Analytic Provenance in Practice: The Role of Provenance in Real-World Visualization and Data Analysis Environments

Karthic Madanagopal
Texas A&M University

Perakath Benjamin
Knowledge Based Systems Inc.

Eric D. Ragan
University of Florida

***Abstract*—Practical data analysis scenarios involve more than just the interpretation of data through visual and algorithmic analysis. Many real-world analysis environments involve multiple types of experts and analysts working together to solve problems and make decisions, adding organizational and social requirements to the mix. We aim to provide new knowledge about the role of provenance for practical problems in a variety of analysis scenarios central to national security. We present the findings from interviews with data analysts from domains, such as intelligence analysis, cyber-security, and geospatial intelligence. In addition to covering multiple analysis domains, our study also considers practical workplace implications related to organizational roles and the level of analyst experience. The results demonstrate how different needs for provenance depend on different roles in the analysis effort (e.g., data analyst, task managers, data analyst trainers, and quality control analysts). By considering the core challenges reported along with an analysis of existing provenance-support techniques through existing research and systems, we contribute new insights about needs and opportunities for improvements to provenance-support methods.**

Digital Object Identifier 10.1109/MCG.2019.2933419
Date of publication 5 August 2019; date of current version 1 November 2019.

Published by the IEEE Computer Society

Many practical forms of data analysis involve extracting meaning and insight from diverse and distributed data by progressively reducing the ambiguity in the available information.[1] While analytic techniques and advancements in data science assist by automating certain aspects of the analysis pipeline, the complexity and uncertainty needed for real decision making makes human involvement significant and inevitable.[2] Visualization techniques are often combined with analytic methods to support the integration of human and machine capabilities for data exploration by enhancing sense-making and cognitive capabilities with various types of visual and interactive systems. In the past two decades, a large body of research has been done in the visualization, human–computer interaction, and intelligence communities to understand the problems of data analysis and to develop various innovative solutions to address the needs of the analyst.[3] Significant research has advanced the effectiveness and performance of computational algorithms, and the refinement of new visual designs provides benefits for meaningful extraction of information from data.

Yet, practical data analysis scenarios involve more than just the interpretation of data. Many real-world analysis environments involve multiple types of experts and analysts working together to solve problems and make decisions, adding organizational and social requirements to the mix. Through the processing workflow, the data itself undergoes changes and transformations as it moves through the computational pipeline. The already-complex processes for data analysis are interwoven with further challenges relating to the need for verification, collaboration, and communication. It is no surprise, then, that many efforts in research and practice have focused on supporting the history or *provenance* of data analysis.[4] In general, analytic provenance refers to the history of changes to data, system state, user interactions, and human understanding during analysis.[4]

Many studies have analyzed the process of data analysis and derived needs and requirements for data analysis systems.[4,5] While provenance is generally believed to be of critical importance across an array of areas in data science, the utility of provenance is also different across domains and organizations. Previous work has highlighted the complexity of the many different perspectives and purposes for provenance,[4] but there remains a need to understand how the needs and purposes for provenance apply in real analysis environments.

In our research, we aim to provide new knowledge about the role of provenance for practical problems in a variety of analysis scenarios central to national security. We present the findings from interviews with data analysts from domains, such as intelligence analysis, cyber-security, and geospatial intelligence. In addition to covering multiple analysis domains, our study also considers practical workplace implications related to organizational roles and the level of analyst experience. By considering the core challenges reported along with an analysis of existing provenance-support techniques through existing research and systems, we contribute new insights about needs and opportunities for improvements to provenance-support methods.

Thus, our study addresses the following.

1. Summarize the needs and requirements for provenance support across different roles and domains in real-world national security data analysis scenarios.
2. Identify the primary purposes and methods where provenance is currently captured and utilized.
3. Identify possible opportunities for advancements and new techniques for provenance support in analytic and visualization systems.

Research objectives 1 and 2 are addressed in the Role of Provenance by Tasks section as findings from our interview. The Opportunities section addresses Research objective 3.

RELATED WORK

Before presenting our new findings about the place of provenance in practical data analysis environments, we will provide an overview of related studies and tools.

Sensemaking and Intelligence Analysis

Exploratory data analysis involves deriving useful insights from diverse and distributed sources. It generally involves fusing conflicting or incomplete data from multiple sources and validating against various competing hypotheses. A large and growing body of literature has done investigations on understanding the analysis process and prescribing strategies and designing algorithms to efficiently collect, process, and analyze data.[1] Researchers interacted with various analysts and developed process models that help to understand the cognitive nature of the analysis process.[6] Most of them are iterative and nonlinear in practice.

Data exploration is an intense cognitive task, and recall is an important aspect toward formulating various hypotheses to solve a problem.[3] Cognitive ability and limitations are an important aspect of intelligence analysis, as it influences balancing limitations in human processing with strategies for leveraging experience and expertise.[7] Cognitive psychology is a scientific field that focuses on studying various factors that contribute to the information processing aspect of the human brain. Cognitive limitations can be a major reason why people sometimes fail to link decisions with evidence.[2]

Analysts are constantly challenged to work on multiple tasks at a time or sometimes asked to work on quick response tasks. During quick response tasks, an analyst will have to completely stop what they were doing and start working on the new, high priority task. After the quick response task is done, they will have to go back to their previous task. Limited attention is one of the major factors that limits a human's ability to perform multiple tasks at the same time. Attention is generally considered to be a cognitive effort that depends on various factors like environment, type of tasks, and the individual's mental state.[6] Since intelligence analysis is a complex cognitive process, it is not easy to get back to the previous state of mind quickly. During the analysis process, analyst takes quick notes to help them recall important information in future.

Knowledge elicitation is a crucial step in the context of psychological research in understanding the experts.[8] There is a large volume of published studies describing various knowledge elicitation methods to understand how experts perform their mental process in the field of decision making, information analysis, etc. However, there is a lack of research in the area of visual analytics on how to support knowledge elicitation in complex data analysis. Most of the data analysis enterprises work on multiple shifts with multiple analysts working on the same task in different shifts. In those kinds of scenarios, every analyst will have to transfer their knowledge captured during their shift to the next shift analyst and vice versa.[1] Developing more rational processes and supporting them with good interface designs has proven to garner better results in enhancing the decision-making behavior of human analysts. Our research focuses on the role of provenance for operational activities in data analysis environments, which not only include data analysis, but also account for broader associated activities of real-world settings.

Analytic Provenance

A considerable amount of research has been done on capturing behavioral data during the analysis task that could be used for various purposes, such as externalizing analyst thought process,[5] reproducability of analysis workflow,[9,10] recall analysis steps, and results.[11] These types of interaction histories that help an analyst to recall the computational sequences and the states traversed to arrive at the result was termed as analytic provenance by the visual analytics community.[4] Various studies have shown the value of interactions and their benefits in information sense making.[12,13] Detailed studies on the definitions of provenance and their categorizations were discussed in various papers.[4] Ragan *et al.*[4] presented an organization framework by studying different types of provenance and their purpose. Most of these provenance based studies were mainly focused on understanding provenance from task perspective, i.e., how provenance is useful for the task to be done effectively. Human mental process is one of the most important factors in performing exploratory data analysis. The reflection of insights gathered at each step is very important in progressing to the next step. Since the reflection happens in the analyst's head, there is no direct way to capture this information. We wanted to understand the concept of provenance and its purpose from the analyst's perspective. It will be helpful to understand how they are currently capturing and using provenance, as well as to identify the gaps in the technologies in terms of provenance support.

Visualizing or summarizing the provenance collected will help an analyst to have a quick understanding of the provenance data. Depending on the type and amount of provenance collected, understanding or making sense of the provenance data can be time consuming. CLUE (Capture, Label, Understand, Explain) provides a model to integrate provenance captured during the analysis stage and construct visual stories called "Vistories" using key terms and annotations.[14] Provenance is collected at each stage of the analysis in the form of interactive histories and evidences, and the insights and annotations done by analysts are important for reasoning.[13] With more provenance collected, analysis becomes very difficult. Ragan *et al.*[15] has cataloged different types of provenance and their purpose in the field of visual analytics. However, little work has investigated the sense making of provenance itself.

Provenance Support in Data Analysis Tools

Various visual analytic tools have been built with provenance support that help analysts to reason with various evidences collected during the analysis process.[13] A vast majority of the tools that have provenance support, provide provenance capture through manual annotations. The manual annotations expect an analyst to provide detailed information on each of the activities they perform on the data. Although the manual annotation method is very effective, one of the problems with this kind of approach is the lack of consistency. The quality of the provenance depends on the analyst's ability to annotate. In order to mitigate the issue of inconsistent annotations, various studies have been conducted to infer insight provenance from lower level interaction data like mouse movement, clicks, drag and drop, filter, etc. Visual history of the analysis steps have proven to be effective in process recall, and various methodologies have been developed to evaluate how visual tools influence process memory.[4] Automatic provenance capturing capabilities are built into various visual analytic tools like GraphTrail[10] and SenseMap.[11] Provenance collected during the exploration process has been used to understand the sensemaking process[11] and gauge the utility of visual analytic tools.[10]

Many studies have illustrated the value of history tools in the context of individual use[10] and also in group use, such as collaborative data visualization.[16] In collaborative data analysis, it is important for an analyst to know what analysis has been done and what is left. Visualization histories have proven to be efficient communication mediums in collaborative scenarios, and various studies have been performed in understanding the importance of dimension-oriented views in understanding provenance data.[16] Semantic models can be used to analyze lower level interaction logs and identify higher level functions like filtering, aggregation, bookmark, etc. Graphtrail

Domain	Yrs. of Exp	M/F	Role	Highest Edu. Level	Type of Analysis Data	Team Size	
						Min	Max
Intelligence Analyst	6	M	Data Analyst	Bachelor	Text & Quantitative Data	2	12
Intelligence Analyst	12	F	Analyst Manager; Qual. Ctrl	Bachelor	Text & Quantitative Data	2	3
Intelligence Analyst	18	M	Quality Control Analyst	Bachelor	Text & Quantitative Data	1	5
Intelligence Analyst	21	M	Trainer	Master	Text & Quantitative Data	1	2
Cyber Security Analyst	5	M	Data Analyst	Bachelor	Text & Quantitative Data	2	6
Cyber Security Analyst	11	M	Trainer	Bachelor	Text & Quantitative Data	1	1
Cyber Security Analyst	15	M	Data Analyst	Bachelor	Text & Quantitative Data	2	6
Geospatial Analyst	10	F	Analysis Task Manager	Bachelor	Video & Imagery	2	8
Geospatial Analyst	11	F	Quality Control Analyst	Bachelor	Video & Imagery	2	2
Geospatial Analyst	16	M	Analysis Task Manager	Master	Video & Imagery	2	4
Process Consulting	5	F	Data Analyst	Master	Quantitative Data	1	2
Finance	5	M	Data Analyst	Bachelor	Quantitative Data	1	2
Meteorology	8	M	Analysis Task Manager	PhD	Video & Imagery	1	5
Manufacturing	8	M	Data Analyst	PhD	Quantitative Data	1	1

Figure 1. Summary of participant backgrounds and analysis experience.

capture provenance in the form of workflow during data analysis and visual exploration.[10] These integrated workflows help to recall the exploration history in an intuitive interface, though unfortunately they are limited to sequential workflows. Given that a majority of the analysis tasks are iterative in nature, this kind of approach is restrictive.

APPROACH

In order to explore the needs and requirements of the data analyst and identify areas of concern that are not entirely addressed with existing tools, we conducted semistructured interviews with data analysts from different domains. Our study aims to understand a holistic picture of the various tasks performed by analysts in their day-to-day operations and study how provenance data is currently used and could be better supported to improve job efficiency and quality.

Participants

In order to make this a comprehensive study, we selected a diverse cross-section of data analysts from various domains, age groups, and expert levels (see Figure 1). The participants are primarily selected from three different domains: first, intelligence analysis, second, geospatial-imagery analysis, and third, cyber-security analysis. Four of the participants are from other domains, such as finance, manufacturing, process consulting, and meteorology. The type of data and the characteristics of data analysis are different between these domains. Intelligence analysts synthesize data from multiple sources and derive insights from the fused intelligence; it is one of the primary domains that deals with textual data. Geospatial analysts primarily deal with analyzing satellite imagery for various tasks, such as threat monitoring, weather forecasting, etc. Cyber-security analysts monitor network logs on a real-time basis to uncover any security incidents in the network. Multisource information fusion, unstructured data analysis, and real-time analysis are the three areas that differentiate these three groups of analyst.

All participants chosen for this study possess at least five years of experience in data analysis and they have played different roles during their data analysis career: first, Data Analyst, second, Analysis Task Managers, third, Data Analyst Trainers, and fourth, Quality Control Analyst. Some of the participants have been in multiple job profiles at some part of their career. All participants had completed at least a bachelor's degree, and everyone had sufficient skills and experiences using computers in their everyday operations. Some of the experts were already known to us through various collaborations in the past, and others were recommended by other analysts.

Interview Procedure

Data collection was based on a semistructured interview with our expert participants. Interviews followed an interview guide consisting of sets of questions organized into key areas of interest, though the procedure was flexible to allow the participant to provide additional information or allow interviewers to ask for further explanation or elaboration, as necessary. In other words, questions were meant to kindle the discussion, and off-topic elaborations were allowed during the interview. While our study aimed to study the role and uses of provenance in operational settings, we note that we did not explicitly introduce the term *provenance* in our interview questions to avoid leading questioning.

To aid data collection, interview questionnaires were sent to the analysts a day before the interview and the interviews were conducted over phone or video conferencing with the exception of four who were able to be interviewed in-person at their work location. Each interview lasted for about 60–90 min. Five analysts who could not finish all the questions because of the time constraint were requested to send the filled-in questionnaire to us through email. Even though the questions that were asked during direct interviews are the same as the ones sent through email, the written responses were not as detailed as the direct interviews. During direct interviews, we did not strictly follow only the questions that were formulated. Sometimes we asked detailed questions to better understand the responses from participants. To avoid concerns regarding sensitive data, capture notes were taken during interviews, but no audio or video was recorded.

The interview was designed such that approximately 10 min was allocated for questions or discussion related to each topic. A brief introduction was provided at the start of each section before beginning questioning. Due to the flexibility of the semistructured approach and the flow of the interview, we noted that questions and aspects of topics were not always covered strictly in the same order for all participants. In addition, since many practical data analysis domains relate to sensitive information, participants were encouraged to provide generic examples and descriptions of analysis and operational practices to avoid any potential issues with information disclosure. Interview topics and questions were available in a supplemental section.

Method of Results Analysis

Hand-written notes were taken during the interview and converted into a digital document afterward. For a few interviews (5) that could not be completed due to time constraints, we sent questions to the analyst, and they provided additional information by emailing the completed questionnaire. We cleaned up these responses and combined them with our interview notes.

To understand the analysts' operations and the intersection with provenance, we followed a process similar to theoretical sampling[17] with multiple rounds of coding generated conclusions by analyzing the collected data using multiple rounds of data review. The analysis was conducted by a single coder. First, we performed open coding by reading through the notes line by line and tagged entities, tasks, job profiles, and problems included in their responses, along with marking key sentences. Then, tentative labels were created for key sentences that summarize the theme. Next, we performed a separate full pass of descriptive coding in which we labeled statements with descriptive summaries. Various themes were derived by analyzing the descriptive codes. Within each theme, we organized comments based on the associated tasks and problems.

FINDINGS

In this section, we summarize the major findings distilled from our analysis of interview data. Participants provided a breadth of information about analysis processes and workplaces. Here, we focus on the primary analysis tasks and challenges most relevant to workflow and provenance issues in the realm of national security analysis. While the focus is data analysis, issues extend beyond simple isolated cases of human–machine interactions. Differences in tasks, analysis positions, and levels of experience influence the implications and applications for provenance.

One of the themes that emerged, focused on various tasks that are performed by data analysts on a day-to-day basis. We did a deeper study of the tasks and compared them against

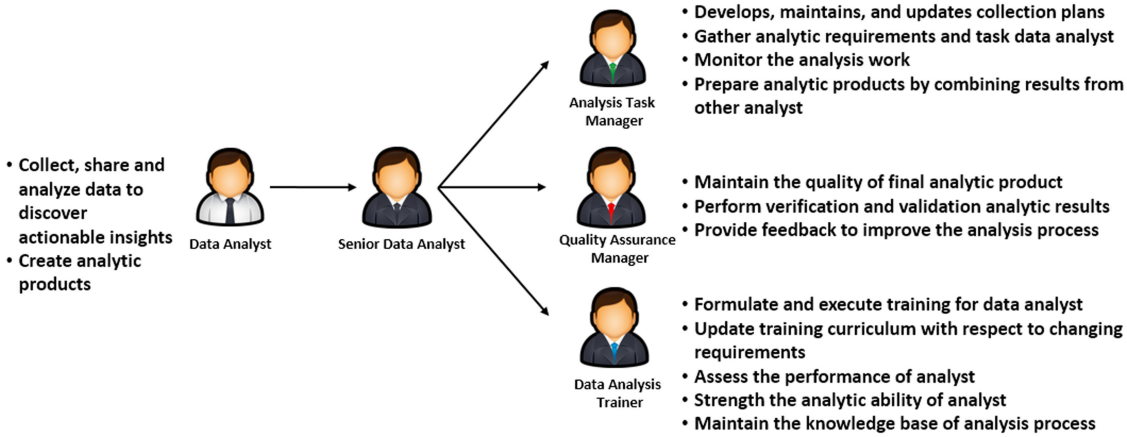

Figure 2. Different analyst role types identified during interview process.

analysis processes that were studied by various researchers in the past.[5,18] We then marked down the tasks that will have significant impact if the provenance or logging data are available. Detailed descriptions of the identified key tasks are presented in the Role of Provenance by Tasks section. Another theme that emerged from the identified tasks are the analyst job profiles. There appeared to be a clear separation of job responsibilities among the identified job profiles. We searched online for various data analyst positions and collected their job descriptions. The collected online job descriptions are compared against the job descriptions we obtained through our interview process. Deductive approach is followed to elaborate on the evidence that was collected for each of the identified tasks. Detailed descriptions of the identified analyst job profiles are presented in the Roles and Careers in Analysis Work section.

Roles and Careers in Analysis Work

The primary job responsibility of a data analyst is to synthesize data from multiple sources and derive useful insights. Generally, data analysts get their tasking from their data analysis manager or their operations manager. Depending on the type of data used for analysis, analysts can be divided into subcategories, such as geospatial intelligence (GEOINT) analyst, intelligence analyst, all-source analyst, cyber-security analyst, etc. The analysis type and the tools used for analysis are greatly affected by the type of data, but the objective of data analysis still remains the same—progressively analyze the available data and derive useful insights.

Often, research considers that the primary responsibility of a data analyst is to follow the analysis process and uncover the truth hidden in the data, but this is not the complete picture. Our interviews indicate that there are various types of tasks that are performed by analysts apart from data analysis itself. Apart from the core responsibility of performing data analysis, there exists other job responsibilities that a certain class of data analysts are expected to fulfill.

- *Data analysts* are responsible for collecting, organizing, and analyzing data from multiple sources.
- *Senior data analysts'* primary duties are to validate the results of other data analysts and create analytic products/reports.
- *Data analysis managers or operations managers* are responsible for gathering requirements and giving tasks to other analysts.
- *Quality control analysts* are responsible for verifying the analysis reports that are generated by data analysts and making sure that the data and the results are valid and up to standards.
- *Data analytics instructors or trainers* are responsible for assessing performances of analysts, gauging the effectiveness of analysis operations, providing hands-on exercises, and introducing analysis tools.

Figure 2 shows a summary of roles with a common career progression. Different positions

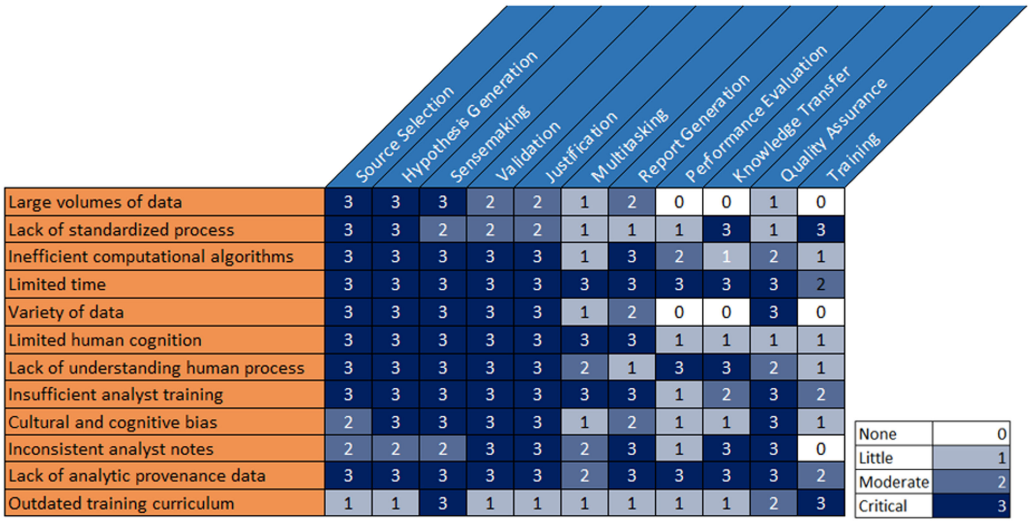

Figure 3. Mapping of analytic tasks to analyst problems.

focus on different tasks, but each task is integral to the holistic data analysis process. Recently, Alspaugh *et al.*[19] interviewed data analysts from various domains and discussed the various challenges involved in performing different types of data exploration activities. A majority of the studies[20] were conducted to better understand data analysis processes in various domains and did not focus on understanding the difficulties faced by different types of analysts in performing their day-to-day activities. Little research has been done in understanding different job duties performed by an analyst and their challenges in performing their day-to-day operations. We primarily focused on studying the challenges in performing various analysis tasks and whether provenance based solutions are helpful in performing their job duties.

Tasks and Challenges

For each of the analytic tasks that were determined through the interviews, common challenges and issues were identified. After our coding analysis of interview results, we conducted a secondary round of data collection with the participants to better understand the relationship between analytic tasks and identified problems. We created a scoring sheet and circulated it to all our participants to get their feedback. In our scoring sheet, possible relevance levels that can be assigned for each mapping are *critical*, *moderate*, *little*, and *none*. To establish a common understanding of terms, participants were also given descriptions of items for clarification. A total of 11 out of 14 participants returned scoring responses, which we used to create the tables presented in Figures 3 and 4. For each mapping, we identified the majority relevance level and assigned it as the final relevance score. The final relevance mappings between analytic tasks and analyst problems are presented in Figure 3.

From the mappings (Figure 3), it is quite evident that the lack of analytic provenance data is seen as an underlying problem for a majority of analytic tasks. We were a little suspicious about the results because it seems too good to believe that lack of analytic provenance data is

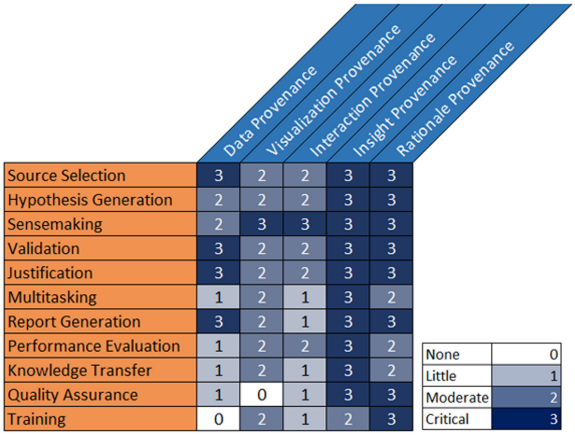

Figure 4. Mapping of tasks to types of provenance information.

considered critical for almost all the tasks. We reached out to our participants and understood their reasoning behind the relevance score. Details of our findings and our participant feedback are presented in the Role of Provenance by Tasks section.

Role of Provenance by Tasks

In this section, we summarize the core task areas related to the data analysis process as determined through our interviews, and we discuss the role of provenance based on existing challenges and approaches common in each area.

Source Selection An analyst will have to understand clients' requirements and identify the best sources that could help them in answering their analysis questions quickly and with high confidence. Most analysts mentioned source selection as a critical task, as the chosen source will inevitably influence the results of the analysis. While our participants were a mix of people who have worked on postevent analysis, real-time analysis, and preevent forecasting, source selection was common across the board. Data profile is a common name that refers to a list of preferred or identified data sources. Every analyst will start with a small set of standard data profiles and add additional sources to it based on their experiences in different scenario. Some analysts keep their data profiles private because it is considered as a secret sauce that differentiates them from other analysts.

Some of the participants with less than eight years of experience felt that the senior analysts have an edge over them in terms of the quality of analysis mainly because of the rich data profiles. The analysts are trained on algorithms and procedures that will help in source selection, such as use of quality of information (QoI) metrics, trust metrics, etc., but still every analyst has their own source selection strategies. When asked specifically about their methods, one of the analysts mentioned that having a comprehensive view of all the data will enable them to filter and focus on the data sources that are more valuable to their analysis. Analysts in the cyber-security domain had an issue with this method because the real-time nature of their work means that not all the data are available at once and it is difficult to get a complete picture.

In the intelligence domain, not every analyst will have access to all data, and sometimes a separate group of analysts (source collection analysts) are assigned the task of finding new data relevant for the analysis problem. Many of the participants lacked proper training in source selection methods. Due to the variety of strategies, it can be helpful to capture different methods and come up with a set of best strategies that every analyst can use. From our participants, we found that the current method of understanding source selection strategies are done manually by conducting interviews with experts. We contend that capturing analytic provenance or interaction logs can be used to infer the source selection strategies that are used by different analysts.

Hypothesis Generation Analysis of competing hypothesis (ACH) is one of the analysis methods that was commonly referred by most of the participants. ACH helps to avoid cognitive bias by generating various competing hypothesis and validating each purely based on the evidences available.[3] Even though ACH is efficient in supporting better analysis by covering more ground, analysts with less experience reported difficulty in the following ACH. The challenging task in performing ACH is coming up with an exhaustive list of hypotheses, as very few have the rare skill of hypothesis generation that one mainly gains through experience. Our participants felt that the experience gained by working on various analysis problems over the years helps to acquire the skill of generating various hypotheses. Our participants with less than eight years of experience mentioned that they struggle to come up with competing hypotheses and have very little support in terms of technology.

Some of the analysts mentioned that there is a danger of people being biased in the name of hypothesis generation. Analysts need to be careful in selecting a hypothesis, and this is commonly resolved through constant validation with fellow analysts. Many hypothesis generation techniques are done through statistical inference of looking at the input and the output of various analysis results. Understanding the

intermediate steps will help in coming up with algorithms that are aligned to the mental model of an analyst. From the limited set of participants we interviewed, we realized that the current generation tools lack the ability to capture various hypotheses tested by the analyst. Even the tools that capture the hypothesis do not have an intuitive way to present the data.

Sensemaking Through our interview, we realized there are various perspectives of sensemaking among the analysts. One of the analysts mentioned that sensemaking is the process of resolving contradictions that were identified in the data. "Connecting the dots" is a common metaphor that was referenced by analysts when discussing the sense-making process. Another analyst mentioned "Data analysis is a process of making data tell the story." In a nutshell, what our participants meant by sensemaking is the process of understanding the data and using the information derived to answer the analysis question. Various known factors were coined during the interview that affects the sensemaking process, such as lack of data, contradictions in the data, volume and veracity of data, and unstructured nature of data. Out of various themes that emerged, we focused mainly on the challenges where provenance data can be useful in training an analyst on the sensemaking processes.

Training an analyst on how to think is a challenging task as different analysts have different backgrounds and skills. Understanding the thought process of an analyst is imperative when training another analyst on how to perform the same task. Numerous studies have been done in the past to understand an analyst's thought process[3,7,18] and various methods have been prescribed,[5,11] but it is still an unsolved challenge due to the complexity of analysis and human problem solving. Most of the studies in understanding the thought process of an analyst is done either through expert interviews[5] or exercise based studies using think-a-loud observations.[5,15] Our participants felt that the current method of understanding sensemaking model is shallow and does not provide additional information other than the models prescribed in the existing body of the literature.

Collaborative sensemaking is another topic that was commonly referred by our participants during the interview. Given the information-rich, complex, and dynamic environment, most of the analysts mentioned working as a team is a much more effective way to quickly make sense of the data when compared to an analyst working alone. Since the importance of collaborative sensemaking is well studied by various researchers, we focused mainly on how provenance data could be helpful in collaborative sensemaking. Our participants mentioned efficient sharing of data, results, and details about analysis steps are the most important aspects of collaborative sensemaking. A few of the analysts mentioned that their analysis counterpart worked on other shifts and at remote locations. In those situations, the analysts were required to document the entire analysis process at the end of each shift, so that the other analysts will have enough information to continue the analysis. Participants mentioned that the new table-top interfaces that are being used in their organization are useful in performing collaborative analysis and also supports sharing virtual boards with analysts located remotely. Even though new tools are available to support collaborative sensemaking, some of our trainer participants felt that those tools were too primitive to help us to understand the sensemaking process. Since the current generation of collaborative tools are designed only to provide a platform for analysts to collaboratively work together, the rich data that are captured by these tools are less effectively being used. Less work has been done in exploring the use of provenance data collected during the collaborative sensemaking process.

Validation Analysts who had experience with validating analytic results mentioned "Thinking about thinking is an important business and is often overlooked." Red-teaming our strategy for sense making and where we might have bias, vulnerability to deception, or be wrong is often not programmed into the manpower studies, and it is the first thing to fall off. A major percentage of our interviewees defined data analysis as a process of validating certain assumptions through

evidence collection. Every analyst goes through a series of validations, as they discover key insights from the data. First, the evidence collected during the analysis is validated against the set of hypotheses. Self-validation is very important in data analysis because it guides the next steps in the analysis process. Then, the validation is done by consulting with senior analysts or quality control specialists. A senior data analyst pointed out that the final results are not good enough to validate the analysis, and the steps in which the analyst arrives at the results are very crucial. It is important to make sure that the analyst has explored all areas of the data and that there are no flaws in their hypothesis. The only way to make sure that the analysis is complete is by exploring the provenance of the data analysis. There is a lack of consistency in the level of provenance that is captured by analysts. This lack of consistency is mainly caused by the human in the loop nature of provenance capture systems. If the provenance of analysis is not captured or presented in a standard way, it is very tedious to validate any analysis findings.

Identifying and neutralizing bias is another challenge in data analysis. Hindsight is an important aspect of the human brain that helps an analyst to uncover the true facts that are hidden in the noisy, incomplete data. Not all hindsight has proven to be useful, but in cases of negative outcomes, especially in the case of intelligence analysis, hindsight has been proven to have a greater impact.[3] This is mainly because of the human minds natural tendency to focus on things that have negative outcomes. Given that hindsight bias is useful in some contexts and not useful in others, makes validation a challenging task.

Knowledge Transfer Most of the participants who were specifically intelligence analysts and cyber-security analysts worked in shifts. In those environments, the same analysis tasks are shared by groups of analysts who work at different shifts. At the end of every shift, an analyst will have to share their results with their counterpart who will continue the analysis in the next shift. Analysts who have worked on those kinds of scenarios mentioned that every day they spend about an hour doing the briefing, which is very inefficient, but necessary to do their collaborative work. When an analyst is doubtful of the results obtained by their counterpart, they will have to wait a whole shift for them to be available. Some of the participants mentioned that knowledge transfer between work shifts is their primary problem and any solution that can help to improve will save time and increase their productivity.

Performance Evaluation and Training Performance evaluation is a formal and productive procedure that helps to determine the skill level of an analyst and identify training requirements to improve their performance. Data comprehension, critical thinking, systems analysis, judgment and decision making, active learning, complex problem solving, attention to detail, and written communication were some of the required skills mentioned by our participants. When it comes to data analysis, many skills are interwoven. Sensemaking of the big picture is different than persuasively writing it up. Writing is different than giving a live presentation and handling follow-up questions in-person. The key performance indicators in the data analysis community are specificity, timeliness, accuracy, relevance, and clarity. Trainers generally do scenario based evaluation to better understand the strengths and weaknesses of an analyst. The testing scenarios are also not updated over a long time, which makes the evaluation less accurate. The trainers felt that creating a summary of the work done by an analyst that encompasses the evaluation factors would be required.

Training is a process through which the analytic capabilities of analysts are strengthened as the operational scenario changes. Standard training curriculum is generally used to train entry level analysts, which includes the best practices that have been followed by seasoned analysts for years. There are initial courses that usually focus on some sort of basic creative thinking, but mostly on data types, how to get them and understand them, and an introduction to the target systems the new analysts will be involved with. It is often encyclopedic and the biggest challenge is that there is no unanimity on how to best train an analyst. Most of the analysts felt on-the-job experience is the best

training ground to hone their skills. Brainstorming with senior analysts is another way to learn new skills on the job. Like the performance evaluation, identifying and updating training curriculum should be organically derived from the analyst rather than artificially adapted from other domains. The biggest challenges are how to level the experiences of a team of human beings, all with different gifts/abilities and preferred work styles.

Quality Control The primary goal of the intelligence community (IC) is to generate intelligence reports that will be used by policy makers and decision makers. Because of its high stakes, it is important to make sure the intelligence reports are of high quality. Organizations that primarily focus on data analysis tasks will have quality control teams that are responsible for assuring the quality of the final analysis report by making sure all the data points are verifiable, data sources are trustworthy, and the conclusions derived by the analyst are plausible with high standards. In order to assess the verifiability of the intelligence reports, every analyst is expected to maintain high standards of fact checking and reference maintenance. Specificity, timeliness, accuracy, relevance, and clarity are the same rubrics used to evaluate the quality of data analysis reports.

Some of our senior analyst participants mentioned that in the past, the quality check on the reports are done at the final stages of the analysis, but in more recent times it has been integrated into the analysis process, where the data analysis tools themselves have built-in quality control checks, which makes it easy to maintain high quality standards. Generally, the quality assurance team is primarily comprised of senior data analysts. Analysts by nature have the desire to understand every detail on how the results were obtained. Some of our participants mentioned that quite a bit of time is spent in maintaining high quality of documentation to support the quality control process. Improved techniques could increase the quality and utility of annotations to enable the capture of provenance information that might not be automatically captured—such as analysts' rationale, hypotheses, and findings.

Reporting The main purpose of the report is to present all key insights that are derived from the data with proper justifications. The level of detail for the reports generally differs by the target audience. In the case of postevent analysis kind of scenarios, the reports are very detailed. Most of the analysts perceived analysis as a process of building arguments, and all of the analysis reports need to meet the military standards for specificity, timeliness, accuracy, relevance, and clarity (commonly referred to as STARC). The report needs to cover every hypothesis that is considered, and justification needs to be provided for the conclusive hypothesis.

Most of the data analysis tasks are complex, and it takes longer periods to complete the analysis. The longer period of analysis causes people to forget some of the main steps or the steps that helped them arrive at the results. In those kinds of situations where an analyst does not remember all the analysis steps, report writing becomes challenging. Depending on the type of organization, the report goes through rigorous review by the quality control team. The quality control team consists of a set of senior analysts that work to make sure the analysis satisfies the core requirement and is presented in a proper way.

Multitasking Traditionally most of the data analysis tasks are done in isolation where an analyst works on one task and is not disturbed while performing their work. Numerous analysts during our interview mentioned that they are expected to multitask irrespective of whether they are doing a complex or a simple task. One of the intelligence analysts specifically mentioned about quick response tasks, where some high priority tasks are assigned to analysts and they are expected to work on it immediately. Even though the quick response tasks are critical from a mission point of view, from an analyst's point of view it affects the momentum of their analysis process. Depending on where an analyst is in the analysis workflow, an analyst will have to take quick notes before switching to a quick response task. The notes taken before switching tasks play an important role in determining how much time it

takes for an analyst to get back to the same state of thinking. Some cybersecurity analysts who perform network analysis specified that some of the analysis tools they use can take a snapshot of their network structure during exploration and reviewing those snapshots help them in getting back to their task in quick time. The snapshots that are captured by the exploration tool is an example where the provenance information can make switching between tasks easy and efficient.

Resolving Annotation Inconsistency One of the common strategies toward solving a new analysis problem is to look out for previous analyses that have some similarities with the current task. Reports and other forms of documentation help to understand the assumptions and details of the analysis process. Apart from the general reporting standards, there are no strict requirements for the level of detail that needs to be captured in the analytic products. One of the major issues pointed out by our analysts is the inconsistency in the level of details captured in the documentation or reports. Some of the documentation only contains the input data and the final results, but it rarely covers sufficient details of the analysis process. Apart from analytic product and its documentation, there are no other provenance logs that are captured to understand the analytic process. One of the analysts, who is currently active in performing data analysis, mentioned that he prefers script based analytic platforms like Matlab and R instead of visual analytic tools. Instructions written on the scripts are the only way to manipulate data and the instructions themselves are self-explanatory in detailing the analytic process, whereas the visual analytic tools are efficient in understanding the data through powerful visualization and interactions, but often lack the ability to capture the steps involved in arriving at the final results without explicit analyst annotation.

OPPORTUNITIES

Based on an analysis of the previous literature along with the findings from our interviews, this section identifies opportunities for provenance support during different aspects of collaborative intelligence analysis. In particular, we identify and describe opportunities where visual analytics support for analytic provenance will likely provide high value for practical needs. In Figure 4, we mapped the tasks identified from our analysis (Role of Provenance by Tasks section) to the provenance types described in previous research by Ragan *et al.*[4] From the table, we can infer the type of provenance needed by analyst role types. For example, *rationale provenance* is required for *quality control analysts* to understand the reasoning and intent behind decisions made during the analysis, and *insight provenance* is required by all types of analyst for knowledge transfer and process recall.

From the mapping of analytic tasks to analyst problems (Figure 3), it is evident that the provenance data are critical for most of the analytic tasks, where many tasks are heavily associated with multiple issues, whereas others (e.g., training) have more specialized needs. Provenance-based design aspects, such as provenance capture, provenance storage and retrieval, and provenance visualization need to be designed with the identified tasks in mind. We contend that a good understanding of the need for provenance support over particular tasks can guide design requirements to shape future research and development of provenance visualization and management tools.

Provenance Capture

Several of the intelligence analysis requirements that were identified in the previous sections may be addressed through the capture of activity logs and insight provenance by visual analytic tools. Analytic provenance captures the history and lineage of all the actions performed by an analyst during a data exploration process. The captured analytic provenance may be used to reproduce the computational sequences that are performed by an analyst in arriving at the results. Existing tools and prior research efforts have focused on supporting the capture of provenance, but this not a solved problem. Automated methods are imperfect, and there is a lack of clarity on the details of the provenance that needs to be captured. The interfaces for future visual analytic systems need to be designed with the provenance capturing requirement as a

significant design focus. Unlike data provenance, analytic provenance is still at a relatively nascent stage where standards and best practices are yet to be defined.

Annotations added in the analytic product while performing data analysis represent an efficient and effective way to capture important exploration process knowledge. Support for annotations in current data analysis tools and analysis toolkits usually manifest as free text editors with some support for attachments. Some of the analysts we interviewed mentioned the need for auto-annotation support in terms of templates augmented with context questions. Instead of an analyst writing annotations from scratch, the system can identify the context of the annotations and prescribe annotation frames that may be easily and rapidly filled out by an analyst. These annotation frames may be derived by processing previous annotations submitted by the analyst. Such annotation support using assistive writing also solves the problem of inconsistent annotations.

Provenance Management

Management of provenance involves categorizing the types of provenance, identifying representation mechanisms, and building scalable storage and retrieval methods. Improved visual analytic support for provenance management would be beneficial for various data analysis tasks across numerous roles. For example, trainers could use provenance data to evaluate the analyst performance and quality control analysts can use provenance data to validate analysis results by understanding the reasoning and intent behind decisions made during the analysis.

Various prior studies have identified different types of provenance and their significance with respect to the types of information that may be derived from the provenance.[4] Depending on the type of provenance and the set of attributes it supports, retrieval methods need to be appropriately designed. Support for heterogeneous properties and efficient search and maintenance of temporal coherence are two of the key objectives that have to be considered for the design of interfaces for provenance retrieval.

Presentation and Summarization

Our study found challenges with communication, and presentation of provenance information is affected by issues with limited time, difficulty in provenance capture, data scale, and limited algorithmic support, which makes these areas ripe for potential improvement. Provenance data that are captured for analytic processes are often considered as state diagrams, where the state of the data changes after each interaction. Usually the analytic provenance is visualized as graphs using nodes and branches. Branches represent pivot points where the analyst tried different hypotheses. As the size of the graph grows big, it becomes increasingly difficult to understand the provenance graph. New scalable visualization methods need to be developed to visualize analytic provenance in a simple and intuitive fashion. When the volume of data that is captured for analytic provenance is moderate, it is easy to synthesize and derive useful insights. As the size of the data grows, managing and making sense of the collected provenance becomes a key challenge. More focus on the computational methods and visual analytic tools for provenance sensemaking is required. Depending on the type of information that needs to be derived, summarization methods are needed.

Machine learning has become an integral part of visual analytics in extracting patterns from data that will be rendered as intuitive visualizations to enable sensemaking. Applications of machine learning algorithms in extracting patterns from provenance data are less studied by the research community as compared to other research areas, such as image analytics, topic and sentiment modeling from text, etc. If machine learning algorithms are the tools to understand the user's mental model of the analysis, then, provenance data are the glue. Provenance capture and management functionality of visual analytic tools needs to be carefully designed with consideration to how the collected provenance will be used by various algorithms and visual analytic tools that align with the analyst's mental model, which may vary depending on the analyst role or current state in the analysis pipeline. For example, active data analysts may wish to see more detailed

records of the analysis process, whereas managers may benefit more from a higher-level summarization.

Adaptive Training

Our findings indicate that the training will likely help improve the productivity of an analyst. By analyzing analytic provenance from past analyses, efficient methods could help identify areas of concerns for different analysts during training. Adaptive training could be beneficial by helping analysts undertake remedial courses or exercises in specific areas that have been identified as their weak spots. There would also be potential benefits to performing utility assessments of tools on a frequent basis using end-users' behavioral data. Provenance captured during the analytic process—when compared against process benchmarks—could help understand where an analyst is performing well and where additional training is needed. Similar approaches may be developed to derive curriculum requirements for new analysts.

CONCLUSION

Most data science and visualization techniques for data analysis focus on tasks involving the extraction of meaning or understanding from data, but practical analysis environments and workflows involve complex sociotechnical systems. From an interview study with analysis experts spanning several areas of the intelligence analysis and security communities, our research establishes a greater foundation of knowledge for practical problems in analysis efforts. This paper highlights core issues involving the use of provenance in data analysis scenarios. Many issues are currently known by the research community and supported by existing tools and techniques, but significant challenges remain for many tasks. Furthermore, our study reveals opportunities for improvement across critical tasks having limited research and tool support. Advancements are needed to support common workflow needs, such as the presentation of analysis history, the training of analysis strategies, and quality assurance methods.

ACKNOWLEDGMENTS

The authors thank Tim West and Brian Veneklase for their help and expertise in formulating this research. This material is based on work supported by NSF 1565725 and the DARPA XAI program N66001-17-2-4032. This paper has supplementary downloadable material at http://ieeexplore.ieee.org, provided by the authors.

REFERENCES

1. *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. McLean, VA, USA: CIA Center for the Study of Intelligence, 2009.
2. A. H. Morris, "Human cognitive limitations. broad, consistent, clinical application of physiological principles will require decision support," *Ann. Amer. Thoracic Soc.*, vol. 15, no. Supplement 1, pp. S53–S56, 2018.
3. R. J. Heuer, *Psychology of Intelligence Analysis*. Morrisville, NC, USA: Lulu. com, 1999.
4. E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 31–40, Jan. 2016.
5. Y.-a. Kang and J. Stasko, "Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2011, pp. 21–30.
6. R. V. Badalamente and F. L. Greitzer, "Top ten needs for intelligence analysis tool development," in *Proc. Int. Conf. Intell. Anal.*, 2005, pp. 1–6.
7. G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 70–73, Jul./Aug. 2006.
8. R. Hoffman, N. R. Shadbolt, A. M. Burton, and G. Klein, "Eliciting knowledge from experts: A methodological analysis," *Org. Behav. Decis. Processes*, vol. 62, no. 2, pp. 129–158, 1995.
9. C. T. Silva, J. Freire, and S. P. Callahan, "Provenance for visualizations: Reproducibility and beyond," *Comput. Sci. Eng.*, vol. 9, no. 5, 2007, Art. no. 82.
10. C. Dunne, N. H. Riche, B. Lee, R. Metoyer, and G. Robertson, "GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 1663–1672.

11. P. H. Nguyen, K. Xu, A. Bardill, B. Salman, K. Herd, and B. W. Wong, "SenseMap: Supporting browser-based online sensemaking through analytic provenance," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2016, pp. 91–100.
12. A. Wexelblat and P. Maes, "Footprints: History-rich tools for information foraging," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1999, pp. 270–277.
13. A. Toniolo *et al.*, "Supporting reasoning with different types of evidence in intelligence analysis," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2015, pp. 781–789.
14. S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit, "From visual exploration to storytelling and back again," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 491–500, 2016.
15. E. D. Ragan, J. R. Goodall, and A. Tung, "Evaluating how level of detail of visual history affects process memory," in *Proc. ACM 33rd Annu. Conf. Human Factors Comput. Syst.*, 2015, pp. 2711–2720.
16. A. Sarvghad and M. Tory, "Exploiting analysis history to support collaborative data analysis," in *Proc. 41st Graph. Interface Conf.*, 2015, pp. 123–130. [Online]. Available: http://dl.acm.org/citation.cfm?id=2788890.2788913.
17. B. G. Glaser, A. L. Strauss, and E. Strutzel, "The discovery of grounded theory; strategies for qualitative research," *Nursing Res.*, vol. 17, no. 4, 1968, Art. no. 364.
18. P. Pirolli and S. Card, "Information foraging," *Psychol. Rev.*, vol. 106, no. 4, 1999, Art. no. 643.
19. S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst, "Futzing and moseying: Interviews with professional data analysts on exploration practices," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 22–31, Jan. 2019.
20. E. Kandogan, A. Balakrishnan, E. M. Haber, and J. S. Pierce, "From data to insight: Work practices of analysts in the enterprise," *IEEE Comput. Graph. Appl.*, vol. 34, no. 5, pp. 42–50, Sep./Oct. 2014.

Karthic Madanagopal is a Ph.D. candidate in the Department of Computer Science, Texas A&M University. He is also a senior programmer analyst with Knowledge Based Systems Inc. His research interests include human–computer interaction, visual analytics, and intelligence analysis. He is a member of the IEEE Computer Society. Contact him at karthic11@tamu.edu.

Eric D. Ragan is an Assistant Professor with the Department of Computer and Information Science and Engineering, University of Florida. He directs the Interactive Data and Immersive Environments (INDIE) lab, which conducts research of human–computer interaction, visual analytics, 3D interaction, and virtual reality. He received his Ph.D. in computer science from Virginia Tech. He is a member of the IEEE Computer Society. He is the corresponding author of this article. Contact him at eragan@ufl.edu.

Perakath Benjamin is a Vice President of R&D at Knowledge Based Systems, Inc. (KBSI), where he manages and directs the R&D activities. He received his Ph.D. in industrial engineering from Texas A&M University. Contact him at pbenjamin@kbsi.com.