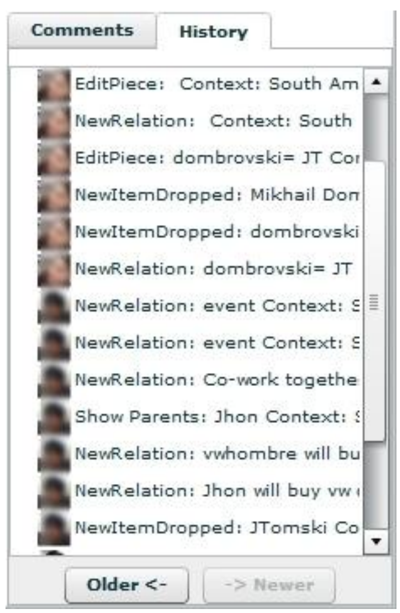

Figure 11 History panel shows the history of analysts' activities involving the stories in the StorySpace

Figure 12 Comments on pieces of analysis can be used for indirect communication between analysts

panel (Figure 11) was designed to show the history of pieces inside a story by determining who has done what to which piece as the minimum data that can be necessary for the analysts.

5.2.5 Comments

Analysts need to discuss about their perspectives and their analysis process to resolve possible conflicts, and clarify ambiguities. They may also want to plan for closer collaboration. All of these issues require distributed analytics systems to support some sort of explicit communication. Despite trying to limit the scope of our proof-of-concept prototype, we decided to support asynchronous communication through putting comments (Figure 12).

5.3 Activity Stream

Activity Stream was designed to provide awareness about analysts' activities and a browsing mechanism for finding pieces of analysis. Being aware

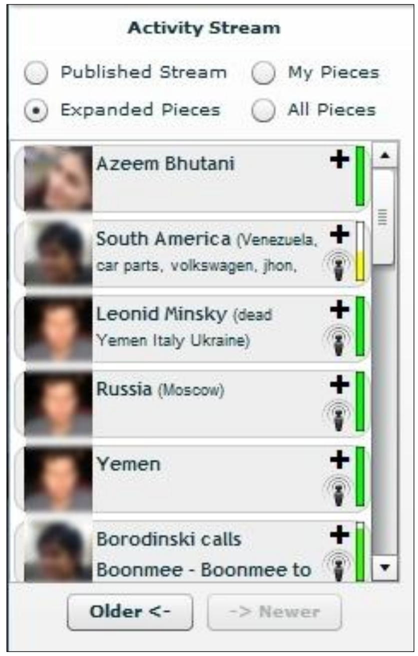

Figure 13 Activity Stream provides awareness of other analysts' recent activities

of the point of attention of other analysts helps analysts to better allocate their cognitive resources (division of labour).

Activity Stream provides four different modes for browsing pieces of analysis including browsing one’s own pieces, all of the analysts’ pieces, expanded pieces and finally published pieces. Published pieces of analysis are the ones that analysts preferred other analysts to be aware of. The default mode of the Activity Stream is showing the published pieces so that analysts could be aware of the latest activities of other analysts.

Figure 14 Analysts can drag and drop pieces from other components into Scratchpad for later use

5.4 Scratchpad

Scratchpad (Figure 14) is a free space that analysts can drag and drop pieces to and from other containers such as Activity Stream, Related Pieces panel, Narrative Views, etc. Scratch Pad supports spatial organization of the pieces and was aimed at providing a personal space like a desktop that analysts could use for various purposes.

5.5 Usage Scenario

In this section, we walk through an analysis scenario to demonstrate more specifically how AnalyticStream supports an analyst. Suppose that an analyst has received documents regarding illegal arms dealing in Africa. He starts with reading documents and whenever he finds an important piece of information or an interesting event, he creates a new piece of analysis and enters its contents, title, and related tags. He also sets his level of certainty about each piece based on his level of trust to the source of information. For example, while reading blog

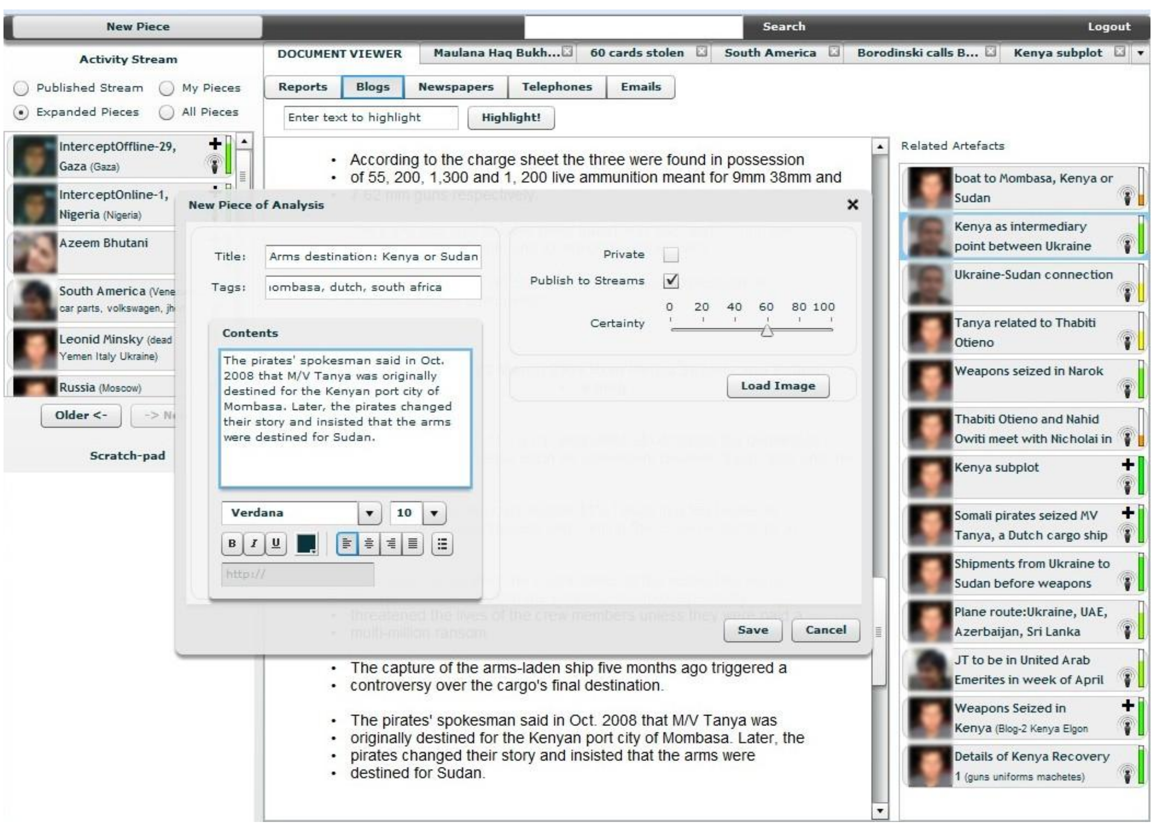

Figure 15 Creating a new piece extracted from the blogs

posts he notices an interesting piece of information about the pirates' leader that changed his saying about destination of a stolen ship: first Kenya, then Sudan. Therefore, he creates a new piece representing it (Figure 15). While reading, he notices that one of the suggested pieces in the Relevant Artefacts panel indicates that Kenya seems to be an intermediate destination, so he adds that piece to his scratchpad (Figure 16).

He keeps on reading, creating pieces and adding suggested pieces to his Scratchpad for possible future use. After having a better sense of what is

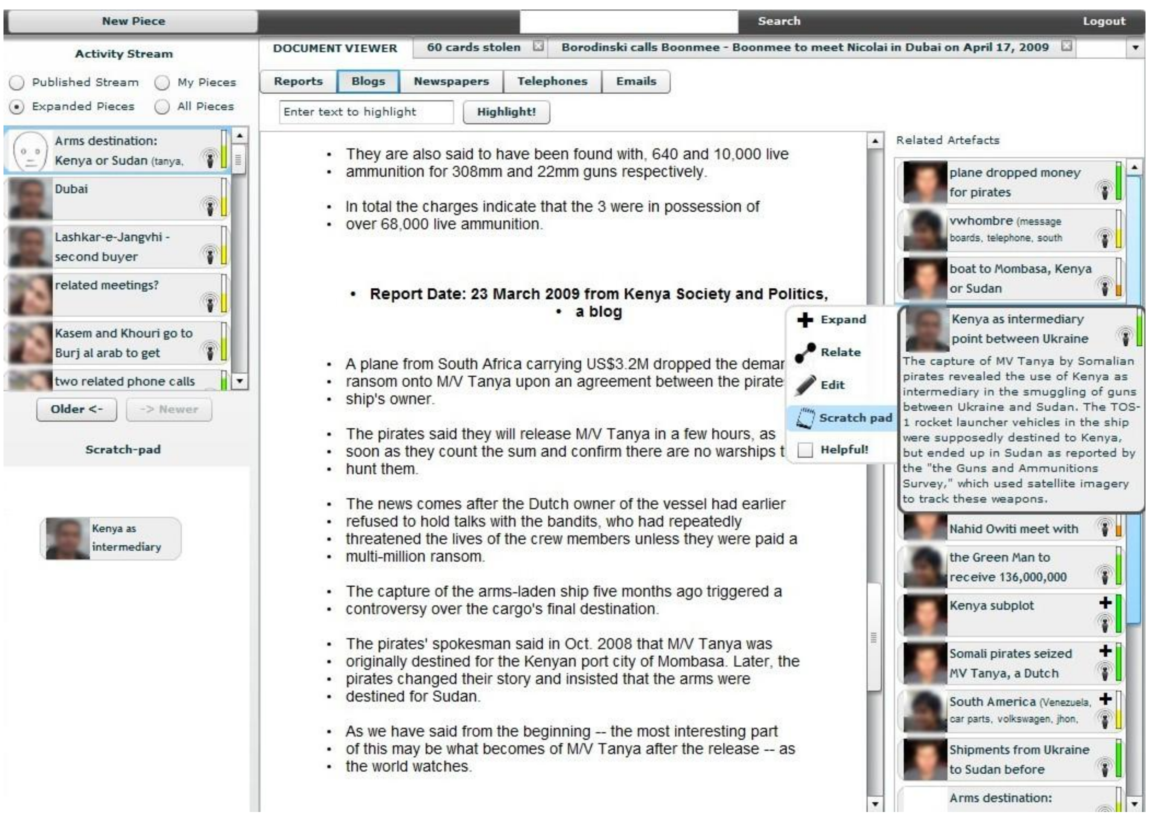

Figure 16 Finding a relevant piece and adding it to the ScratchPad

happening in Kenya, he creates a new piece called “illegal arms in Kenya" for summarizing his findings and expands that piece (Figure 17). While adding his collected pieces to the Narrative view, he notices some new relevant suggestions including more details about pirates, mentioning that they were Somali Pirates. Based on one of the pieces that mentioned the ransom of the ship’s owner to the pirates and the other piece indicating the conflicting assertions of pirates’ leader regarding the destination of the ship, he hypothesizes that the ship’s owner has tried to bribe the pirates to silence them from talking about the route of the ship

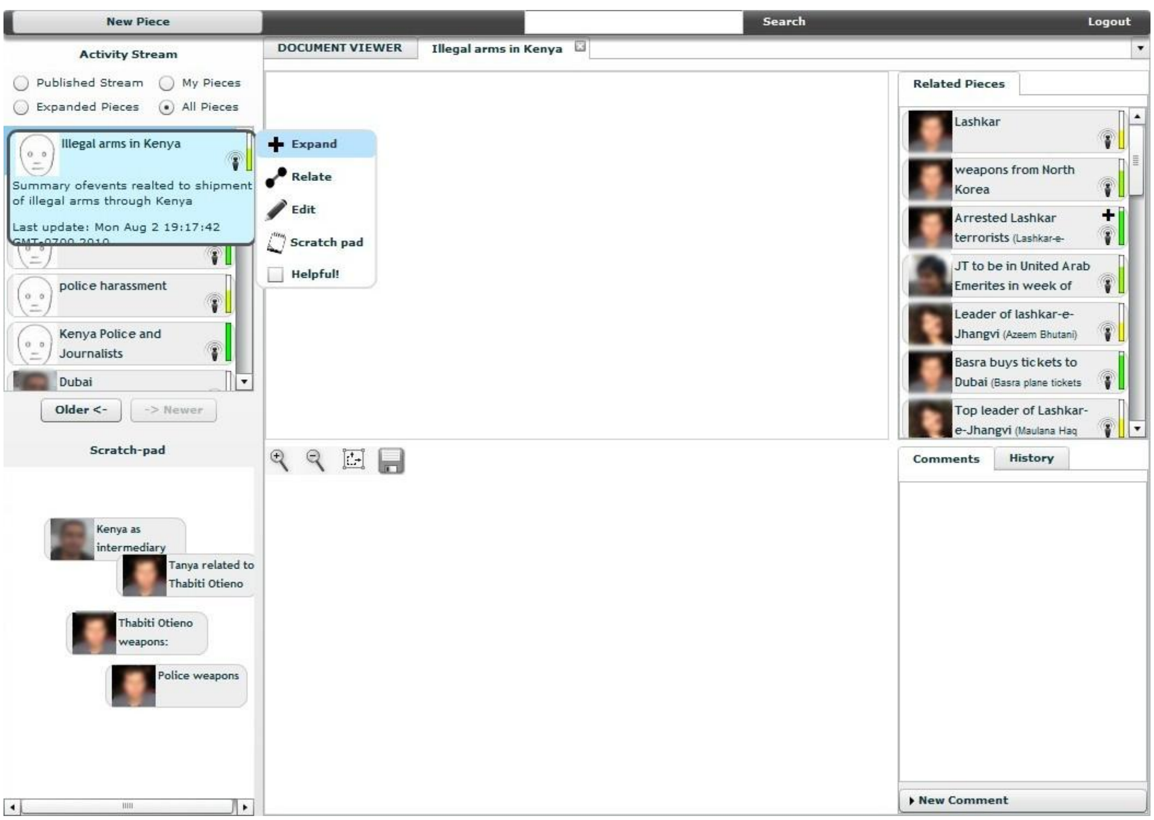

Figure 17 Expanding a piece of analysis

and its cargo. Therefore, he adds a relation between the two pieces to record his hypothesis (presented in the right side of the graph view in Figure 18). Another suggested piece indicates the involvement of Kenyan Police in dealing arms. Therefore, he hypothesizes that Kenya is both a midpoint and a destination of illegal arms and creates a new piece for recording his hypothesis. He also connects this piece to the supporting pieces, to clarify the reason behind his hypothesis and arranges the pieces in the narrative view to better represent his line of reasoning (Figure 18).

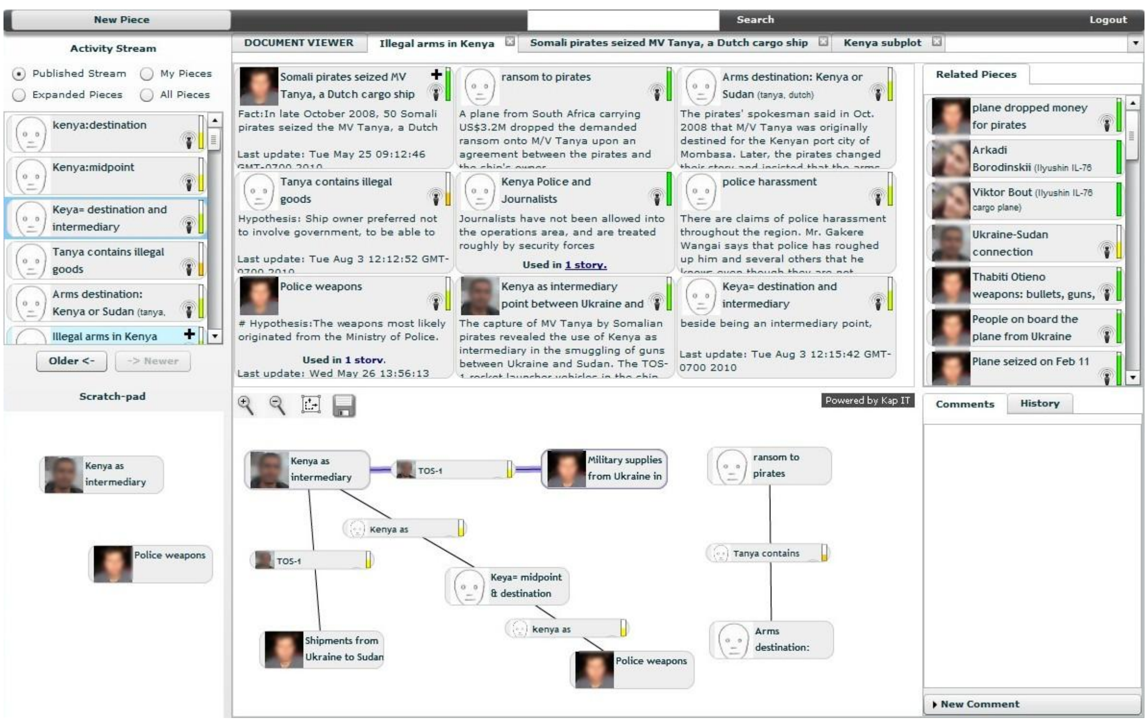

Figure 18 Building a story by dragging and dropping pieces from Scratchpad and Activity Stream and creating relations between pieces

5.6 Implementation notes

In this section, we briefly point out the various architectural highlights of AnalyticStream and explain the limitations of current implementation.

5.6.1 Software Architecture

AnalyticStream is designed as a web application so that analysts can use it whenever they have access to the web. In order to support rich interactions, the user interface is implemented in ActionScript/Flash, and Java/J2EE is used for implementing the backend of the system. Flash user interface and the Java backend are connected using BlazeDS middleware that provides remote object invocation from the Flash user interface to the server-side java objects, and Kap Lab's Visualizer is used for visualizing the graph of relations between pieces of analysis.

AnalyticStream is deployed on Google App Engine (GAE) infrastructure, which is a platform for hosting Java and Python web applications in Google-managed data centres. Using GAE is free up to certain limits (e.g. 1,300,000 requests/day, 1GB/day of outgoing and 1GB/day of ingoing bandwidth, etc.), which were sufficient for the case study evaluation of AnalyticStream. GAE infrastructure is a scalable hosting platform and as the required resources for a web application increase, more resources will be allocated to it and the limit of allocated resources can be determined (and paid for) by the developer of the system; that is developers are able to purchase additional resources whenever needed. However, the automatic allocation of resources by GAE sometimes causes problems for smaller web applications that often require little or no resources. For example, often AnalyticStream was not consuming any resources causing GAE to unload it and this issue increased the average loading time of AnalyticStream and resulted in participants' complaints about performance of AnalyticStream.

Other GAE-related issues that caused minor problems for AnalyticStream were limitations on request time and request size. The size of an HTTP Request should be less than 1MB and all HTTP requests should be responded in less than 30 seconds. Because of the request size limitation, users cannot upload visualizations larger than 1MB. Additionally, because of the limit on request execution time, AnalyticStream could not perform some of the time consuming tasks such as indexing and entity extraction as part of the original request for saving a piece of analysis; therefore, we defined scheduled tasks (cron jobs) to be executed in background for performing the time consuming parts of requests.

In addition, for improving the performance of AnalyticStream, despite using Java Persistence API for object-relation mapping, all of the relationships and foreign keys representing them are handled manually.

5.6.2 Finding Relevant Artefacts and Search Engine

We used Compass library to support full-text search in AnalyticStream. Compass is an open-source library built on top of Apache Lucene, which is a high-performance open-source search engine library. Compass facilitates the integration of search into java applications. Java classes should be annotated to determine the properties that should be indexed and their relative importance in the search.

Analytic Stream used content similarity as a measure of relevance for finding the pieces of analysis that are related to analysts' workspace. To assess content similarity, AnalyticStream extracts the contents of the pieces of analysis in an analyst's workspace and searches for them to find related pieces. The search is automatically performed whenever the workspace is updated, for example adding a piece to the workspace or removing or editing it triggers the search process.

Additionally, we used Stanford Named Entity Recognizer (Finkel, Grenager, & Manning, 2005) to extract the entities form the pieces of analysis and give them more weight in the search process, to improve the relevance assessment.

5.7 Summary

AnalyticStream is a web-based collaborative distributed analytics system that employs several suggestions from chapter 4 and provides the basic functionality required for enabling distributed collaborative reasoning. The design of AnalyticStream is aimed at facilitating and fostering the reuse cycle and to achieve this goal, it attempts to improve the discoverability of pieces of analysis by suggesting related pieces of information. This mechanism can be seen as a distributed reminding mechanism in which the system reminds analysts about their or other analysts' pieces of information that are related to the workspace based on content similarity.

AnalyticStream provides a narrative view for representing the expanded pieces of analysis that can enhance comprehensibility and thus possibility of reuse, by providing a clear flow for reading pieces of analysis and enabling grouping and sequencing them for various purposes.

Small-size representations for pieces of analysis, suggests creating smaller, easier to integrate pieces of analysis. Having several small pieces of analysis rather than one large piece that convey the same information can facilitate the integration of analysts' efforts.

Finally providing a certainty indicator was aimed at encouraging the sharing of less certain pieces of analysis. Analysts can determine their certainty about their assertions to avoid misleading other analysts, while sharing all their findings. Several other features such as determining the number of times that a piece is reused or providing a checkbox for analysts to determine if a piece is helpful for them, and showing the activity stream were aimed at encouraging the sharing of pieces of analysis.

Other features of the AnalyticStream such as comments and activity history are implemented not to impede distributed collaborative sensemaking processes. Based on our analysis in chapter 3 and chapter 4 we knew that despite not focusing on those features, we need to provide a basic way of communication and a mechanism for tracking the history of pieces of analysis to facilitate collaborative sensemaking processes.

AnalyticStream as proof of concept working prototype was designed first to show how some of the suggestions based on the theoretical analysis using Activity Theory and Distributed Cognition Theory could be applied, and second to enable us conduct a case study for understanding the effects of applying suggestions based on a mixed methods analysis of cases. The next chapter describes this study.

6: METHODS

The purpose of this case study is to understand and describe how various design decisions contribute to the reuse cycle and collaborative sensemaking processes. The choice of research methods is usually associated with to the choice of paradigm stance. As a pragmatic researcher, I decided to employ both qualitative and quantitative methods in this study with more emphasis on qualitative analysis. Pragmatic researchers focus on what works and what is available and is not committed to any system of philosophy and reality (Creswell, 2009). Consequently, pragmatic researchers employ various research techniques as necessary, based on practicality and productivity. This study adopted a mixed methods approach, which is a type of "research design in which qualitative and quantitative approaches are used in type of questions, research methods, data collection and analysis procedures, and/or inferences."(A. Tashakkori & C. Teddlie, 2003). Considering that case studies aim for describing one or multiple cases as accurately as possible, the most complete description requires the researcher to seek for all the how, why, what, how many, how often, etc. questions involved in the case, that requires employing a mix of methods.

We used a concurrent embedded mixed methods design (Figure 19) to collect and analyze qualitative and quantitative data. The qualitative data collected through interviews, participants' diaries and usage logs, are the core of the study and the quantitative measurements derived from the usage logs and a

Likert-scale questionnaire have been used to answer more specific questions about usefulness of system's features and validity of design decisions.

The goal of this study was to understand how the design decisions embodied in AnalyticStream can support and foster the reuse cycle and distributed collaborative sensemaking processes. The design of AnalyticStream is aimed at supporting collaborative sensemaking processes in a distributed setting, in which analysts start working independently while they may take advantage of each other's contributions during the process. An important sub-goal of AnalyticStream is to foster the sharing and reuse of pieces of analysis in a distributed setting. The relative effectiveness of information discovery mechanisms (search and recommendation) is investigated using quantitative data from Likert-scale questions and usage logs. Due To our interest in understanding how AnalyticStream supports distributed collaboration and reuse

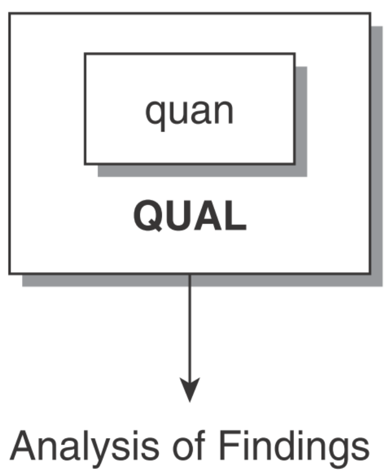

Figure 19. A concurrent embedded mixed methods design with qualitative method as the the primary method and quantitative method as the secondary method is employed.

process, we decided to emphasize on qualitative data and we used the quantitative data to measure variables such as usage frequencies and subjective satisfaction ratings about various features of the system. Qualitative data can also help in better understanding the mechanisms that led to our quantitative findings.

Although a controlled lab study could provide less noisy data for analysis and could enable us to use techniques such as think-aloud which can be helpful in unravelling the analysis process, we decided to ask our participants to conduct analysis in their comfort zones and the times that work better for them to enable them to engage in the analysis process and feel that they are performing analysis rather than attending an experiment. Additionally, collaborative analysis process in natural settings cannot happen in a one or two-hour session and we wanted the participants to actually get involved in analysing the dataset to enable the formation of natural collaboration and collaborative sensemaking processes. We expected that AnalyticStream can foster the sharing and reuse processes and ultimately may lead to more efficient collaborative analysis process and higher quality analysis products; however, we did not expect the effects to be easily measurable and we preferred to focus on understanding processes and characterizing them. We believe that the results of this study can help in identifying measurable variables that can facilitate future quantitative experiments.

6.1 Research Questions

6.1.1 Qualitative question

1. How does AnalyticStream support the distributed collaborative sensemaking processes?

6.1.1.1 Sub-questions

a. How does recommendation of relevant pieces of analysis affect the analysis process (mainly effects on searching, sharing reusing etc behaviours)?

b. How does various design decisions (mainly privacy options, certainty declaration, activity stream, search and recommendations) shape the sharing and reuse behaviours?

c. How is each of the user interface components used and how does each of them contribute to the sensemaking processes?

d. How does the usage of various user-interface components (mainly desktop, narrative view, recommendations panel) and data structures (mainly comment, piece of information, and story) evolve during the course of analysis?

e. What are the shortcomings of the design?

6.1.2 Quantitative questions

The quantitative hypotheses are:

Hypothesis 1: The recommendations are more useful than searching, for finding relevant pieces of analysis during the analysis process.

Hypothesis 2: Narrative/Sequential view of expanded pieces of analysis is more useful for understanding a set of related pieces of analysis, than their graph view.

Also the usage logs (described in section 6.4) provide various descriptive variables about the usage of user interface components, which will be used for finding possible associations between variables and behaviours. Additionally, the descriptive statistics will be used to describe and characterize the analysis process and the effectiveness of various design decisions and support the qualitative analysis.

6.1.3 Mixed methods question

Mixing the quantitative and qualitative results can help us understand how the design decisions and features of AnalyticStream support analysts in their collaboration and analysis process and which of the features of the system are worth further experimentation.

6.2 Participants

A total of 6 subjects, 3 males and 3 females, were recruited by calling for volunteers through emails to research groups and researchers with related areas of interest. All of the participants were university researchers, between 25 and 38 years of age, with science and engineering backgrounds and five of them had experience with visual analytics tasks.

6.3 Procedure

6.3.1 Preliminary preparation

Before the main experiment, all of the participants were asked to answer to a pre-experiment questionnaire (Appendix 1) to provide basic demographic information (age, nationality, course of study, and gender) and explain their experiences with information sharing systems and analytical reasoning tasks. This information was collected to help us describe individual differences in the final analysis, if required. Additionally, participants were trained for 5-10 minutes to use the system and to make sure that they understand all of the features of the system and its user interface.

6.3.2 Experiment procedure

The participants were asked to use AnalyticStream for multiple sessions to make sense of the datasets. We initially started with asking them to use the system for specific amount of time in every session, but later we noticed that because of not being in a lab environment we cannot dictate the details of their participation. Considering that AnalyticStream was designed to support different-time, different-place collaboration, we did not specify any timing for participants' participation and some of them started their analysis process several days later than others. As we primarily focused on collecting qualitative data, it was desirable for us to have participants in different situations. The usage logging component of AnalyticStream that is called AnalyticSensor recorded user interactions (the details are available in the section 6.4.1) that are one of the main sources of data for analysing users' behaviour. In addition, they were asked to fill a semi-structured diary during and after each analysis session. After the experiment period, participants were asked to fill a post-test questionnaire, to reflect on the overall experience and evaluate various features of the system. Finally, based on their responses, we interviewed them to clarify the ambiguous points and validate our understanding (based on the usage logs, the diaries and the questionnaire) about their behaviour during the course of study.

6.3.3 Task

The participants had been asked to provide a forensic analysis of illegal arms dealing. The datasets contained intelligence reports and materials drawn from other sources that were important in their analysis. In particular, we asked them find the hidden plot in the data, which involves summarizing the activities that happened in each country with respect to illegal arms deals, based on a synthesis of the information from the different sources, and presenting their hypotheses. In addition, they needed to analyze the associations among the players in the arms dealing.

6.3.4 Dataset

Participants were given a text corpus including five large documents that contain intelligence reports, emails, blog posts, telephone transcripts, and news articles. They were asked to read and make sense of them and understand the story behind the data. The dataset used for the experiment was synthetic: that is, it was a blend of computer- and human-generated data. The dataset was acquired from Visual Analytics Science and Technology (VAST) contest 2010 (Grinstein, Plaisant, Scholtz, & Whiting, 2010).

6.4 Data Collection

Data collection methods that were used to enable us make sense of users' behaviour are usage logging, participants' diaries, enter and exit questionnaires and a final interview. The following sections explain the benefits of these data collection methods and describe why and how they are implemented in this research.

6.4.1 Usage logs

Analyzing interaction logs is one of the efficient ways of analyzing users' behaviour and human-computer interaction researchers have used it extensively. The data is automatically collected and users do not have to work in a laboratory setting, therefore the results can have greater ecological validity. The major caveat of this technique is the complexity and effort required for analyzing huge amount of usage logs. We used Tableau information visualization system to visualize usage logs. The visualizations were helpful for analyzing patterns of usage and finding interesting sequences of interactions, which are discussed in the section 6.7.

Tracking users' behaviour in web-based systems has traditionally been easier as the server can keep track of users' navigation between pages. However, most of the recent web applications use various technologies such as AJAX, Flash, etc. that enable users to interact with the system without navigating between pages. As AnalyticStream uses flash for supporting rich interactions, we implemented a logging component called AnalyticSensor that captures the context of events as well as the details of the events. Log events may include, *object* of event, which refers to the reasoning artefact involved in the event, *action* that refers to the manipulation that has happened to the object, *analyst* who has performed the action and temporal and spatial settings (X and Y of the start and end point, where applicable) of the event. Spatial settings may be important because the reasoning environment is visual and the spatial settings may provide insights into the analysis process. Therefore, *source* and *destination* of events (the location of the artefact before and after the event) are recorded. However, most of the events do not have spatial properties and this information is only recorded for the Scratchpad and the Graph view. Moreover, the order of elements in narrative view is recorded

Events recorded from the user interactions include:

- The amount of time spent on analysis (we asked users to login before starting their work and log-out when they were done)
- Interaction with components:
    - Interacting with desktop
    - Interacting with Activity Stream
    - Interacting with Hierarchical History
    - Interacting with Comments panel
- Number of countable activities:

- Creating new Pieces
- Creating relations
- Adding/Removing pieces to/from a story and determining the sources of the pieces: from whom and using which UI component
- Putting comments
- Marshalling pieces in the sequential view
- Arranging pieces in the graph view
- Arranging pieces in the analyst desktop

Analyzing usage logs can be a complex task and we used other collected data to understand participants' behaviour. The combination of participants' diaries and questionnaires provides a more comprehensive view of participants behaviour and enables the cross validation of the analysis. (Gerken, Bak, Jetter, Klinkhammer, & Reiterer, 2008; Kort & Poot, 2005)

6.4.2 Participant Diaries

A journal or diary is an "annotated chronological document, created by an individual who has maintained a regular, personal and contemporaneous record" (Alaszewski, 2006; D. H. Zimmerman & Wieder, 1977). Journals or diaries have long been used by literacy scholars as a source of data that is independently composed; but recently this method has been adapted for use by psychologists, sociologists, etc. as a systematic way of prompting individuals to record the details of events (Toms & Duff, 2002). Following this trend, human-computer interaction researchers have used this method to convey workspace or natural settings oriented studies (Czerwinski, Horvitz, & Wilhite, 2004; Hyldegård, 2006; Palen & Salzman, 2002; Rieman, 1993).

Participants are asked by the investigator to keep track of their experiences, perceptions or events based on a set of instructions or preliminary structure. Typically the information that is requested in journals is hard to be recalled accurately, if they are not recorded during the activity or immediately after it. Therefore they are used to prevent inaccuracy of recalling and to enable achieving to a more accurate understanding of the events and participants' experiences.

For this project participants were asked to take note of their experiences during and after each session of analysis and the diaries were to be used to assist in gaining an understanding of their experiences with their analysis process and various features of the user interface.

Considering that imposing structure to participants' diaries may interfere with the participant's perceptions of their experiences, the requested structure of diary entries only includes unstructured comments and open-ended questions about the analysis session. The participants were asked to include any comments about their analysis process and their experience with AnalyticStream during the analysis or shortly after each analysis session. The open-ended questions are designed to provide illumination of the context of the participant's experiences and to allow comparison between the analysis sessions for each participant and between the participants. The structure of the diary entries and the questions are presented in Appendix 2.

6.4.3 Questionnaire

Questionnaires can be used to collect both qualitative data using open-ended questions and quantitative data using close-ended questions. Close-ended questions enable investigators to measure respondents' opinions and open-ended questions allow for richer feedback that may provide insight into explanations for what happened during the course of study and participants' opinions, attitudes, feelings, perceptions, etc. (Cohen, Manion, & Morrison, 2007). Moreover, unexpected issues may emerge in responses to open ended questions, which can result in interesting findings.

We used a post-test questionnaire to ask participants to reflect on the whole experience in the open-ended questions and to answer several close-ended questions about various features of AnalyticStream. A major benefit of using a post-experiment questionnaire is that it can summarize participants' experiences in quantitative and qualitative terms.

Additionally, we used a pre-experiment questionnaire to collect basic demographic information (age, nationality, course of study, and gender) and their experiences with information sharing systems and analytical reasoning tasks.

6.4.4 Interview

In this method, interviewer asks the participants a series of questions and the interviewer should be nonjudgmental to the responses to reduce his potential biasing effect on the participants (Abbas Tashakkori & Charles Teddlie, 2003). Interviewing offers the flexibility to adapt questioning according to the responses of interviewees, to clarify questions or answers, or to probe answers more deeply with supplementary questions as appropriate, to explore issues that emerge from the respondents (Cohen et al., 2007).

Usage logs and participants' responses to open-ended questions in questionnaire and their diaries are potentially vague and may be affected by participants' verbal skills; therefore, we decided to follow them up by a complementary unstructured interview to both validate our understanding and to collect possible additional information. The interviewer took notes of the participants' answers.

6.5 Qualitative Analysis

Qualitative data analysis is "the process of making sense out of data" (Merriam, 1997) through "organizing and sorting data in light of increasingly sophisticated judgements and interpretations" (Glesne & Peshkin, 1991). Valuable qualitative research relies on conducting rigorous analysis of data and in this section we explain how we have approached the data analysis in this study.

We collected information about how our design decisions affect users' behaviour, which of the factors affect the users sharing/reuse decisions, what are the processes involved in collaborative sensemaking using AnalyticStream and how they affect analysts' cognitive processes. Qualitative data analysis is interpreting participants' meanings and we, as researchers, cannot be separated from this interpretation as our experiences, characteristics, beliefs and biases influence the qualitative data collection and analysis process.

The main strategy of enquiry of this mixed-methods study is instrumental case study. Generally, "how" questions can be addressed in depth, using case studies (Yin, 2008) and AnalyticStream is designed as an instrument to help us understand the effects of various design decisions that we made and the guidelines that we suggested, and nuances of distributed asynchronous collaborative sensemaking processes. In this study, the group of analysts involved in the distributed analytics process is the case under investigation. In another level, each analyst is considered as a case, with specific characteristics that requires exclusive investigation.

The sources of data that contain narrative descriptions need to be rigorously analyzed and perhaps coded if the relation between what participants say and what researchers infer is not clear. Those sources are participants' diaries and post-test interviews. In addition, usage logs can be considered as automatically coded record of participants' behaviour. The amount of work required to process a diary depends largely on how structured it is. In this study, participants' diary format was consisted of a set of open-ended questions, which made it much easier to understand, comparing with unstructured diaries.

6.5.1 Trustworthiness and reliability

The triangulation of the usage-logs and daily questionnaire helped in assuring the validity of our interpretations of users' behaviour. Moreover, the ambiguities or uncertain interpretations of participants' behaviour and answers to open-ended questions are disambiguated in post-test interviews through member checking; that is we asked participants to validate our understanding of what they say or how they behaved, whenever the interpretations were not clear.

6.6 Quantitative Analysis

We employed a concurrent embedded mixed method approach and quantitative method, as the secondary method, was less emphasized. In this study, we used quantitative methods to address specific questions different from the qualitative ones as well as for triangulation and verifying the qualitative findings. The mixing and comparison of results (whenever practical) are presented in section 6.7. Additionally, some of the quantitative results are used to enrich the descriptions of qualitative data.

In this study, individual characteristics such as analysis skills, experience in knowledge sharing, desire for sharing, etc not only affect the participant's behaviour directly, but also interact with other characteristics which make it harder to understand and remove their effect. Although a repeated measures design may control variation in participants' individual characteristics, the advantage of repeated measures design can be impaired by various environmental sources of noise. This is particularly prominent in a multiple-session long-term study like this, because we did not want to impose a specific time of the day for their analysis, and during the six-week period of the analysis process, several factors such as various aspects of analysts' cognitive states as well as their stage in the analysis process may change, which make it hard to identify significant difference, while maintaining a desired level of ecological validity. Therefore we decided to focus on quantitative measurements as a complementary source of data for enriching our qualitative analysis.

6.6.1 Measures and Analysis

The usage logs provide several quantitative measures about participants' behaviour. More details on the user interactions that are recorded are provided in section 6.4.1. Descriptive analysis of participants' usage logs provides a summary of users' behaviour and facilitates comparing the participants' behaviours. Moreover, the usage logs are visualized in Tableau to facilitate exploring usage logs for detecting patterns of usage and interaction sequences.

The second source of quantitative data is the post-test questionnaire that aims at evaluating the usefulness of narrative view in comparison with the graph view of stories, and the usefulness of related artefacts panel in comparison with searching for pieces of analysis. Several Likert-scale questions are dedicated to each of the comparisons.

Chi-square analysis is employed to identify the differences in usefulness of recommending related pieces versus search, and comparing the function of narrative/sequential view versus graph view of expanded pieces of analysis.

6.6.2 Threats to Validity

6.6.2.1 Internal Validity

Considering that the study was mainly designed for collecting qualitative data, our research design does not control many of the extraneous variables such as level of expertise, time of the day that they analysts work with the system, artefacts that are available in the system, etc. Additionally, the research design imposes some extraneous variables such as the data available in the system, which is different for different analysts, as some of them started their analysis later than the other analysts. However, this is quite common in distributed analysis where analysts are not aware of each other and they may analyze their shared information in different time frames. Moreover, regardless of the time that analysts start investigating on an information space, the artefacts that are available to them for their analysis are different. However, considering that we did not have any between-subject comparisons, this issue is of less importance.

While the usage logs provide objective measurements about usage of user interface components, the Likert-scale questions that address our hypotheses, are subjective measurements. For each hypothesis, we asked several questions from the users to make sure that the questions are setting the stage for a fair judgment from the user.

6.6.2.2 External Validity

Although the goal of doing a case study is to expand and generalize theories, they usually do not represent a sample. The generalisability of case studies can be increased by appropriate case selection. In this study we can consider the whole collaborative analysis process as one case, or in another level, each of the analysts participating in the collaborative analysis process can be analyzed as a case. Our participants joined the collaborative analysis process at different stages to enable us evaluate the generalisability of our observations regarding the stage of collaborative analysis. Also some of our participants were experienced with using visual analytics tools and performing analytical tasks, while some of them were new to analytical tasks.

We also tried to improve the ecological validity of our case study comparing to controlled lab studies by allowing our cases to work with the system in their comfort zones and at the times they feel ready for focusing on their analysis rather than in a controlled lab setting that users may feel more like they are attending an experiment rather than performing analysis.

6.7 Results and Findings

In this section, we first report the results of the close-ended questions, and we continue with presenting our analysis of usage behaviour and answers to open-ended questions.

6.7.1 Comparison of narrative/sequential view and graph view

The distributions of participants' answers to close-ended questions are presented in the following bar charts:

How do you rate the usefulness of Narrative view for gaining an overall understanding of expanded pieces of analysis?

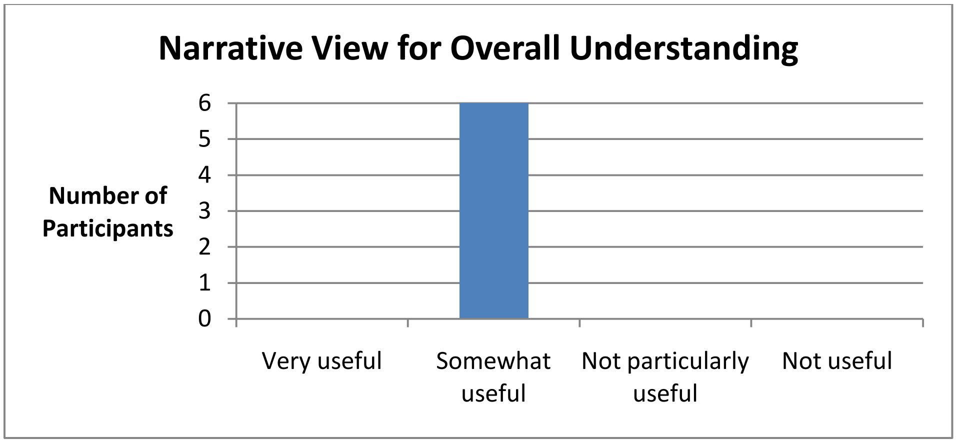

Figure 20 Usefulness of Narrative view for gaining an overall understanding of expanded pieces of analysis

How do you rate the usefulness of Narrative view for understanding details of expanded pieces of analysis?

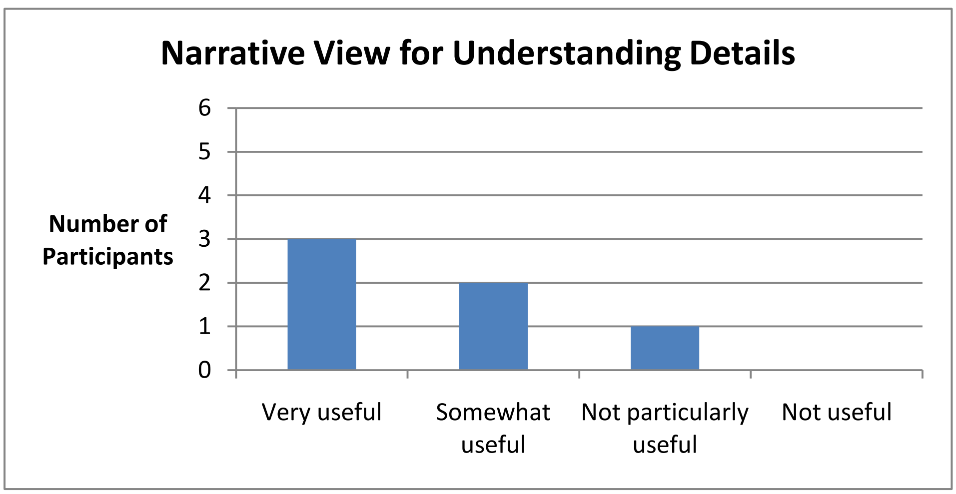

Figure 21 Usefulness of Narrative view for understanding details of expanded pieces of analysis

How do you rate the usefulness of Graph view for gaining an overall understanding of expanded pieces of analysis?

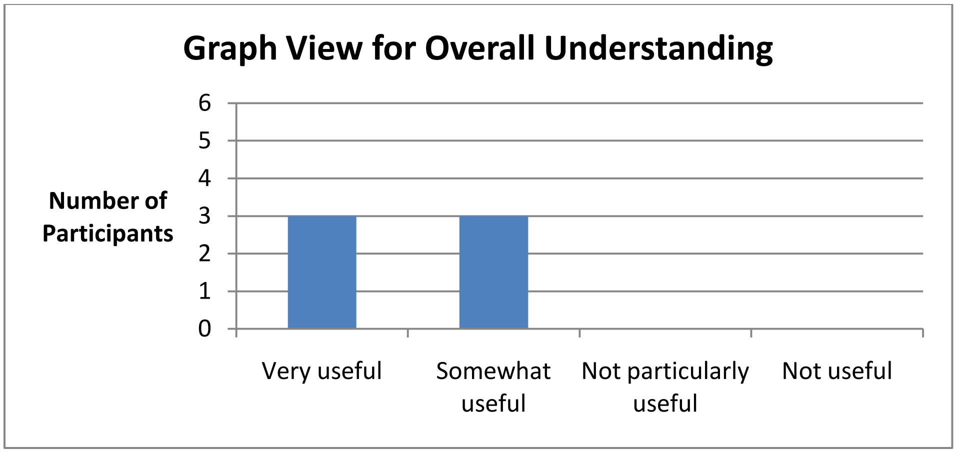

Figure 22 Usefulness of Graph view for gaining an overall understanding of expanded pieces of analysis

How do you rate the usefulness of Graph view for understanding details of expanded pieces of analysis?

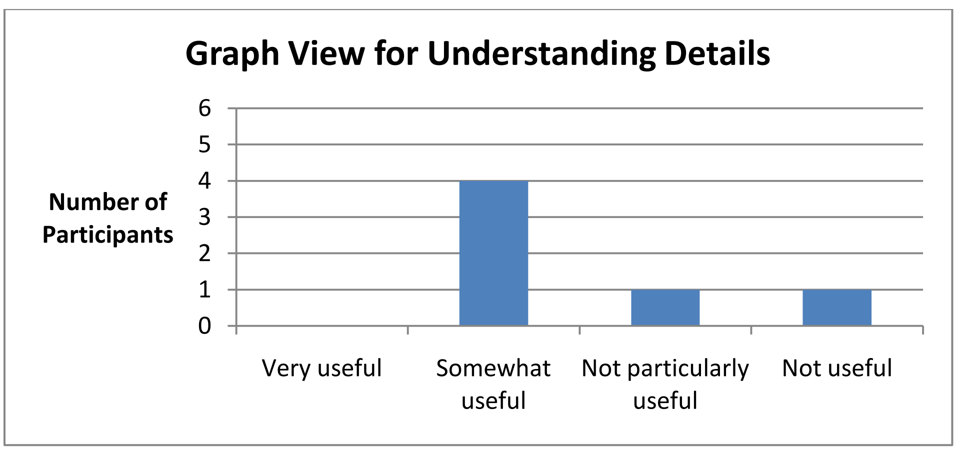

Figure 23 Usefulness of Narrative view for understanding details of expanded pieces of analysis

Which of the views of an expanded piece was more useful for gaining an overall understanding?

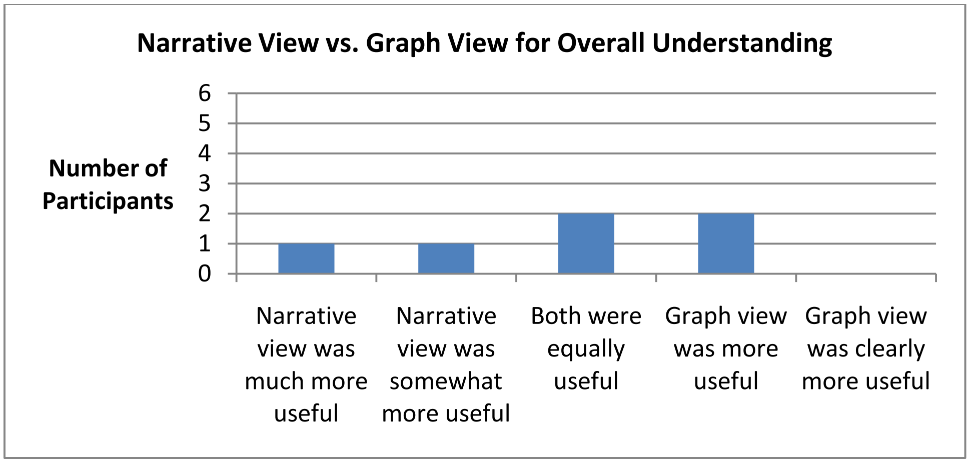

Figure 24 Comparing the Usefulness of Narrative view vs. Graph view for gaining an overall understanding of expanded pieces of analysis

Which of the views of an expanded piece was more useful for understanding its details?

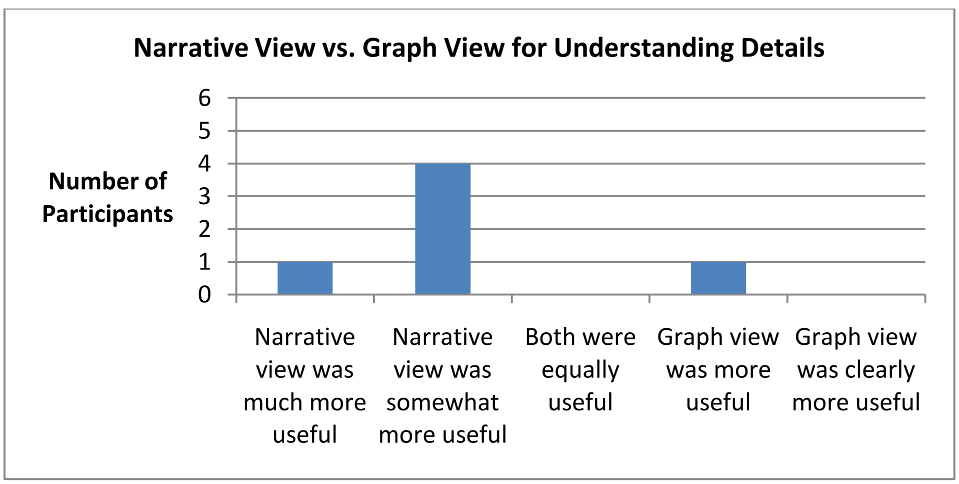

Figure 25 Comparing usefulness of Narrative view vs. Graph view for understanding details of expanded pieces of analysis

For identifying the possible significant differences between the Narrative view and the Graph view, the 4-point Likert scale questions are binned into two categories (useful or not useful) and 5-point Likert scale questions are binned into 3 categories (Narrative view, Graph view or no difference) and chi-square is applied. Considering that our questions have addressed different aspects of the design and different possible benefits, significance level of 0.05 is used and experiment-wise error rate is not calculated.

The results are following:

Table 1 Results of the 4-point Likert scale questions evaluating the functionality of Graph view and Narrative/Sequential view

	Not useful	Useful	X^2	p-value
Narrative view for details	1	5	1.5	0.22
Narrative view for overall understanding	**0**	**6**	**4.16**	**0.041<0.05**
Graph view for details	2	4	0.16	0.69
Graph view for overall understanding	**0**	**6**	**4.16**	**0.041<0.05**

Based on table 1, both Graph view and Narrative view were significantly useful for gaining an overall understanding of pieces of analysis.

Table 2 Results of the 5-point Likert scale questions comparing the functionality of Graph view and Narrative/Sequential view

	Narrative view	No Difference	Graph view	X^2	p-value
Better for details	**5**	**0**	**1**	**7.07**	**0.029<0.05**
Better for general understanding	2	0	4	4.04	0.13

Based on table 2, narrative view was significantly more useful for understanding details of pieces of analysis and graph view seemed to be more useful for gaining an overall understanding, though it was not statistically significant.

6.7.2 Comparison of related artefacts panel and search

The distributions of participants' answers to close-ended questions are presented in the following bar charts:

How do you rate the usefulness of Related Pieces panel for finding your own pieces of analysis?

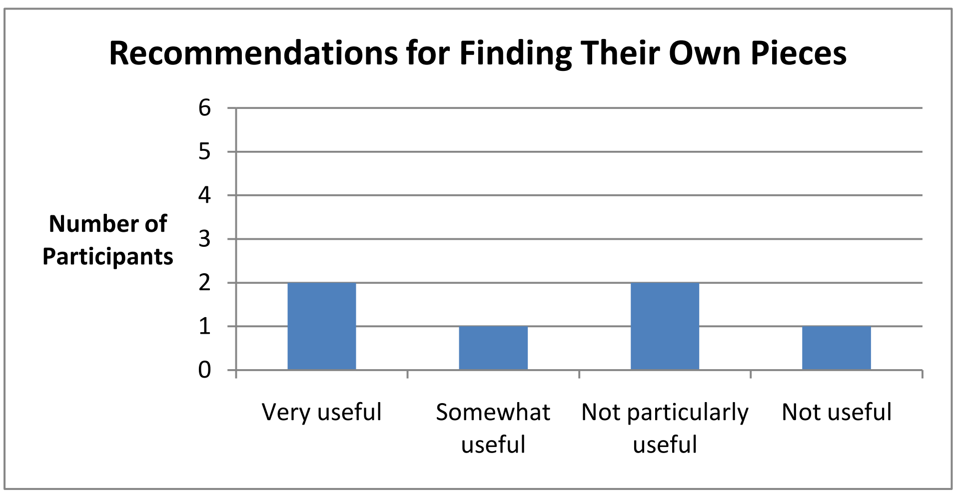

Figure 26 Usefulness of recommendations for finding participants' own pieces of analysis

How do you rate the usefulness of Related Pieces panel for finding other analysts' pieces of analysis?

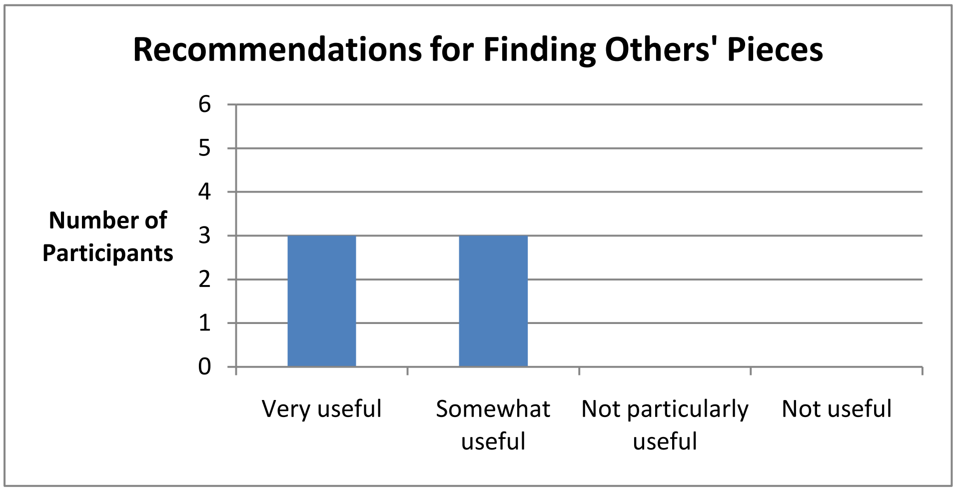

Figure 27 Usefulness of recommendations for finding other participants' pieces of analysis

How do you rate the usefulness of search feature for finding your own pieces of analysis?

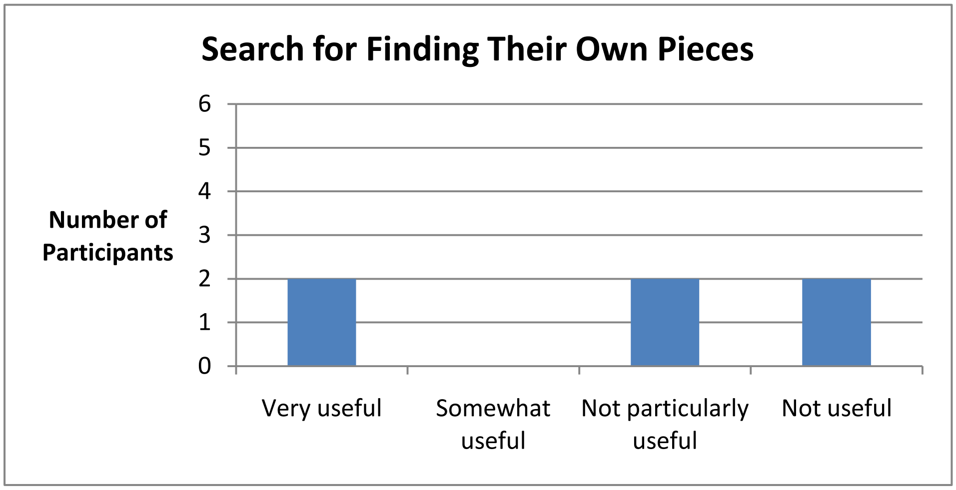

Figure 28 Usefulness of search for finding participants' own pieces of analysis

How do you rate the usefulness of search feature for finding other analysts' pieces of analysis?

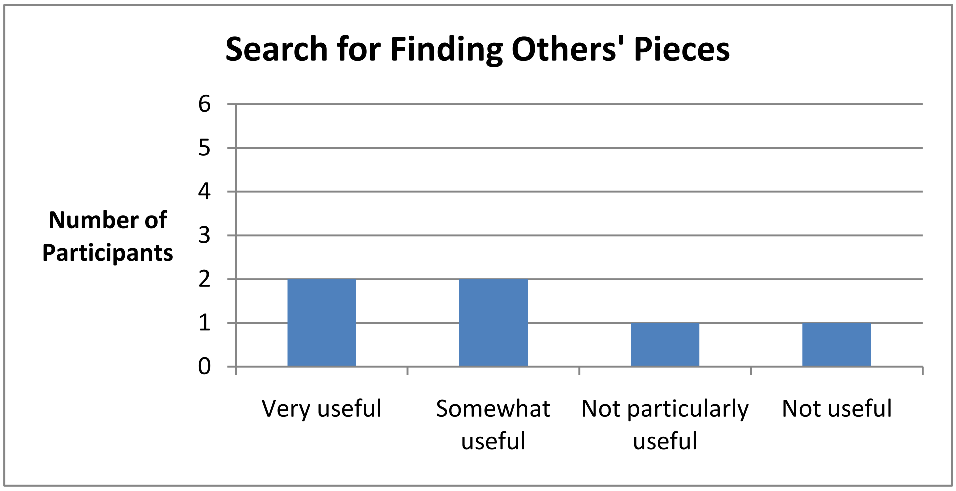

Figure 29 Usefulness of search for finding other participants' pieces of analysis

Which of the features of search and Related Pieces panel are more useful for finding your pieces?

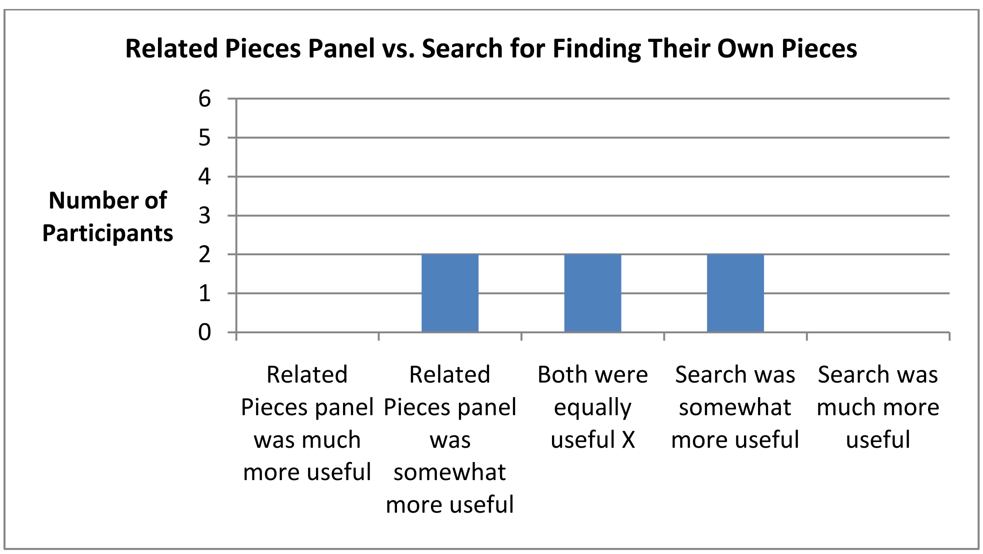

Figure 30 - Comparing usefulness of recommendations vs. search for finding participants' own pieces of analysis

Which of the features of search and Related Pieces panel are more useful for finding other analysts' pieces?

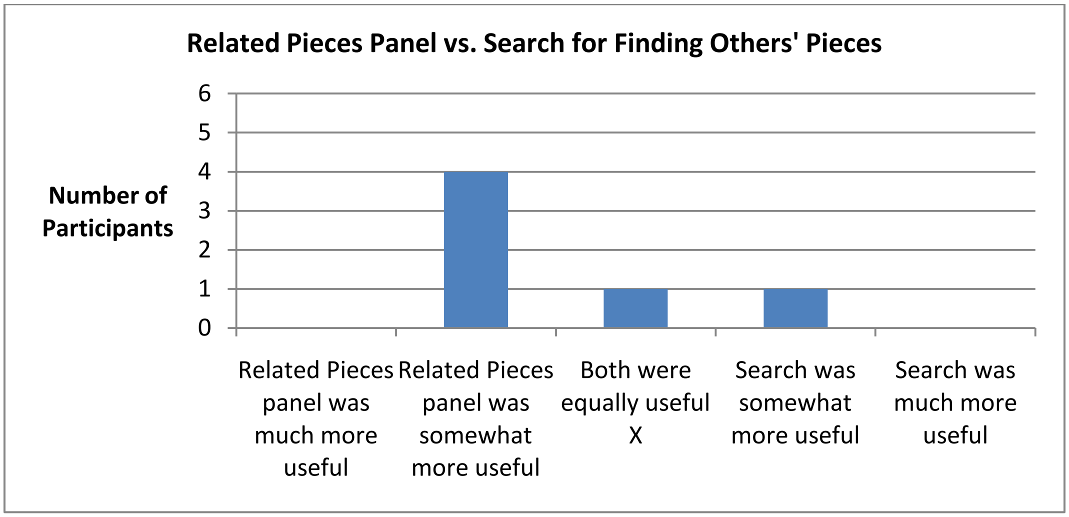

Figure 31 - Comparing usefulness of recommendations vs. search for finding other participants' pieces of analysis

Similar to the previous section, for identifying the possible significant differences between the two features, the 4-point Likert scale questions are binned into two categories (useful or not useful) and 5-point Likert scale questions are binned into 3 categories (recommendation, search or no difference) and chi-square is applied. Similar to the previous set of questions, considering that our questions have addressed different aspects of the design and different possible benefits, significance level of 0.05 is used and experiment-wise error rate is not calculated.

The results are following:

Table 3 Results of the 4-point Likert scale questions evaluating the functionality of Related Pieces panel and Searching

	Not useful	Useful	X^2	p-value
“Related Pieces” for finding their own pieces	3	3	0.00	1.00
“Search” for finding their own pieces	4	2	0.16	0.69
“Related Pieces” for finding others’ pieces	**0**	**6**	**4.16**	**0.041<0.05**
“Search” for finding others’ pieces	2	4	0.16	0.69

Based on table 3, recommending related pieces of analysis was significantly useful for finding other participants’ pieces of analysis. The qualitative data in section 6.7.4 clarifies how the recommendations could help the participants.

Table 4 Results of the 5-point Likert scale questions comparing the functionality of Related Pieces panel and Searching

	Search	No Difference	Related Pieces	X^2	p-value
Better for finding their own pieces	2	2	2	0.00	1.00
Better for finding others’ pieces	1	1	4	3.03	0.22

Both qualitative and quantitative data suggest that suggesting relevant pieces of analysis is more helpful than search for finding other analysts’ pieces of analysis. However, a more powerful quantitative study with a larger sample size is required to validate the comparative usefulness of recommending relevant pieces vs. search.

6.7.3 Findings from usage logs and answers to open ended questions

Our findings are categorized based on the qualitative research questions, presented in section 6.1.1:

6.7.3.1 How does recommendation of relevant pieces of analysis affect the analysis process (mainly effects on searching, sharing reusing etc behaviours)?

Based on the quantitative results, all of the participants found the recommendations useful for finding other analysts pieces of analysis (statistically significant) and most of the them (4 out of 6) found the recommendations more useful than searching for finding other analysts' pieces of analysis. Considering that the only participant that found it less helpful than search was the least active participant of the study (based on the number of interactions with system), we hypothesize that the more analysts engage in their analysis, the more recommendations become useful, as analysts are focused on their analysis and often they do not spend time exploring others analysis products.

One of the participants found recommending relevant artefacts useful for connecting subplots when analysts were working on different subplots and the subplots had something in common. This was one of the main anticipated benefits of recommendations. Another important anticipated benefit was enhancing the discoverability of artefacts for facilitating the reuse process. All of the participants browsed the recommendations and read many of the recommended pieces, several times and all of them rated it as somewhat useful or very useful for finding others' pieces of analysis. This was a rather weak indicator of reuse but reviewing the usage logs showed that the more active half of the participants reused the recommended pieces of analysis in building their stories or expanded pieces.

Another participant also mentioned that recommendations were more useful because she "was not looking for any specific thing". As the participants were mostly in the initial phases of analyzing the dataset, where they had an ambiguous model of the problem space and they could not distinguish key entities, recommendations could help them in identifying important entities by recommending pieces of analysis that contain similar entities. This important benefit of recommending related pieces of analysis was not anticipated.

6.7.3.2 How does various design decisions (mainly privacy options, certainty declaration, and format of pieces of analysis) shape the sharing and reuse behaviours?

Certainty levels enabled the participants to share their doubts and suspicions. One of the participants explicitly mentioned that he was not sharing this information on the wiki that they have been using for sharing findings with analysts that have not been participating in the study, because he did not think it is ok to include non-important or uncertain intelligence. This feature helped the participants to share more intermediate products, which made it easier to uncover their analysis process.

One issue was that participants were not using same criteria for assigning certainty levels. That is some of the analysts tend not to be 100% certain about anything and some of them tend to trust all of the sources of information. Moreover, some of the participants were not used to entering their certainty, which was misleading. Analysis support systems should not provide a default certainty value and the pieces that their producer has not explicitly expressed his certainty about them should clarify that instead of showing the default value.

Having the unpublicized privacy level (public but not published to the activity steam) enabled analysts to create helper pieces or pieces that they wanted to use for a special purpose such as creating a relation. These pieces did not have any independent meaning or value, but they were produced to clarify another piece or to relate two other pieces.

Also two of the analysts preferred to set some of the pieces with non-important information as unpublished. Among them were the pieces that were extracted pieces from the datasets. Again, the default value played an important role and some of the pieces that were not worth publishing were published probably because it was the default privacy level. Some of the participants created private pieces but they removed most of them during their analysis process.

6.7.3.3 How is each of the user interface components used and how does each of them contribute to the sensemaking processes?

All of the analysts browsed the Activity stream right after logging in most of their sessions. However, in the interviews they mentioned that sometimes it was aimed at remembering their own activities rather than others'.

Comments panel was gone unnoticed and only one of the analysts used it. One of the analysts mentioned that he preferred to put everything in new pieces or add extra information to the pieces. Another possible reason is that the analysts did not perceive each other as collaborators as the system supports very limited communication.

Graph view was useful for connecting pieces of information and gaining an overall understanding of expanded pieces of analysis. Moreover, participants created pieces as representatives of entities and added new relations to show how they are related. More importantly, graph view was helpful in the early stages of analysis when participants could only identify few relations and they could not form a story using their findings.

The narrative view was used to aggregate pieces of analysis in the form of groups, sequences, cases and stories. The narrative view was used as an ordered set or just a group at early stages when analysts could not yet make a coherent story of the events, and sometimes it was used as a story including a sequence of related events. Although sometimes the ordering of pieces was confusing for some of the analysts, analysts that entered the experiment later than the others were happy to have prepared stories and one of them explicitly mentioned in her diary that "At this stage [early stage of the analysis], I'm more interested in seeing stories rather than individual artefacts". An interesting issue was the use of deictic expressions that we had not anticipated. Pieces have different locations in the narrative view of different expanded pieces and sometimes analysts used deictic expressions such as "the fact above" to point to another piece of analysis; however, in case that piece was being reused in another story, the deictic expression was not valid anymore and it was confusing for the analysts reading the second story. Moreover, the narrative view was scalable in a way that the number of pieces in a row was determined based on the width of AnalyticStream's window; therefore, resizing the window may lead to rearrangement of pieces and only their sequence is preserved.

Both graph view and narrative view were helpful for demonstrating how various analysts' findings and subplots were related to each other and many of the expanded pieces contained pieces from various analysts. Some of the hypotheses and complex connections were easier to identify in the graph view as one of the analysts mentioned: "The graph view was better than the narrative view for seeing hypotheses and connections across disparate data since they usually involve many different pieces with different relationships."

6.7.3.4 How does the usage of various users interface components (mainly desktop, narrative view, recommendations panel) and data structures (mainly comment, piece of information, and story) evolve during the course of analysis?

As we mentioned in section 5.1, a piece of analysis could be expanded to a sequence of pieces of analysis, represented in the Narrative view. The expansion could have an arbitrary meaning and arbitrary relation to the piece; however it was originally aimed at describing the story behind a piece of analysis or a story that can explain how or why a piece of analysis is produced. The participants started using expansions the way we described, but during the course of the study they used them as a grouping mechanism and sometimes as a sorted set or a sequence of pieces of analysis. Considering that analysts were in the initial phases of analysis and building hypotheses and cases was not easy in that phase, it was interesting to see how they appropriated the data structure to address their need for categorization, sequencing, etc. and therefore many of the expanded pieces are not stories. However the participants did arrange the pieces in the narrative view even if they were not stories, which suggests that the ordering was meaningful.

Another interesting usage of pieces was for representing relations that should connect more than two pieces of analysis. One of the analysts figured out that she can create a “master relationship node” and connect all of the related pieces to that.

6.7.3.5 What are the shortcomings of the design?

The design of AnalyticStream was particularly focused on supporting sharing and reuse of pieces of analysis to facilitate the integration of analysts’ efforts, therefore as a collaborative visual analytics system it had several shortcoming which analysing them can inform the design of future distributed collaborative visual analytics systems.

The main weakness of AnalyticStream was the lack of support for various levels of collaboration. The design of AnalyticStream assumes that analysts are not explicitly collaborating, rather they are working on their own tasks and sometimes they find each other’s analysis products useful for their own tasks and implicitly or explicitly reuse them. Therefore, AnalyticStream does not support direct communication between analysts. The only way of communication was through pieces of analysis and comments on them, which clearly were not enough and some of the participants mentioned their need for asking questions, consultations, and direct communications. Although the comments feature could solve the problem to some extent, it was gone unnoticed, because the duration of the experiment was limited and participants were not thinking of each other as collaborators and also they rarely checked their previous pieces as they were trying to make sense of new information as much as possible. Collaborative visual analytics systems should support various levels of collaboration ranging from implicit collaboration through shared pieces of information or explicitly with direct communication even perhaps in same place and at the same time.

Participants also asked for more personal views of the information, for example they wanted to see everything that they have created in a single view to be able to summarize and review them.

Moreover, participants had several suggestions for improving the design of AnalyticStream. Among those were: enabling customization of pieces of analysis so that they can better serve for various purposes such as representing entities, providing entity-centric workspace so that they can set specific visual properties for entities and instances to make them easier to memorize and recall, and quick preview of pieces relations without requiring expanding them.

Additionally two of the participants preferred shorter list of suggestions and also providing the reason behind suggestions so that the analysts can faster and better understand the relation between the suggested pieces and his task or workspace.

6.8 Summary of findings and Discussion

Generally, the results of our study showed that suggesting the relevant artefacts made those artefacts easily discoverable for the participants and helped them to identify the more important entities in their information space based on their occurrence in multiple pieces. Additionally, relevant artefacts panel provided context-specific awareness of other analysts' activities, which was a welcome addition to Activity Stream for providing awareness. The recommendations were significantly useful for finding other analysts pieces of analysis.

The narrative view provided the participants with a place for categorizing and ordering heir pieces of analysis as well as telling stories based on several related events. Due to the efficient and information-dense comic-like representation of pieces of analysis, the participants found the narrative view more helpful than the graph view for understanding details of pieces of analysis (statistically significant) and graph view seemed to be better for gaining an overall understanding of expanded pieces (not statistically significant). Moreover, they found narrative view as well as graph view helpful for gaining an overall understanding of expanded pieces of analysis (both statistically significant). The default visual size of pieces of analysis made participants to create smaller pieces of analysis and use a sequence of them for telling a story or building a case. This was especially useful for improving the reusability of pieces as analysts could drag and drop each other's pieces in their own stories or groupings, without the need for decomposing them.

Our case study confirmed many of the theoretically grounded suggestions in two ways:

- Several features that facilitated the distribution of memory, reasoning and attention facilitated the collaborative sensemaking processes.
- Due to the limited resources and limited scope of our implementation, we could not apply many of the guidelines that our analysis suggested. As a result, we could clearly understand some of the problems that occurred during our study such as the problem of easy integration of efforts with the community's efforts, which are mentioned below.

In chapter 3, we discussed that: "A distributed collaborative system should support a spectrum of collaborative processes, starting from the point in which users do not know each other, to the point that users explicitly and consciously interact with each other to achieve shared goals". Unfortunately, AnalyticStream does not support the entire spectrum, though it did provide some support for the ends of that spectrum:

- It supported the first through awareness mechanisms such as a shared stream of reasoning artefacts, basic history mechanisms, comments and social navigation by clicking on analysts' profile pictures to see their activities.

- It supported the last by providing participants' email addresses,

However, because intermediate stages were not supported, the transition between the stages was not possible for the participants.

One of the interesting issues that showed up during the evaluation was the availability of analysts for synchronous collaboration. AnalyticStream is able to handle concurrency, but does not provide any special support for the analysts that can collaborate synchronously, or even in same place. This shortcoming of AnalyticStream led analysts to use traditional tools like whiteboard and markers to collaborate synchronously in same place.

Another challenge was posed by the lack of support for integration with community. As we mentioned earlier, the analysis task was part of the 2010 VAST challenge and several researchers including two of the participants of this study have been working on it, even before the start of the study. Some of these analysts were not eager to use AnalyticStream and this problem limited the amount of collaboration in the system and made the two participants to use a wiki in addition to AnalyticStream to make sure all of their collaborators can access their findings. AnalyticStream could have provided a read-only (or with special permissions for non- users) version of information and analysis outcomes or an easy way for exporting them to the external wiki, so that people who do not want to use the system can access them.

It is important for a comprehensive collaborative visual analytics system to support various modes of collaboration as well as mechanisms for importing and exporting information.

7: CONCLUSIONS AND FUTURE WORK

Asynchronous collaborative analytics systems facilitate loosely coupled collaboration among distributed analysts to provide a platform for more efficient large-scale analytical collaborations, and enable analysts to address complicated multi-faceted problems that they currently cannot deal with.

We started our analysis of the design space of distributed visual analytics systems by reviewing the theoretical foundations of understanding it including sensemaking models, foundations of reusable reasoning and asynchronous collaboration. Then, we explored the design space of distributed visual analytics systems using the theoretical lenses of Activity theory and Distributed Cognition Theory to provide a more continuous and comprehensive view of the design space comparing to the previous efforts such as (Heer & Agrawala, 2007). In addition, we focused on analyzing distributed analytical processes as well as the process of sharing and reuse of reasoning artefacts. Activity Theory helped us identify the important factors that influence the design situation and Distributed Cognition Theory helps us understand how user interface elements can support and augment the distribution of cognitive processes such as attention, memory and reasoning. While both theories have limitations in providing predictions and prescriptions, they can inform and inspire the design process.

We introduced, AnalyticStream, a proof of concept web-based distributed visual analytics system that embodies some of the theoretically based suggestions, while lacking some others due to the limited scope of this research. The design of AnalyticStream is aimed at fostering the sharing and reuse of reusable pieces of analysis. It facilitates the reuse process through employing the following mechanisms:

- Improving the discoverability of pieces of analysis by suggesting the pieces that are related to an analyst's workspace. This method can be seen as a reminding process distributed over several analysts at different times.
- Implicitly biasing users to create smaller pieces of analysis that do not require decomposition, by providing small-size visual representations for the pieces and allowing users to compose them to build cases and stories
- Providing a narrative/sequential view and graph view to facilitate making sense of a set of related pieces of analysis
- Providing symbolic incentives for sharing pieces of analysis including enabling users to indicate if a piece was helpful for them and showing the number of times that a piece is used in stories
- Fostering sharing of pieces of analysis by allowing users to indicate their level of certainty about the pieces that they share, so that they do not feel excessive responsibility about the pieces that they share

Ultimately, we conducted a mixed-methods case study, collecting various quantitative and qualitative data, to better understand the collaborative sensemaking processes that analysts go through in a distributed asynchronous collaborative setting, to evaluate the effects of our design decisions and to get a sense of the validity of suggestions and guidelines resulted from our theoretical analysis of the design space. Employing a mix of methods enabled us to assess different facets of the complex outcomes of using AnalyticStream and the design decisions embedded in it to yield a broader, richer portrait than one method alone can. The results of our analysis clearly show that both qualitative and quantitative data provide interesting insights for the design of future distributed visual analytics systems.

One of the important venues for future research is extending AnalyticStream to support sharing and reuse of higher-level artefacts and processes such as analysis patterns or strategies. This is an exciting yet complex research problem, which requires an in-depth understanding of human cognitive processes. Moreover, the results of the mixed-methods study presented in this thesis can be used as a precursor to studies with more emphasis on quantitative measurements to verify the generated hypotheses and validate our findings by focusing on accurately measuring the related constructs in controlled lab experiments.

Ultimately, we believe that the findings of this research can inform the design of future distributed visual analytics systems, and inspire further investigations in this exciting area of research.

REFERENCE LIST

Alaszewski, A. (2006). *Using diaries for social research*. SAGE.

Antoniadis, P., & Le Grand, B. (2007). Incentives for resource sharing in self-organized communities: From economics to social psychology. *Digital Information Management*, 756–761.

Aragon, C. R., Poon, S. S., Aldering, G. S., Thomas, R. C., & Quimby, R. (2009). Using visual analytics to develop situation awareness in astrophysics. *Information Visualization*, *8*(1), 30-41. doi:10.1057/ivs.2008.30

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1978). *Educational psychology: A cognitive view*. Holt Rinehart and Winston.

Axelrod, R. (2006). *The evolution of cooperation*. Basic Books.

Bannon, L., & Bødker, S. (1997). Constructing common information spaces. In *Proceedings of the fifth conference on European Conference on Computer-Supported Cooperative Work* (pp. 81-96). Lancaster, UK: Kluwer Academic Publishers. Retrieved from http://portal.acm.org/citation.cfm?id=1241986

Basili, V. R. (1990). Viewing maintenance as reuse-oriented software development. *IEEE software*, *7*(1), 19–25.

Bier, E., Ishak, E., & Chi, E. (2006). Entity workspace: an evidence file that aids memory, inference, and reading. *Intelligence and Security Informatics*, 466–472.

Birsan, D. (2005). On plug-ins and extensible architectures. *Queue*, *3*(2), 46.

Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson, G. (2007).

From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories. *Journal of Computer-Mediated Communication*, *12*(2), 652-672. doi:10.1111/j.1083-6101.2007.00343.x

Boyd, D. (2008). Facebook's Privacy Trainwreck: Exposure, Invasion, and Social Convergence. *Convergence*, *14*(1), 13-20. doi:10.1177/1354856507084416

Brehmer, B. (1991). Distributed decision making: some notes on the literature. In *Distributed decision making: Cognitive models in cooperative work.* NY: John Wiley.

Bruner, J. (1991). The Narrative Construction of Reality. *Critical Inquiry*, *18*(1), 1-21.

Card, S. K., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think* (1st ed.). Morgan Kaufmann.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1-39. doi:10.1016/0010-0277(86)90010-7

Cockburn, A. (2001). *Agile Software Development*. Addison-Wesley Professional.

Cohen, L., Manion, L., & Morrison, K. R. B. (2007). *Research methods in education*. Routledge.

Cosley, D., Frankowski, D., Kiesler, S., Terveen, L., & Riedl, J. (2005). How oversight improves member-maintained communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 11-20). Portland, Oregon, USA: ACM. doi:10.1145/1054972.1054975

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2007). SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces* (pp. 32-41). Honolulu, Hawaii, USA: ACM. doi:10.1145/1216295.1216309

Creswell, J. W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches*. SAGE.

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004). A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 175-182). Vienna, Austria: ACM. doi:10.1145/985692.985715

Dadzie, A. S., Lanfranchi, V., & Petrelli, D. (2009). Seeing is believing: Linking data with knowledge. *Information Visualization*, *8*(3), 197–211.

Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared workspaces (pp. 107-114). Toronto, Ontario, Canada: ACM. Retrieved from http://portal.acm.org.proxy.lib.sfu.ca/citation.cfm?doid=143457.143468

Ellis, J. B., Halverson, C. A., & Erickson, T. (2005). *Sustaining Community: The role and design of incentive mechanisms in online systems (Workshop at Group 2005)* (Workshop Report). Sanibel Island, Florida, USA.

Engestrom, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Orienta-Konsultit Oy.

Engeström, Y., Miettinen, R., & Punamäki, R. (1999). *Perspectives on Activity Theory*. Cambridge University Press.

Erickson, T., Halverson, C., Kellogg, W. A., Laff, M., & Wolf, T. (2002). Social translucence: designing social infrastructures that make collective activity visible. *Commun. ACM*, *45*(4), 40-44.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363-370). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from http://portal.acm.org.proxy.lib.sfu.ca/citation.cfm?id=1219885

Fischer, G. (1987). Cognitive View of Reuse and Redesign. *Software, IEEE*, *4*(4), 60-72.

Frakes, W., & Terry, C. (1996). Software reuse: metrics and models. *ACM Computing Surveys (CSUR)*, *28*(2), 415–435.

Gerken, J., Bak, P., Jetter, H. C., Klinkhammer, D., & Reiterer, H. (2008). How to use interaction logs effectively for usability evaluation. In *CHI 2008 Workshop BELIV'08 Beyond Time and Errors: Novel Evaluation Methods for Information Visualization* (Vol. 8). Presented at the CHI 2008 Workshop BELIV'08 Beyond Time and Errors: Novel Evaluation Methods for Information Visualization.

Glesne, C., & Peshkin, A. (1991). *Becoming Qualitative Researchers: An Introduction*. Longman Publishing Group.

Greenberg, S., & Buxton, B. (2008). Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 111-120). Presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Florence, Italy: ACM. doi:10.1145/1357054.1357074

Grinstein, G., Plaisant, C., Scholtz, J., & Whiting, M. (2010). IEEE VAST 2010 Challenge. Retrieved August 2, 2010, from http://hcil.cs.umd.edu/localphp/hcil/vast10/index.php

Halasz, F. G., Moran, T. P., & Trigg, R. H. (1987). Notecards in a nutshell. *SIGCHI*

Bull., *17*(SI), 45-52.

Halverson, C. A. (2002). Activity Theory and Distributed Cognition: Or What Does CSCW Need to DO with Theories? *Comput. Supported Coop. Work*, *11*(1-2), 243-267.

Haven, K. (2007). *Story Proof: The Science Behind the Startling Power of Story*. Libraries Unlimited.

Heer, J., & Agrawala, M. (2007). Design Considerations for Collaborative Visual Analytics. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*. Presented at the IEEE Visual Analytics Science & Technology (VAST), IEEE. doi:10.1.1.93.8234

Heer, J., van Ham, F., Carpendale, S., Weaver, C., & Isenberg, P. (2008). Creation and Collaboration: Engaging New Audiences for Information Visualization. In *Information Visualization* (pp. 92-133). Retrieved from http://dx.doi.org/10.1007/978-3-540-70956-5_5

Heer, J., Mackinlay, J., Stolte, C., & Agrawala, M. (2008). Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1189-1196. doi:http://doi.ieeecomputersociety.org/10.1109/TVCG.2008.137

Heer, J., Viégas, F. B., & Wattenberg, M. (2007). Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1029-1038). San Jose, California, USA: ACM. doi:10.1145/1240624.1240781

Heer, J., Viégas, F. B., & Wattenberg, M. (2009). Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Commun. ACM*, *52*(1), 87-97. doi:10.1145/1435417.1435439

Heuer, R. J. (2005). *Psychology of Intelligence Analysis* (1st ed.). Nova Science.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.*, *7*(2), 174-196. doi:10.1145/353485.353487

Hutchins, E. (1996). *Cognition in the Wild* (New edition.). The MIT Press.

Hyldegård, J. (2006). Using diaries in group based information behavior research: a methodological study. In *Proceedings of the 1st international conference on Information interaction in context* (pp. 153-161). Copenhagen, Denmark: ACM. doi:10.1145/1164820.1164851

Johansen, R. (1988). *GroupWare: Computer Support for Business Teams*. The Free Press. Retrieved from http://portal.acm.org/citation.cfm?id=542298

Kadivar, N., Chen, V., Dunsmuir, D., Lee, E., Qian, C., Dill, J., Shaw, C., et al. (2009). Capturing and supporting the analysis process. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (pp. 131-138). Presented at the 2009 IEEE Symposium on Visual Analytics Science and Technology (VAST), Atlantic City, NJ, USA. doi:10.1109/VAST.2009.5333020

Kaptelinin, V., Nardi, B. A., & Macaulay, C. (1999). Methods & tools: The activity checklist: a tool for representing the “space” of context. *Interactions*, *6*(4), 27–39.

Kaptelinin, V., & Nardi, B. A. (2009). *Acting with Technology: Activity Theory and Interaction Design*. The MIT Press.

Karahalios, K. G., & Donath, J. S. (2004). *Social Catalysts: enhancing communication in* *mediated spaces*. Massachusetts Institute of Technology.

Kirschner, P. A., Buckingham-Shum, S. J., & Carr, C. S. (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making* (1st ed.). Springer.

Klein, G., Moon, B., & Hoffman, R. (2006). Making Sense of Sensemaking 1: Alternative Perspectives. *Intelligent Systems, IEEE*, *21*(4), 70-73. doi:10.1109/MIS.2006.75

Kort, J., & Poot, H. D. (2005). Usage analysis: combining logging and qualitative methods. In *CHI '05 extended abstracts on Human factors in computing systems* (pp. 2121-2122). Portland, OR, USA: ACM. doi:10.1145/1056808.1057117

LaBerge, D. (1997). Attention, Awareness, and the Triangular Circuit. *Consciousness and Cognition*, *6*(2-3), 149-181. doi:10.1006/ccog.1997.0305

Leont'ev, A. N. (1978). *Activity, Consciousness, and Personality*. Prentice-Hall.

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

Merriam, S. B. (1997). *Qualitative Research and Case Study Applications in Education: Revised and Expanded from I Case Study Research in Education/I* (Rev Sub.). Jossey-Bass.

Neisser, U. (1967). *Cognitive Psychology* (1st ed.). Prentice Hall.

Nobarany, S., & Haraty, M. (2009). Improving Accuracy of Tagging Systems Using Tag Qualifiers and Tagraph Vocabulary System. In *Metadata and Semantics* (pp. 223-226). Springer. Retrieved from http://dx.doi.org/10.1007/978-0-387-77745-0_21

Norman, D. A. (1994). *Things That Make Us Smart: Defending Human Attributes In The Age Of The Machine*. Basic Books.

Palen, L., & Salzman, M. (2002). Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 87-95). New Orleans, Louisiana, USA: ACM. doi:10.1145/587078.587092

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual review of psychology*, *49*(1), 345–375.

Parent, W. A. (1983). Recent Work on the Concept of Privacy. *American Philosophical Quarterly*, *20*(4), 341-355.

Paul, S. A., & Morris, M. R. (2009). CoSense: enhancing sensemaking for collaborative web search (pp. 1771-1780). Boston, MA, USA: ACM. doi:10.1145/1518701.1518974

Pike, W., Bruce, J., Baddeley, B., Best, D., Franklin, L., May, R., Rice, D., et al. (2008). The Scalable Reasoning System: Lightweight visualization for distributed analytics. In *Proceedings of the 2008 IEEE Symposium on Visual Analytics Science and Technology* (pp. 131-138). Presented at the Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on.

Pike, W., May, R., & Turner, A. (2007). Supporting Knowledge Transfer through Decomposable Reasoning Artifacts. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences* (p. 204c). IEEE Computer Society. Retrieved from http://portal.acm.org/citation.cfm?id=1255916

Pirolli, P., & Card, S. K. (2005). The Sensemaking Process and Leverage Points for

Analyst Technology as Identified Through Cognitive Task Analysis.

Preece, J., & Maloney-Krichmar, D. (2005). Online communities: Design, theory, and practice. *Journal of Computer-Mediated Communication*, *10*(4).

Richard J. Boland, J., Maheshwari, A. K., Te'eni, D., Schwartz, D. G., & Tenkasi, R. V. (1992). Sharing perspectives in distributed decision making (pp. 306-313). Toronto, Ontario, Canada: ACM. doi:10.1145/143457.143536

Rieman, J. (1993). The diary study: a workplace-oriented research tool to guide laboratory efforts. In *Proceedings of the INTERCHI '93 conference on Human factors in computing systems* (pp. 321-326). Amsterdam, The Netherlands: IOS Press. Retrieved from http://portal.acm.org/citation.cfm?id=164935

Robinson, A. C., Weaver, C., & Center, G. (2006). Re-Visualization: Interactive Visualization of the Process of Visual Analysis. In *Workshop on Visualization, Analytics & Spatial Decision Support at the GIScience conference*.

Rotem-Gal-Oz, A., Bruno, E., & Dahan, U. (2007). *SOA Patterns*. Manning Publications Co.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking (pp. 269-276). Amsterdam, The Netherlands: ACM. doi:10.1145/169059.169209

Schank, R. C. (1983). *Dynamic Memory: A Theory of Reminding and Learning in Computers and People* (1st ed.). Cambridge University Press.

Schank, R. C. (1999). *Dynamic Memory Revisited* (2nd ed.). Cambridge University Press.

Schnotz, W., Bannert, M., & Seufert, T. (2002). Toward an Integrative View of Text and Picture. In *The Psychology of Science Text Comprehension* (p. 385).

Shrinivasan, Y. B., Gotzy, D., & Lu, J. (2009). Connecting the dots in visual analysis. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (pp. 123-130). Presented at the 2009 IEEE Symposium on Visual Analytics Science and Technology (VAST), Atlantic City, NJ, USA. doi:10.1109/VAST.2009.5333023

Simon, H. A. (1996). *The Sciences of the Artificial - 3rd Edition* (3rd ed.). The MIT Press.

Slade, S. (1991). Case-based reasoning: A research paradigm. *AI magazine*, *12*(1), 42.

Stasko, J., Görg, C., & Spence, R. (2008). Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, *7*(2), 118-132. doi:10.1057/palgrave.ivs.9500180

Sumner, T., & Dawe, M. (2001). Looking at digital library usability from a reuse perspective. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries* (pp. 416-425). Roanoke, Virginia, United States: ACM. doi:10.1145/379437.379742

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social & behavioral research*. Sage Publications, Inc.

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social & behavioral research*. SAGE.

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

Tolcott, M. A., & Holt, V. E. (1987). *Impact and Potential of Decision Research on Decision Aiding*. Retrieved from http://stinet.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA192276

Trigg, R. H., Suchman, L. A., & Halasz, F. G. (1986). Supporting collaboration in notecards. In *Proceedings of the 1986 ACM conference on Computer-supported cooperative work* (pp. 153-162). Austin, Texas: ACM. Retrieved from http://portal.acm.org.proxy.lib.sfu.ca/citation.cfm?id=637069.637089&coll=portal&dl=GUIDE&type=series&idx=SERIES296&part=series&WantType=Proceedings&title=CSCW

Viégas, A. B., Wattenberg, M., Van Ham, F., Kriss, J., & Mckeon, M. (2007). Many eyes: A site for visualization at internet scale. *Proceedings of Infovis*. doi:10.1.1.91.7922

Viegas, F., & Wattenberg, M. (2006). Communication-Minded Visualization: A Call to Action. *IBM Systems Journal, 45*(4).

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Weaver, C. (2004). Building highly-coordinated visualizations in Improvise. In *Proceedings of the IEEE Symposium on Information Visualization* (pp. 159–166).

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409.

Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity* (1st ed.). Cambridge University Press.

Wolfinger, R. (2008). Plug-in architecture and design guidelines for customizable enterprise applications. In *Companion to the 23rd ACM SIGPLAN conference on*

Object-oriented programming systems languages and applications (pp. 893–894).

Wright, W., Schroh, D., Proulx, P., Skaburskis, A., & Cort, B. (2006). The Sandbox for analysis: concepts and methods. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 801-810). Montréal, Québec, Canada: ACM. doi:10.1145/1124772.1124890

Yin, R. (2008). *Case study research: Design and methods*. Sage Pubns.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, *21*(2), 179-217. doi:10.1016/S0364-0213(99)80022-6

Zimmerman, D. H., & Wieder, D. L. (1977). The Diary: Diary-Interview Method. *Journal of Contemporary Ethnography*, *5*(4), 479-498. doi:10.1177/089124167700500406

APPENDICES

Appendix 1: Pre-test Questionnaire

1. Age:
2. Gender:
3. Nationality:
4. Course of Study (Degree and Program):
5. Have you ever contributed to any social/collaborative web-based information sharing systems such as Wikipedia, Yahoo answers?

   If yes, which system?

   If yes, how often? (How many times in a day/month/year)

More than 1/day	**1-6/week**	**1-3/month**	**1-11/year**	**Less than 1/year**

6. How many posts/files/videos/news/music items (on average) do you usually share on social networks such as Facebook, MySpace, etc.?

More than 4/day	**1-3/day**	**1-5/week**	**1-3/month**	**Less than 1/month**

7. Briefly explain your experiences with analytical reasoning tasks (e.g. analyzing news stories, participating in VAST contests, analyzing various datasets, etc.) ?

Appendix 2: Participants' Diary Structure

1. Notes during the analysis:
2. Briefly explain what did you do in this session?
3. Anything that comes to your mind about your analysis process:
4. Have you noticed anything about other analysts? Briefly explain anything that comes to your mind about other analysts & what they did:
5. Have you noticed any new functionality of the system in this session? If yes, briefly explain.
6. Have you found a new reason for using one of the UI components? If yes, briefly explain.
7. Have you found a new way of doing something? If yes, briefly explain.
8. Why did you use ...(if you did):
    a. Activity Stream:
    b. Narrative View:
    c. Graph View:
    d. Relevant artefacts:
    e. Search:
    f. Scratchpad:
    g. History:

h. Comments:

9. Any functionality that you were happy to have:

10. Any functionality that you wish you had:

11. Anything else that you want to mention:

Appendix 3: Post-test Questionnaire

1. For what purposes have you find the sequential/narrative view useful?
2. For what purposes have you find the graph view useful?
3. How do you rate the usefulness of Narrative view for gaining an overall understanding of expanded pieces of analysis?
   - Very useful
   - Somewhat useful
   - Not particularly useful
   - Not useful
4. How do you rate the usefulness of Narrative view for understanding details of expanded pieces of analysis?
   - Very useful
   - Somewhat useful
   - Not particularly useful
   - Not useful
5. How do you rate the usefulness of Graph view for gaining an overall understanding of expanded pieces of analysis?
   - Very useful
   - Somewhat useful
   - Not particularly useful
   - Not useful
6. How do you rate the usefulness of Graph view for understanding details of expanded pieces of analysis?

- Very useful
- Somewhat useful
- Not particularly useful
- Not useful

7. Briefly explain for what purposes narrative/sequential view was more useful than the graph view and for what purposes the graph view was more useful than the narrative/sequential view?
8. Which of the views of an expanded piece was more useful for getting a sense of it?
    - Sequential/narrative view was much more useful
    - Sequential/narrative view was somewhat more useful
    - Both were equally useful
    - Graph view was more useful
    - Graph view was clearly more useful
9. Which of the views of an expanded piece was more useful for understanding its details?
    - Sequential/narrative view was much more useful
    - Sequential/narrative view was somewhat more useful
    - Both were equally useful
    - Graph view was somewhat more useful
    - Graph view was much more useful
10. How do you rate the usefulness of Related Pieces panel for finding your own pieces of analysis?

- Very useful
- Somewhat useful
- Not particularly useful
- Not useful

11. How do you rate the usefulness of Related Pieces panel for finding other analysts' pieces of analysis?
    - Very useful
    - Somewhat useful
    - Not particularly useful
    - Not useful
12. How do you rate the usefulness of search feature for finding your own pieces of analysis?
    - Very useful
    - Somewhat useful
    - Not particularly useful
    - Not useful
13. How do you rate the usefulness of search feature for finding other analysts' pieces of analysis?
    - Very useful
    - Somewhat useful
    - Not particularly useful
    - Not useful

14. Which of the features of search and Related Pieces panel are more useful for finding your pieces?
    - Related Pieces panel was much more useful
    - Related Pieces panel was somewhat more useful
    - Both were equally useful
    - Search was somewhat more useful
    - Search was much more useful
15. Which of the features of search and Related Pieces panel are more useful for finding other analysts' pieces?
    - Related Pieces panel was clearly more useful
    - Related Pieces panel was slightly more useful
    - Both were equally useful
    - Search was clearly more useful
    - Search was slightly more useful