The Influence of Visual Provenance Representations on Strategies in a Collaborative Hand-off Data Analysis Scenario

Jeremy E. Block, Shaghayegh Esmaeili, Eric D. Ragan, John R. Goodall, G. David Richardson

Abstract— Conducting data analysis tasks rarely occur in isolation. Especially in intelligence analysis scenarios where different experts contribute knowledge to a shared understanding, members must communicate how insights develop to establish common ground among collaborators. The use of provenance to communicate analytic sensemaking carries promise by describing the interactions and summarizing the steps taken to reach insights. Yet, no universal guidelines exist for communicating provenance in different settings. Our work focuses on the presentation of provenance information and the resulting conclusions reached and strategies used by new analysts. In an open-ended, 30-minute, textual exploration scenario, we qualitatively compare how adding different types of provenance information (specifically data coverage and interaction history) affects analysts' confidence in conclusions developed, propensity to repeat work, filtering of data, identification of relevant information, and typical investigation strategies. We see that data coverage (i.e., what was interacted with) provides provenance information without limiting individual investigation freedom. On the other hand, while interaction history (i.e., when something was interacted with) does not significantly encourage more mimicry, it does take more time to comfortably understand, as represented by less confident conclusions and less relevant information-gathering behaviors. Our results contribute empirical data towards understanding how provenance summarizations can influence analysis behaviors.

Index Terms—Analytic provenance, sensemaking, information transfer, visualization, workflow summarization, user studies

1 Introduction

Exploratory analysis involves the process of gathering information, identifying patterns, and investigating hypotheses. Due to the open-ended nature of exploratory analysis, there can be uncertainty in understanding the thought processes and the factors contributing to conclusions [5, 51]. With multiple ways of working through the data, individuals might arrive at different conclusions. Often this work happens in collaborative sessions or team environments [37, 75], so communicating how a discovery is reached is critical to all parties maintaining understanding with each other.

By assisting with tracking the analysis history, software can help facilitate collaboration and hand-offs of information across shifts. By tracking the *analytic provenance* [5, 44, 52] of an investigation, other analysts can later review the history to reveal what information was considered—or not considered—and how connections in the data led to the development of hypotheses or conclusions. While the potential value of provenance information is strong, core challenges remain with how to process provenance data and design effective representations to support easy human understanding. As has been well documented in the visualization community, the representation of information can have dramatic effects on human interpretation of data [16, 20, 64, 68]. Furthermore, there is relatively limited empirical knowledge of how provenance information is used in hand-off scenarios where a second analyst continues an analysis with provenance records from a prior analyst [75]. And while it is expected that awareness of a previous analyst's thought process would influence the analysis strategies for a new analyst, there is a need to better understand *how* the approach might be affected. Additionally, our research studies how different forms of provenance representation might influence continuing analysis behaviors in collaborative hand-offs.

We conducted a user study on how people use provenance information when continuing an analysis after a previous analyst's partial progress. The study is situated in the context of an intelligence analysis scenario with a text data set. To explore different analyst responses to provenance representation, the study compares two types of summarized provenance representations (based on interaction history and data coverage) along with a control condition without explicit provenance information. Through this study, we characterize patterns in participants' analysis findings, interaction behaviors, and approaches to using the provenance information in an information hand-off scenario. We contrast the effects of two provenance representations to show how confidence in one's conclusion relates to the level of detail in handed-off data representations, describe interaction metrics for investigation behaviors, and confirm analysis strategies identified in prior work.

- *Jeremy E. Block is with University of Florida. E-mail: j.block@ufl.edu.*
- *Shaghayegh Esmaeili is with University of Florida. E-mail: esmaeili@ufl.edu.*
- *Eric D. Ragan is with University of Florida. E-mail: eragan@ufl.edu.*
- *John R. Goodall is with Oak Ridge National Laboratory. Email: jgoodall@ornl.gov.*
- *G. David Richardson is with Oak Ridge National Laboratory. E-mail: richardsongd@ornl.gov.*

Manuscript received 31 March 2022; revised 1 July 2022; accepted 8 August 2022. Date of publication 26 September 2022; date of current version 2 December 2022. Digital Object Identifier no. 10.1109/TVCG.2022.3209495

2 Related Work

Here we describe the nature of sensemaking, the uses for provenance information to represent how users understand problems, and the various techniques used to evaluate and correct for bias in analyst processes.

2.1 Collaborative Sensemaking

From the early work of decision theory [10, 63], the goal was to understand how users arrived at a clear conclusion when working with ill-structured data. Researchers continued this work with the study of the *sensemaking* process, which covers the human tendency to oscillate between foraging for new information and schematizing how it fits with what one already knows [48]. While multiple definitions have evolved from this preliminary work, the general understanding is that sensemaking "involves the ongoing retrospective development of plausible images that rationalize what people are doing" [69]. Of interest to the work discussed in this paper is the *Data-Frame* model proposed by Klein et al. [34]. In their work, they explain *how* something is introduced can have a direct impact on how users frame and continue their analysis. Intelligence analysis [9, 18], medical diagnosis [23, 39], and even humble internet research [41, 42, 49] all share tasks where complex data requires thoughtful consideration and artifact synthesis to arrive at and present a formal conclusion [58]. So, while common analytic settings clearly employ parts of the sensemaking process, they also define operations involving collaborative teaming or hierarchical units [37]. Collaborators can distribute workload, but they need to establish a common ground to understand the investigation status and contribute toward the goal [53].

Even without the added complexity of coordination and knowledge sharing among collaborators, sensemaking tasks are often nonlinear by nature. Insights and connections may be discovered independent of a strict method or procedural scaffolding, making it more challenging to systematically describe how concepts build on each other or explain the overall relationships. Add to the combination that collaborating analysts are dealing with multiple layers of uncertainty and trust [55], and the situation becomes even more complicated.

Among the many purposes targeted for provenance support in visual analysis applications [50], aid for collaborative work is bolstered by helping people recall what they know, clearly communicate the steps taken, and making it possible to reproduce past work [58]. As demonstrated by Mathisen et al. [38] in their description of the *InsideInsights* tool, capturing a user's interactions as they worked and providing an interface for the addition of annotations, improved the collaboration and established common ground. This technique of enhancing interaction data with analyst-generated annotations is a common way to help maintain context for an analyst or different audience members [46]. Still, often the process of annotating (e.g., writing notes, tagging information) distracts from the gathering of information because it requires users to synthesize their fuzzy concepts into specific terms that may or may not communicate their exact meaning upon later review [58]. At the same time, the process of exploratory data analysis is inherently dynamic, leading to the requirement for plans to constantly change with the situation [61]. Without the transcription of accurate mental schemes, details can be forgotten, leading to false conclusions and inaccuracies [51, 52]. By partnering with computers, human analysts can focus less on annotation tasks, like recording how they arrived at different concepts, and shift their attention toward directing the analysis and hypothesizing relationships between discovered ideas [11, 19, 49]. In this study, we examine how the representation of a user's process impacts the sensemaking processes of new analysts.

2.2 Provenance and Visual Summarization

Analysis tools that capture analytic provenance information aim to improve the understandability [4], reproducibility [47], and transparency [14,28] of insight generation over time by providing the "story" of data exploration and interpretation [46]. Yet, there are a variety of provenance types that serve different purposes depending on the context [29, 50]. Provenance can aid in the recall of past work, the verification of others' work, the recovery of past actions, the review and optimization for future analysis, the presentation of new information, and ultimately the communication of insights between collaborators [50].

Visual provenance summarizations describe events that occur over time, and myriad approaches have been demonstrated for both algorithmic summarization and visual representation to ease human interpretation of the workflow, especially in collaborative settings. There is a need for accurate techniques that compress temporal events that occur while matching an appropriate level of temporal granularity and summarization to best serve different audiences. Some techniques focus on preserving the timing and order of events to allow for the review of specific analysis turning points (i.e., **History** representations) [17, 42, 75] while others provide a high-level summary of topics reviewed and remove elements of timing completely to make it easy to see what has been explored and what needs further analysis (i.e., **Coverage** representations) [20, 57, 68]. Provenance helps collaborators to maintain common ground as they work synchronously [14], or asynchronously [72, 75]. To assist in collaborator communication, many provenance visualizations focus on providing a reference to a user's interaction history or data actions using a timeline [75] or branching trees [17, 28, 41]. Others take a drastically different approach by removing the temporal aspect entirely and instead aim to describe *what* data was explored rather than *when* it was explored. By ignoring time in the representation, viewers can focus on the context of data coverage and patterns (e.g., [57, 68, 72, 74]). These techniques trade a higher level of summarization for less emphasis on analysis step replication and the specific actions taken to arrive at an analysis state.

The potential value of provenance summarization features in analysis tools is well justified by prior literature [28, 45, 50, 71, 72, 74]. However, it is less clear how differences in the way the provenance information is summarized and visualized can affect an analyst's process and decision-making. Numerous studies from the visualization community have demonstrated that differences in visual representation or the addition of new information can influence user bias. For example, Dimara et al. [16] have shown how allowing users to intentionally remove salient data (e.g., outliers) can lead to less bias and more rational decisions. Also, Wall et al. [68] have shown how providing users a summarization of what they have reviewed (i.e., an overview of their data coverage) can increase some user's awareness of unconscious biases, while potentially encouraging others to amplify their biases by intentionally focusing their analysis on specific areas or hypothesis too.

Considering the common aim of using provenance visualization to support collaboration among multiple analysts [14, 56], and the impact that visualization can have on user interpretations [13], there is a need to better understand the ways the availability and representation of provenance information from one analyst may influence another analyst's behaviors, biases, or conclusions. In addressing this gap, our work draws from lessons learned from existing empirical studies on visual design, provenance, and users' analysis choices to further understand how to best provide provenance information to users.

2.3 Empirical Studies of Visual Design Influences

Within the visualization community, there is a long history of evaluating the effects of using visual tools on user performance which is consistent with provenance representations. Of specific interest are the behavioral effects when making history information available. For example, Zhao et al. [75] focused on the various strategies participants used when examining the work of collaborators, suggesting that different strategies lead to different degrees of investigation completeness. Similar work has evaluated the types of strategies users employ when completing sensemaking tasks [32] and there are concerns that the strategies are influenced by the interface.

This becomes quite obvious when considering the types of interactions afforded to users in an interface. For example, tools like *Hindsight* clearly describe recall provenance through scented widgets to make it clear what parts of the data individuals have already explored in more detail [20]. By lowering the opacity for parts of the data recently examined, unexplored areas were, therefore, more prominent, and users were influenced to extend their examination. Xu et al. [72], in a more collaborative setting, similarly invited participants to review what data had been previously inspected and how it was visualized by prior analysts. By spatially distributing others' attempts in a "constellation," finding commonly used data and alternative areas of interest become more available to the user. This is a direct result of the interface and its influence on user interactions. Consistent with this line of work, multiple studies have shown how cognitive factors (like the anchoring effect [12]) can significantly impact users' insights and exploration in decision-making tasks [15].

It is clear that representations influence analyst behaviors [55]. One of the goals of analytic provenance research is to prevent the effects of belief perseverance [12], base rate biases [40], misuse of representative heuristics [1], and resolving conflicting insights [35]. As an example, visualization techniques have tried to prevent selection bias by showing an overview of how the data has been filtered, and data type comparisons to help users recognize when they may be examining a subset of data too closely or with too much emphasis [8]. Similarly, by showing the work already complete and suggesting ways for the analysis to continue, SOMflow also helped analysts complete a more thorough investigation [54]. These techniques rely on provenance information to help users recall what has been explored and potentially rectify their cognitive biases. Applicable to this study, Sarvghad and Tory [56], compared how coverage and timeline representations improved an analyst's accuracy, and the amount of data explored when analyzing structured numerical data. To extend their findings, we compare effects in textual data analysis and also inspect the strategies employed. Although there is evidence that provenance representation influences user strategies and performance, there are outstanding questions about how different provenance representations compare, especially in textual data analysis.

3 Experiment

To understand the influence of provenance representations on continuing open-ended investigations, we conducted a between-subject experiment. Participants were asked to complete an exploratory data analysis task started by a prior investigator with different representations of provenance information available. We describe the key factors of interest and experimental design in the next section.

3.1 Motivation and Study Design

Designs for summarizing analytic provenance can take a wide variety of forms. For the focus of our study, we consider two common classes of provenance summaries as generalizations of designs found in the research literature. Specifically, we distinguish provenance representations into either *interaction history* or *data coverage* summaries. Both designs bring uniquely different benefits to analysts in practice.

Provenance summaries using an **interaction history** approach tend to provide a timeline of events or describe which data are processed over time. This type of provenance helps identify critical moments of failure, reproduce results, or provide additional data transparency, but it takes more time to review because there is often more data to make sense of. Unfortunately, while computing systems are capable of capturing interaction events, this process can quickly bloom into large sequences that are challenging to summarize. Many types of provenance visualization tools do not distill meaning from raw interaction logs, choosing instead to visually represent all interactions or analysis stages in interactive tools to uncover patterns or flows (e.g., [14, 28]).

In contrast to provenance designs emphasizing the temporal flow of the analysis, we describe a separate generalized class of provenance summary as **data coverage** designs, which typically represent an overview of *what* data was explored instead of *when* it was explored. By compressing time, users can see what has been explored and what remains [56, 57, 68, 72, 74]. These techniques provide a higher level of summarization, focusing on providing a sense of context but lack enough detail to clarify or recreate past work [7].

While modern techniques still implement examples from both data coverage and interaction timelines [60, 66, 75], prior work has seen greater emphasis on using machine learning to extract patterns and assist in the summarization of time [24, 27, 59, 74] and less emphasis on comparative studies investigating the implications of different provenance summaries with people. Questions remain about how best to summarize interaction histories in digestible ways that help analysts by balancing content and cognitive load. Many works have shown that provenance representations influence the analytical behaviors of users [4, 13, 20, 25], yet they do not directly compare the effects of the prototypical provenance representations we discussed. A more direct comparison between interaction history and data coverage would be beneficial since both techniques summarize and present the past in a digestible way and future automation techniques would benefit from guidance on selecting the appropriate level of detail for a user's task.

Therefore, we designed a between-subject experiment to study how individuals work with different forms of provenance information while completing a textual data investigation. We address the following research questions to help direct our analysis:

1. RQ1: How does the inclusion of analysis history or data coverage influence the **conclusions** reached by a secondary analyst?
2. RQ2: How do secondary analyst **behaviors** differ when provided the analysis history or data coverage from a prior analyst?
3. RQ3: How does the inclusion of analysis history or data coverage influence the **types of strategies** a secondary analyst uses to solve the problem?

As a basis for the study, we used a browser-based, direct manipulation interface to display documents and record participant interactions. Since provenance representations are commonly used in collaborative data analysis scenarios, we simulated a hand-off scenario where users pick up and finish the analysis started by someone else. In an online study, participants were asked to review a set of documents and describe

Fig. 1. The textual analysis interface consisting of teal headers that open to reveal documents. Participants were given a summary from a prior analysis session. Some participants also received a representation of the prior analyst's provenance information as described in Figure 2.

any associations they were able to make. Provenance representations informed by the same prior analyst allowed for comparing analyst conclusions and strategies. Based on prior work with provenance, we hypothesized two main effects on behavior in collaborative hand-off cases. We expected history summaries to encourage the continuing analyst to engage in more verification of the prior analyst's progress due to the inclusion of a more complete record of the prior analyst's work (*Hypothesis H1*). And since data coverage summaries provide context about what has been explored at a glance, we expected users to quickly understand the prior analysis and explore other topics (*Hypothesis H2*).

With a combination of a think-aloud protocol [43], screen recordings, interaction logs, and semi-structured interviews, we sought to identify differences in participant conclusions reached and strategies used to better characterize the impact of provenance on collaborative analysis.

3.2 Visual Analysis Task and Tool

In collaborative data analysis tasks, users work together to uncover relationships and share results. Frequently, these analysis tasks require users to communicate ill-structured and potentially relevant information for future analysts to recognize and use. For our study of how a second analyst picks up after a prior analyst makes partial progress, we used a synthetic intelligence analysis scenario based on an existing publicly available dataset from the first micro challenge of the 2010 VAST Challenge data series [26]. This dataset, and others from the VAST Challenge, are commonly used as realistic proxies to simulate analysis tasks for visualization research (e.g. [56, 75]). The data consists of fictional phone transcripts, email correspondence, forum posts, newspaper articles, and other intelligence reports about fictional illegal arms traders. Of the 103 documents in the dataset, only 16 contain information relevant to the solution we asked participants to complete.

Within the tool, users could flexibly move and collapse documents as they explored, similar to other analysis workspace tools [2, 33, 52]. Specifically, users were tasked with determining if illegal arms traders were responsible for the spread of a mysterious pandemic. Participants could right-click to access a context menu. From this menu, they could trigger a search event for the term under their cursor or type out their own query in a text field. Searches highlighted the set of document title bars that contained exact string matches to the queried text. Critical to this experiment was identifying when and what kinds of information was revealed to users. The tool would actively log user events (e.g., document opens, mouse enters and exits, and searches conducted) as they explored the dataset. Each logged interaction was recorded with the time since the session began, the event type, the element's identifier, and other relevant information (position on screen, content of search, etc.). All participants were given the same set of documents, instructions, a "Summary for Supervisor" field, and a note from the prior analyst (see figure 1). The instructions further reiterated the scenario introduced to the participants and the "Summary for Supervisor" was a blank note where participants would type out their conclusions at the end of the session. We describe the provenance representations and their interactions in the next section.

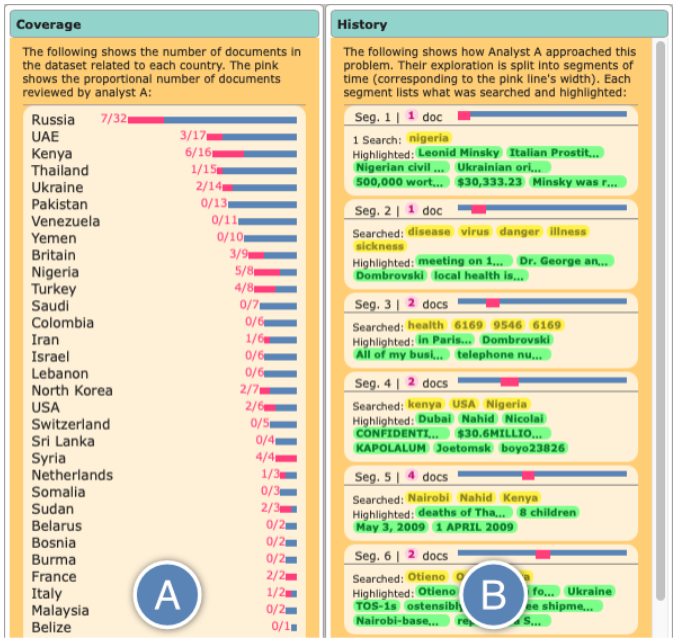

Fig. 2. Participants belonged to one of the three conditions: Control (no view), Control + the Coverage panel (A), or Control + History panel (B). When clicked, the countries (A) or segments (B) would help filter the documents related to the associated data.

3.3 Conditions

As described in Section 2, we define two general categories for provenance representations based on how they communicate time. To further explore how data coverage and interaction histories impact investigation conclusions, behaviors, and strategies, we examined prototypical representations of each. Below, we describe the different provenance representations used in the experiment:

To serve as a **Control**, all participants were given a textual description designed to mimic the series of conclusions or cursory annotations a prior analyst may string together as they completed the same analysis scenario (i.e., the "Notes from Analyst A"). This note served as an easy way to provide context and offer potential entry points for a participant's data exploration. Inspired by similar methodologies [56, 75], the note was based on the interaction log of a researcher's pantomimed analysis that followed specific details through documents to arrive at a partially correct conclusion. Several statements were intentionally hedged to provide ample openness for where participants could begin. In the pantomimed interactions, 15 documents were opened, along with 31 searches and 53 highlights. Generally, the referenced documents and approach would uncover the majority of details required to solve the whole solution (i.e., 6/16 documents). These participants were not given additional panels to filter the dataset and had to rely on the search tool described earlier to find information and solve the scenario. We used this same interaction history to construct the other provenance representations discussed below.

Some participants were additionally given a **Coverage** representation (Figure 2 A) of the data explored by the prior analyst. This view showed the countries that received the most attention from the prior analyst as a list with miniature bars. Each country's bar displayed the ratio of documents explored by the prior analyst and the bars were arranged in descending order by the country's mention frequency. The original dataset [26] did not have labels for the countries mentioned in a document, so these had to be hand-labeled by the researchers. Intending to simulate how a hypothetical tool would work, the researchers attributed the first city mentioned in a document to its corresponding country. These hand-labeled countries were also added to the preamble of each document, to help balance the conditions. This way the other conditions could also filter specific countries with the search tool. The Coverage panel made these country-focused searches conveniently accessible by displaying them as a list. By clicking a country from the list, a "tool-use" event would be logged, and the affiliated documents would be revealed—distinguishing those explored by the prior analyst and those not reviewed—by coloring the documents' title bar. Clicking a selected country again removed document coloring. Participants could select a country to filter the documents and help direct their investigation.

Other participants were provided a **History** representation (Figure 2 B) that summarized the steps the prior analyst took to complete their task. This view gathered and displayed the searches and highlights of analyst A as different segments of the analysis. The interactions from the baseline log were augmented with a handful of additional highlight terms or relevant search terms to provide a bit more content for the users of the History panel. Participants could scan through the History panel to help get a feel for the general terms (searches) and evidence (highlights) the prior analyst cared about over time. While manually segmented, the segmentation was done systematically. Segments were delineated based on a search event and contained the corresponding highlight events that took place between searches. In total, there were 11 segments, each labeled with the number of documents reviewed as well as a small timeline in the right corner of each segment to visualize its duration and placement in the prior analysis. Much like the Coverage representation, when a participant clicks a segment, the corresponding documents from that segment would be revealed by changing the color of affiliated document titles. Clicking a search or highlight term would run the interface's search command and simulate the results the prior analyst would have seen. Clicking a segment again removed the applied colors. Participants could select segments to filter the documents and help review aspects of the prior analysis.

3.4 Procedure

This research was approved by the organization's institutional review board (IRB). Participants joined a virtual meeting room and were asked to complete a demographic questionnaire capturing age, gender, academic program, and self-report measures for the ability to complete analysis tasks and their communication abilities. The experimenter then explained the web application interface and its functions to participants using a set of slides via screen-share. Within the tutorial slides, participants were introduced to the think-aloud protocol and then asked to demonstrate the technique with a short, irrelevant document. The researcher offered feedback and ways to help improve their think-aloud (e.g., they were asked to read aloud and verbalize their plans for what they would do next). Also, it is known that first impressions can have a large degree of influence over what people focus on [62], so all participants were read the same starting scenario (see section 3.2) to control for variations in wording. Participants were invited to ask questions about interactions throughout the tutorial, and a researcher was present with them in the interface to answer interface function questions as they worked. Participants were not told to explicitly "strategize," but rather to "choose what to read with intention" because "they would not have enough time to read everything."

Starting a screen recording, participants were given 30 minutes to complete the task and asked to think aloud [21] as they read and contemplated what associations they saw with the expectation that reading aloud would encourage deeper reflection [6]. At 10, 20, 25, and at the end of 30 minutes, participants were warned about the time elapsed and reminded of their task: "Try to identify what associations may exist and prepare your summary for your supervisor." After the analysis concluded, a post-task interview (included in the session recording) helped capture their mental model and opinions of the tools they used. With their final analysis workspace still visible, Participants were asked a series of semi-structured interview questions to further specify their understanding. These questions asked participants about the relationships they were aware of, identify retroactive strategies they used to arrive at their conclusions, their thoughts about the prior analyst's work, and the provenance views as applicable. Critically, they were asked to make a judgment on the relationship of arms dealing with the Nigerian disease (referred to as their conclusion).

3.5 Participants

We recruited 41 undergraduate university students from an upper-level computer science course as participants in the study. Participants were compensated with course credits. Data from five participants were excluded from analysis due to technical problems, communication uncertainty because of language barriers, or misunderstanding of task instructions. Of the remaining 36 participants, they were initially randomly distributed, before researchers assigned later participants to the smallest groups to finish with experimental groups of equal size (12 participants per condition). Fourteen participants (38%) identified as female, and 1 (3%) identified as non-binary/third-gender. All but 6 students (83%) were completing a degree in computer science, computer engineering, or software engineering.

The majority of participants were between 18 and 24 years of age (77%), 5 participants (14%) were within the 25–34 age category and 3 participants (8%) were older than 35. We asked participants to self-report their ability to complete data analysis on a scale from 1–10; an average reported score of 3.14 signified general inexperience in completing analysis tasks, indicating the participant sample might be considered analogous to novice analysts with no or limited experience.

4 Results

In this section, we present the results and insights drawn from our quantitative and qualitative data analysis. How people find and interact with the data can directly impact how they arrive at their conclusions and what conclusions they can make. To understand the ways people use provenance representations, we looked at participants' analysis behaviors who pick up from a prior analyst's progress. We studied behaviors captured through video recordings, think-aloud comments, interaction logs, and post-study interview responses.

4.1 User Findings and Confidence

We studied whether the availability of the provenance views affected the participants' findings and final conclusions (RQ1) to look for implications of early bias influencing analysts' ability to correct their preconceived expectations. A single author scored each participant's written conclusions using a four-level rubric, according to four factors, including the **accuracy** of their reported findings; the **recognition of errors** made in the prior analysis; the **number of findings and amount of detail** provided; and the **depth of relationships or connections** among entities and events in the data.

These qualities were a set of features we expected differences in based on provenance representations, yet the scores varied greatly. Due especially to the open-ended nature of user-directed analysis, the diversity of user aptitude with analysis and the various ways participants could write their findings, our analysis did not find systematic differences in participants' written conclusions caused by the provenance conditions. For example, when summarizing key findings, some participants formatted a formal report in prose, while others opted for a series of evidentiary bullet points. Often the written conclusion left out parts of the analysis and therefore was not a fair representation of the area's a participant explored. For a similar reason, we do not report on participant encounters with the 16 solution-relevant documents, because we wanted to understand what information "stuck" and was reported in their concluding thoughts. Due to the range of individual differences among participants and personal styles for reporting, the quality and completeness of written conclusions varied greatly. We did not find meaningful differences in analyst written conclusions. Therefore, we prioritized the analysis of data from the personalized post-study interview questions as a basis for identifying differences in participant conclusions and analysis behavior.

As part of this analysis, we assessed participants' confidence in the post-task interview when describing their final answer. High confidence implies that the user has convinced themselves of a specific relationship, and we want to see if that behavior varied systematically with the experimental conditions. One author coded participants' interview responses to the question regarding their final answer about the possible existence of a relationship in the data (i.e., the main investigation goal in the analysis scenario). With the support from a second author to review

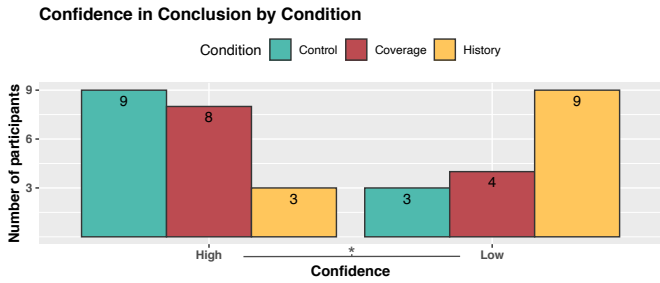

Fig. 3. The distribution of participants' confidence in their conclusions after their analysis. While the condition appears to have an influence on the confidence of a participant, post hoc comparisons cannot determine which conditions lead to different confidences.

the handful of edge cases, we separated participant responses into two categories (**high** and **low confidence**) based on the number of hedge statements in their verbal conclusion. To do this, in the interview, we asked participants if they identified a relationship between the concepts they were investigating. Participants who stated clearly that there was or was not a relationship were designated as *high* confidence (55%). Alternatively, those who suggested only a potential relationship, a need for additional time to review the data, or uttered more than two hedged phrases (e.g., "there might be...," "I guess...," "I think...,", etc.), were placed in the *low* confidence category (45%).

Upon first inspection, Figure 3 shows a strong difference in participant confidence. We see a main effect with Fisher's exact test ($p < 0.05$) by condition. Yet, post hoc comparisons with Bonferroni correction fail to identify the specific differences between conditions. While we cannot state one condition varied significantly from the others, those with a Coverage representation tend to exhibit conclusions more similar to the Control than those with the History representation. This is to say that the majority of conclusion qualities are not as dramatically impacted by the representation of provenance information. We do see some impact on participants' confidence when they verbally explain their conclusion, but without significant interaction effects, we instead turn our attention to the behaviors exhibited by participants to better explain how their analysis differed.

4.2 Quantitative Indicators of Behaviors

The way individuals interact with an interface indicates how they make sense of the data. With the variety of user activities, we choose to use a quantitative indicator to more clearly describe behavior patterns in participant approaches (RQ2). As captured in their interaction logs, we turn to the various analysis events and actions participants ran in the interface as a quantitative proxy for their understanding and to tease out the influence of provenance representation on user analysis behaviors. We look at three key representations of interest: the degree of *similarity*, and *difference* to the prior analyst's review, and the *rate of filtering* for specific information.

Of concern to the inclusion of provenance information is the influence on how much repetition is in subsequent analyses, or the amount of similarity to the prior analyst. While verification can be beneficial when auditing the veracity of a result, in most cases repeated work is not encouraged. We wanted to compare the analysis of each participant to the referenced prior analysis to understand if the provided provenance representation influenced how they addressed the problem of "continuing the analysis." If participants chose to look into the same concepts as the prior analyst, how similar was their review (*overlap*), and if they looked at different things, how unique was their investigation (*independence*)? To quantify the amount of overlap and independence in participants' behaviors we looked at which documents in a participant's interaction history were opened. Since critical information in the underlying dataset is separated into individual documents, we can determine how similar an investigation was to another by examining the set of documents opened. Both provenance representations were based on the same set of 15 documents, and we

calculated two separate, but similar ratios (overlap and independence) from the sets of documents participants reviewed. We calculated a participant's **overlap ratio** by considering the intersection between the set of documents a participant reviewed that was also reviewed by the prior analyst and divided by the number of documents the prior analyst reviewed ($= \frac{\{User\} \cap \{Analyst\}}{\{Analyst\}}$). An overview ratio of 1.0 implies that a participant saw all of the documents that analyst A reviewed.

The **independence ratio** captures the proportion of documents a participant reviewed that were different from the set reviewed by the prior analyst ($= \frac{\{User\} - \{Analyst\}}{\{User\}}$). An independence ratio of 1.0 implies that a participant only reviewed documents that the prior analyst did not. Although these ratios examine similar participant behavior properties, they are not inverse because they compare different document sets.

When we compare the degree of overlap among participants (Figure 4a), we do not see a strong difference between the conditions, implying that the amount of overlap a participant may have with a prior analysis has more to do with individual differences in approach and investigation intentions. But, some expected trends are visible. For example, those in the Control condition straddle the center (∼50%) as though they were unaware of which documents were reviewed by the prior analyst. We also see more spread in the Coverage condition, likely because they could filter documents and choose to follow or avoid the prior analysis. Finally, the History condition has the highest median overlap ratio likely since their representation emphasized how the prior analyst worked through documents.

We see a much different behavior among participants' ratios when we examine how independent their analysis was from the prior analyst (Figure 4b). A Kruskal-Wallis test revealed a significant difference in independence ratios $H(2) = 7.01, p < 0.05$. For those in the Control condition, more than half of the documents they reviewed had not been explored by the prior analyst. The result can be explained since they had no idea which documents were specifically reviewed by the prior analyst, and opened many more documents on average. With more documents opened, the likelihood of a document belonging to the set from the prior analyst goes down, leading to a higher independence ratio because there were many more documents that were not reviewed by the analyst. We also see a long tail for those in the History condition. The pairwise post-hoc Dunn test with Bonferroni adjustments showed to be only significant for the History and the Control ($p < 0.05$) suggesting that they generally spent their session reviewing mostly the same documents as the prior analyst. They also looked at fewer documents overall. With fewer documents reviewed, and an emphasis on documents opened by the prior analyst, these participants were more likely to maintain lower Independence ratios. Overall, we see that while the degree of overlap is not significant, there appear to be some differences in the diversity of documents participants are exposed to when given various provenance representations. This is to say that the affordances provided by different tools can influence which information participants review.

Another aspect of our analysis relates to how participants direct their investigation. We looked at the total number of interactions and divided it by the length of a participant's session to determine an average rate of interaction. The interaction rate (see Figure 4c) can serve as a proxy for the level of control a participant has over the interface and their investigation. Participants with exceptionally low interaction rates may be taking a long time to review documents or working very methodologically, whereas exceptionally high interaction rates may imply they have opened numerous documents at once or are not spending enough time understanding each document. For example, the Control condition opened the most documents (41.5 documents) on median, while maintaining the slowest median interaction rate (0.95 interactions/sec). This contradictory phenomenon may suggest that those in the Control condition opened many more documents on median in an attempt to understand the data, but worked through the documents and the task slowly. Yet, uncovering more descriptive interpretations would require the review of additional metrics.

For example, we can turn to the frequency of filtering events to understand how participants search and reduce the data space (See Figure 4d). While the differences are not significant, there are more median filtering events completed (18.0) by those in the History condition while they also look at the fewest documents (28.5) on average. This is to suggest that those in the History condition were more likely to be methodological and consistent since they were not exposed to as many documents and spent more of their session filtering and refining their investigation criteria. On the other hand, those in the Coverage condition appear to have overall investigation behaviors most similar to the Control, but also have the fastest interaction rate. Those with the Coverage representation may have allowed participants to work faster and independently.

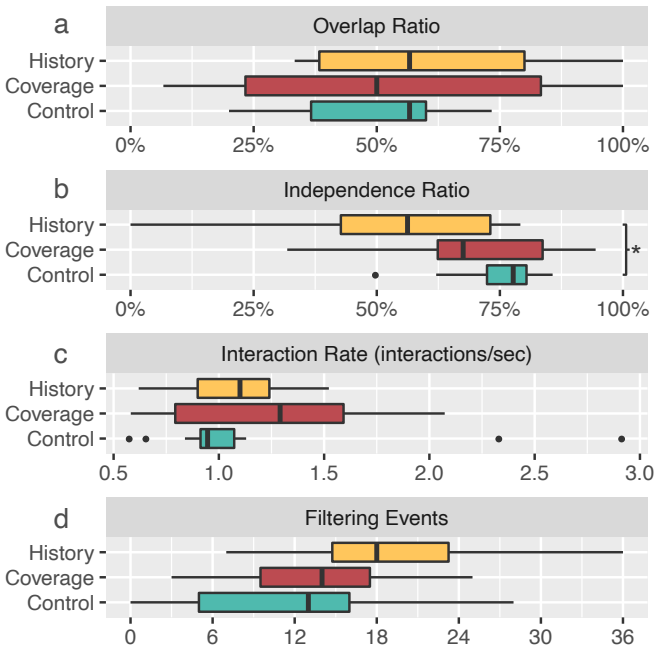

Fig. 4. A user's interaction history is used to calculate the following. Both *Overlap Ratio* and *Independence Ratio* refer to the set of documents each participant opened and compares them to the prior analyst (see Section 4.2). *Filtering Events* refers to interactions that help users find documents and is the combination of clicks in a provenance representation combined with search events. *Interaction rate* is calculated as the ratio of total interactions over the length of analysis (i.e., interactions/sec).

To further describe the ways participants were exposed to information, we examined the timing of users' searches and provenance representation usage. In this case, we define the term **filtering** to describe the sum of searches and provenance panel usage per participant. Although those in the Control condition did not have access to a provenance representation, they relied more heavily on search to find information. This is to say that search and panel usage both helped to reduce the search space in the dataset and their combination provides a more universal comparison. We examine when these behaviors occur to understand when users are selecting and directing their investigations. Since each participant filtered the data their own number of times, we calculate a filter percentage over time. All but 2 participants eventually reached 100% of their filtering events within their 30-minute session, but ultimately, we draw attention to the rate of change. In Figure 5, there is lots of overlap in when participants are completing their filtering behaviors. To clarify the trends, we apply a 4-factor polynomial since it gave the highest r^2 coefficient (0.745) and helps characterize the behaviors observed. As evidenced in the modeled regression, participants from the History condition complete about 50% of their filtering events by about 12 minutes, while it takes 17 or almost 19 minutes for the Coverage and Control conditions to conduct half of their filtering events respectfully. We counted the number of participants who had completed at least half of their filtering events by the halfway point in

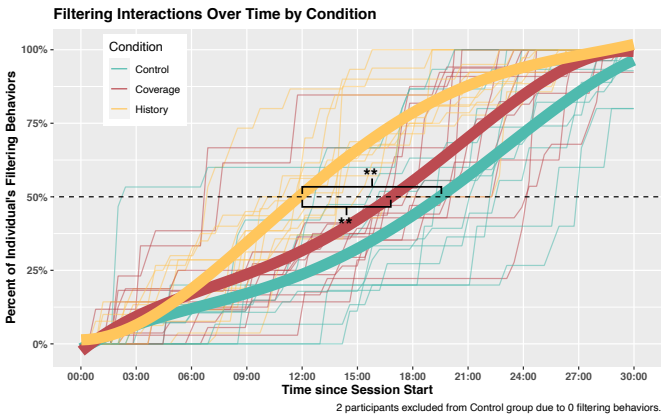

Fig. 5. Trends in percent of a filtering interactions completed over time. The thicker lines represent a four-factor polynomial of trends by condition. Although the History condition takes some time to get started, they conduct more filtering earlier before tapering off later, whereas the Coverage and Control conditions used more filtering later in time and some continued to filter right up to the end.

the investigation (15 minutes) as a proxy for how active participants were selecting data to review. Because the provenance representations provided a more convenient (i.e., a single click) way to filter the data, we see that both the History and Coverage condition completed the majority of their filter interactions before those in the Control condition. A Fisher's exact test confirmed that there were differences in the rate of filtering among the different conditions ($p < 0.001$). Further post hoc pairwise Fisher's exact tests with Bonferroni corrections confirm that the amount of filtering completed by those in the History condition differed from the Control ($p = 0.001$) and Coverage conditions ($p < 0.001$), but no difference was detected between the Coverage and Control conditions at 15 minutes. More specifically, those in the History condition complete the majority of their filtering behaviors before the other conditions (and in the first half of the session).

We believe this difference is due to the interaction pattern required for the use of the History representation. Since information about which documents were reviewed is only accessible after clicking a segment, participants from the History condition would commonly click through multiple segments at once in pursuit of a subset of documents to review instead of intentionally selecting a segment to review or independently typing out their own search query (discussed more in Section 6). This pattern of clicking through each segment in time, or scanning the resulting documents to find something previously reviewed likely led to an increase in filtering events overall (median 18.0 events) and also a higher frequency of events earlier in a session while participants were still gathering information. On the other hand, it appears that the addition of Coverage information for participants did not significantly change when participants would be selecting/filtering data (as compared to the Control). Ultimately, we see some quantitative differences in participant interactions, including how little content those in the History condition were able to review, and how frequently they were filtering the dataset. We now take these quantitative differences and construct some qualitative definitions for user strategies.

4.3 Qualitative Analysis of Strategies

The strategies employed by participants during an open-ended evaluation likely depends on the kinds of information they are presented with (RQ3). Comparing and analyzing the strategies users take when approaching their exploration can shed light on how provenance information is used when investigating the relationships of various data.

In alignment with the work on situated planning [61], of the participants who verbalized a preparatory plan, most were vague and often only consisted of 2–3 steps. These early plans were interesting to us as we wanted to see if the availability of provenance information would influence how plans were made and adapted (see RQ3). With analysis

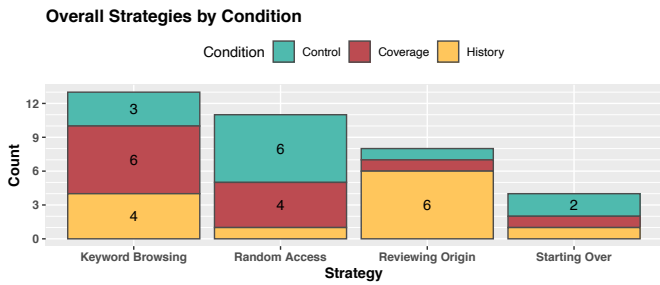

Fig. 6. General strategies used by participants when continuing analysis.

plans being constantly adjusted and renegotiated as new information is acquired, we ultimately simplified our analysis by focusing on initial plans and the actions observed in the first 10 minutes of activity.

From participants' think-aloud and retrospective interviews, we tried to reconstruct participants' intentions as they began their analysis. To do this, one author analyzed the data by conducting two rounds of open coding to establish a set of emergent features from user actions and analysis foci. In a similar data analysis task, Zhao et al. [75] defined a set of analysis **strategies** derived from their own qualitative coding. Their work compared the strategies participants used when constructing knowledge graphs with or without an interactive state timeline. Our work differs in that we compare different provenance representations, while they only evaluate differences with or without provenance features. While their study offered participants different data analysis affordances, we borrow similar qualitative analysis steps in our work as well. From their identified strategy definitions, we cross-referenced our most common tags and refined a set of similar analysis strategies. In contrast, ours are more focused on the amount of similarity to the original analyst's investigation as well as our users' commitment to their investigation plan. We clarify and define our categories below.

- **Keyword Browsing** - These participants set a plan for how they wanted to understand the data before they began or shortly after reviewing the note provided by the analysts. They made an intentional plan, maintained attention to the planned areas of interest, and frequently hypothesized different relationships. They had a medium amount of overlap with the prior analyst and tended to have a higher independence score since they were trying to extend the analysis and were more likely to have a more active role in the investigation.
- **Random Access** - These participants had less structure to their investigations—often working without stating how they wanted to systematically approach the problem—and spending the majority of their time gathering information instead of synthesizing hypotheses. These participants bounced around the dataset with about equal amounts of overlap and independence from the prior analyst.
- **Reviewing Origin** - These participants expressed a plan to verify the work from the prior analyst. Often this was motivated by a lack of trust in the conclusions made in the analyst's summary or started verifying the prior analyst's work and ran out of time for their own investigation. Therefore these participants have noticeably lower independence scores and higher amounts of overlap.
- **Starting Over** - These participants explicitly stated that they wanted to work independent of any influence from the prior analyst. They were worried explicitly about bias or interested in comparing their own understanding with the work by the prior analyst later. These participants intentionally closed the guidance from the prior analyst and waited till at least 15 minutes into the task before they reviewed the analyst's summary or the provenance panels.

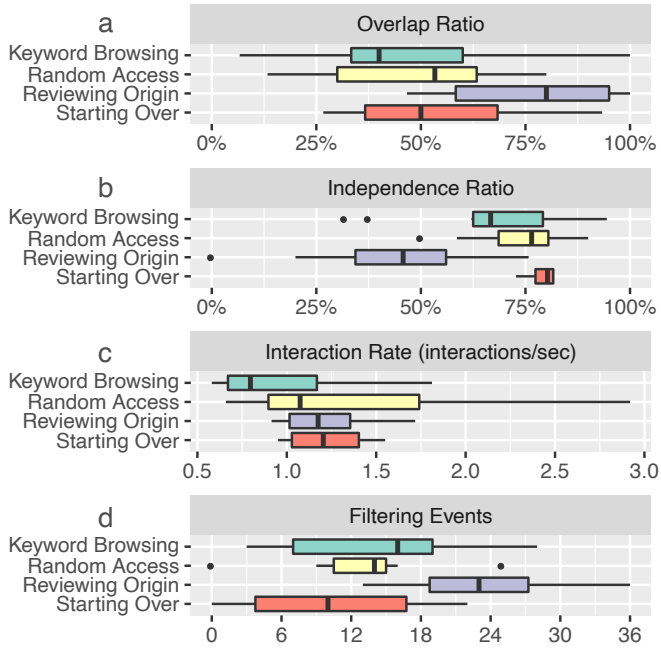

Fig. 7. There are some interaction patterns among the participants in various strategy types. Due to the limited participants in the *starting over* (n=4), we do not compute statistical differences and instead focus on visual analysis and descriptive statistics. For details on the factors described, see Figure 4.

As reported in Figure 7, there are some interesting differences among users and their strategies. With only a handful of participants in the *starting over* group (n=4), we rely on descriptive statistics and visual analysis to describe the groups and their differences instead of standard statistical methods. To begin, all groups appear to open the same number of documents, except for those in the *starting over* group (45.5 document open events). Those in the *starting over* group also had the highest median interaction rate (1.20 interactions per second). Much of this is likely due to the small number of participants in the group, but the way most of these participants worked was often without a stated plan and typically focused on gathering information generally. Similar behaviors were seen in the *random access* group. Without a plan, these participants also had high median interaction rates (1.07 interactions per second) and the highest interquartile range (0.90–1.74 interactions per second). This can be explained by the lack of intention and purposeful activity these participants appeared to execute during the analysis and reduced specificity of what they were looking for. Frequently, these participants had bursts of activity where they would be gathering information slowly before they would find a recognizable piece of information and open multiple documents they already reviewed to find interrelationships and similarities. Without a strong plan set at the beginning and generally less filtering to help find the information of interest, these participants were most likely to require more review of the data.

The other two categories had more explicit intentions for what they wanted to review. Those in the *reviewing origin* group completed the most filtering events on median (23.0 events). Likely due to the high ratio of participants from the History condition, this characteristically high number of filtering events is a direct result of using the History panel to review various segments of time; users had to trigger a filtering event each time they wanted to examine the documents reviewed in each segment of the prior analysis. These participants also maintain a high degree of overlap with the prior analyst (80%) and the lowest median independence (45%). On the other hand, those in the *keyword browsing* group also had a stated plan and had a much different degree of overlap with the prior analyst (only 40%). Most of these participants intended to explore key aspects of the data described by the prior analyst and tended to open more documents the prior analyst had not touched. We noticed that those who used a keyword browsing strategy generally uncovered a more complete and accurate picture of the dataset. We believe this is because they independently verified aspects of interest and followed keywords instead of the work by the prior analyst.

Peculiarly, while there is not a significant difference, the distribution of strategies almost appear to follow with the three conditions(see Figure 6), where *keyword browsing* was mostly associated with those in the Coverage condition, *random access* was mostly members from the Control condition and those in the *reviewing origin* group were from the History condition. Yet, without strong evidence that provenance representations influence the type of strategy participants employ when planning for their analysis, we cannot conclude that these strategies are determined by the affordances in the interface.

5 Discussion

Our work studies how providing provenance summaries can influence future investigators who pick up a partial analysis from a prior analyst. We see evidence that: (1) listing interaction histories can result in a type of secondary sensemaking task for trying to understand the earlier analysis, (2) individual differences introduce a large amount of variation in how people choose to utilize provenance summaries, and (3) the types of strategies exhibited by the continuing analysts show some similarities with analytic strategies predicted by prior work.

Considering the possible strategies participants adopted, we expected those who reviewed the prior analyst's progress would first start at the beginning of the prior analyst's work and then select relevant information to take into their own analysis and establish their own understanding. Yet, the observed results show the opposite trend. For example, the majority of participants in the *reviewing origin* group were also from the *History* condition. Interestingly, our interviews and observations of participants in the *History* condition found they often felt overwhelmed with the extra available information, and we saw the majority of the *History* condition had low confidence in their conclusions (see Figure 3). This corresponds with the frequent triggering and review interaction pattern required to see the documents reviewed by the prior analyst. This is also evidenced by the group's more frequent filtering behavior prior to 15 minutes, which may be due to the amount of information they were tasked to review in the limited time. We find supporting evidence for $H1$, as it appears to take more time to construct an accurate mental model because participants were exploring the events in the data but also understanding how the prior analyst approached the problem. The broader implication is that having History information may exaggerate the task's difficulty by making more content accessible for review and not summarizing enough to serve users hoping to pick up a prior analysis. On the other hand, providing Coverage information to users does not feel exceptionally beneficial beyond what was provided by the Control. While a greater proportion of participants with a Coverage representation may maintain more *keyword browsing* strategies, there are no significant deviations from the control, thus rejecting $H2$.

In open-ended analysis tasks, there are no clear paths that lead to a solution. Among the various metrics collected in our study to characterize analysis, we see the breadth of approaches through the large degree of variation and spread in the data. While some significant differences among conditions emerge (i.e., for the degree of investigation independence), many cross-condition differences may be hidden by the high variability in individual preferences and the way participants adapted their approaches in situ. Though the experiment provided the same instructions for the analysis tasks for all participants, we clearly found four unique types of approaches emerge. While a trend toward a specific strategy appears to align with the conditions (e.g., in Figure 6 we see 6 participants from each condition associated with different strategies), due to lack of significance we cannot conclude a direct relationship between provenance usage and strategy employed. A participant's choice to verify the prior work (i.e., *reviewing origin*) or follow their own set of keywords (i.e., *keyword browsing*) is likely a result of their personal experience completing analysis tasks or interest

in referencing provenance information as well as other factors, and not due to how provenance information is provided. Expanded knowledge on the topic may benefit from future work that considers users' preconceptions and other factors that influence how users develop and enact their strategies.

Finally, among the set of strategies we see, there is evidence that the techniques are in alignment with earlier work. We found participants' initial strategies were markedly similar to the groupings found by Zhao et al. [75]. In a similar data hand-off task with an interaction history-like provenance representation, they identified five typical strategies. Their strategies "random access," "tracing from origin," and "starting over" closely align with our *random access*, *reviewing origin*, and *starting over* categories, respectively. The key difference is that we have combined their "naive browsing," and "hubs and bridges" categories into one group of *keyword browsing*. While they described how some participants adjusted and shifted their strategies, we did not see as many transitions in our shorter 30-minute analysis session, likely due to the limited time participants had to complete the task. Yet, the definitions they used to describe the various interaction strategies and techniques were in alignment with the set of strategies we observed our participants employ. Since their set of categories was also based on traditional, non-collaborative sensemaking strategies, our work further reinforces the idea that hand-off strategies are similar to other data analysis and sensemaking techniques in other settings [32]. Drawing on work for creative collaboration, one way we may learn more from these analysis tasks would be to observe strategies employed in more creative scenarios and the user interactions longitudinally [22]. We also see further evidence in support of the work by Sacha et al. [55]. They identified how skeptical users will only begin verifying another's work if they see anomalies or have hypotheses about where mistakes were made. Those who used a *Reviewing the Origin* strategy were often intrigued by some aspect of the prior analysis and sought to resolve these needs for evidence by reviewing many of the same documents already reviewed. This led to higher ratios of overlap and less independence from the prior analysis. With similarly identified strategies to prior work [75], our findings further reinforce that a handful of common strategies exist for different analysis tasks that seem to be based on a user's situated approach.

5.1 Limitations

Our study of how users pick up an analysis with provenance information from prior collaborators was based on a single analysis scenario and targeted participants with limited data analysis experience. To build further knowledge on the topic, following research is needed with additional data sets and with participants with varying backgrounds from different collaborative data analysis communities like intelligence operators, medical teams, and academic researchers. Studies could also consider how differences in participants' abilities, problem-solving aptitude, or preferences for particular analytic strategies might influence different approaches or interaction patterns.

A challenge with working with open-ended sensemaking tasks is that a user's conclusions may not fit within certain bounds of available conclusions and captured data metrics. We see this in capturing participant conclusions. We set out to capture participant accuracy, error correction, findings made, amount of detail, and depth of relationships but did not have the fidelity in the task complexity nor measures to capture meaningful differences. In future work, more thoughtful care and constraints should be taken to more formally compare these factors and the influence of provenance.

In this work, the researchers describe the tool and scenario to participants. Due to the natural variation in speech, there are potential confounds introduced based on different inflections being interpreted by different users. While all participants were read the same statements, future work could better control for these effects by using a prerecorded procedure or expanding to additional datasets with different contexts to help generalize findings.

While our work introduced the comparison of two prototypical techniques used in the literature, there are also more ways of representing past work and their influence on user performance ought to be explored. for example, some work focuses on generating textual summaries [31,70], while others show branching timelines [8,17,65,67] that communicate how analysis adapts, transforms, and evolves. Still, others use graphs and networks as a way of providing concept maps [36,73,75], while others design comics as a technique for summarizing segments of time [3, 30]. Questions remain not only in the representation provenance should take, but also at what level of detail the provenance should be maintained. The comparison among how these techniques influence user performance as well as variations in the level of detail ought to be explored in the future.

6 Conclusion

In this paper, we examine the effects of provenance representations on future investigator behaviors. In an open-ended textual data analysis task users were given different kinds of providence information visualizations and asked to pick up the analysis. We examine the downstream effects of two prototypical provenance representations for collaborative sensemaking (i.e., Coverage and History). Like the findings of Kang et al. [32], while provenance representations do not appear to have significant impacts on user strategy, it does suggest that there are benefits of providing a succinct representation of provenance to help users pick up where other users left off. While both representations take users time to situate their understanding, we see evidence that the History representation takes more time to comprehend. Because it does not simplify the prior analysis as well as the Coverage representation, the History representation appears to introduce an additional sensemaking task for participants. This appears to be especially true when the purpose of provenance is to collaboratively communicate [50]. It would be interesting to see how these prototypical representations behave when different analysis tasks are evaluated. Our results imply that provenance summaries that reduce the complexity of the analysis will be beneficial in hand-off analysis scenarios. To prevent the overwhelm associated with information overload, perhaps dynamic, user-defined levels of detail in provenance would shed additional light on design guidelines for the analytic provenance community. These results contribute to the refinement of design guidelines for provenance representations and further emphasize the need for provenance summary techniques.

Acknowledgments

The authors wish to thank the participants for their involvement in this study. This work was supported in part by the DARPA Perceptually-enabled Task Guidance (PTG) Program under contract number HR00112220005. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] M. AlKhars, N. Evangelopoulos, R. Pavur, and S. Kulkarni. Cognitive biases resulting from the representativeness heuristic in operations management: an experimental investigation. *Psychology Research and Behavior Management*, 12:263–276, Apr 2019. doi: 10.2147/PRBM.S193092

[2] C. Andrews, A. Endert, and C. North. Space to think: Large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 55–64. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753336

[3] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, and N. Henry Riche. Design patterns for data comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173612

[4] P. Bao. *Sharing Insight Provenance in Collaborative Visual Analytics*. PhD thesis, USA, 2013. AAI3563689.

[5] L. Battle and J. Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum*, 38(3):145–159, 2019. doi: 10.1111/cgf.13678

[6] J. E. Block and E. D. Ragan. Micro-entries: Encouraging deeper evaluation of mental models over time for interactive data systems. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV)*, p. 38–47, Oct 2020. doi: 10.1109/BELIV51497.2020.00012

[7] M. A. Borkin, C. S. Yeh, M. Boyd, P. Macko, K. Z. Gajos, M. Seltzer, and H. Pfister. Evaluation of filesystem provenance visualization tools. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2476–2485, Dec 2013. doi: 10.1109/TVCG.2013.155

[8] D. Borland, W. Wang, J. Zhang, J. Shrestha, and D. Gotz. Selection bias tracking and detailed subset comparison for high-dimensional data. *IEEE Trans. Vis. Comput. Graphics*, 26(1):429–439, Jan 2020. doi: 10.1109/TVCG.2019.2934209

[9] L. Bradel, C. North, L. House, and S. Leman. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 163–172, Oct. 2014. doi: 10.1109/VAST.2014.7042492

[10] G. Bruckmaier, S. Krauss, K. Binder, S. Hilbert, and M. Brunner. Tversky and kahneman's cognitive illusions: Who can solve them, and why? *Frontiers in Psychology*, 12:584689, Apr 2021. doi: 10.3389/fpsyg.2021.584689

[11] N. Chinchor and W. A. Pike. The Science of Anal. Reporting. *Information Visualization*, 8(4):286–293, Jan. 2009. doi: 10.1057/ivs.2009.21

[12] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The Anchoring Effect in Decision-Making with Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 116–126, Oct. 2017. doi: 10.1109/VAST.2017.8585665

[13] I. Cho, R. Wesslen, S. Volkova, W. Ribarsky, and W. Dou. Crystalball: A visual analytic system for future event discovery and analysis from social media data. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, p. 25–35, Oct 2017. doi: 10.1109/VAST.2017.8585658

[14] H. Chung, S. Yang, N. Massjouni, C. Andrews, R. Kanna, and C. North. VizCept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 107–114, Oct. 2010. doi: 10.1109/VAST.2010.5652932

[15] M. B. Cook and H. S. Smallman. Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, 50(5):745–754, Oct 2008. doi: 10.1518/001872008X354183

[16] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri. Mitigating the attraction effect with visualizations. *IEEE Trans. Vis. Comput. Graphics*, 25(1):850–860, 2019. doi: 10.1109/TVCG.2018.2865233

[17] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson. GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pp. 1663–1672. Association for Computing Machinery, New York, NY, USA, May 2012. doi: 10.1145/2207676.2208293

[18] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proc. of the SIGCHI Conference on Human Factors in Comput. Systems*, CHI '12, pp. 473–482. Association for Comput. Machinery, New York, NY, USA, May 2012. doi: 10.1145/2207676.2207741

[19] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, 43(3):411–435, Dec. 2014. doi: 10.1007/s10844-014-0304-9

[20] M. Feng, C. Deng, E. M. Peck, and L. Harrison. HindSight: Encouraging Exploration through Direct Encoding of Personal Interaction History. *IEEE Trans. Vis. Comput. Graphics*, 23(1):351–360, Jan. 2017. doi: 10.1109/TVCG.2016.2599058

[21] M. E. Fonteyn, B. Kuipers, and S. J. Grobe. A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, 3(4):430–441, Nov. 1993. doi: 10.1177/104973239300300403

[22] J. A. Gonzales, C. Fiesler, and A. Bruckman. Towards an appropriable cscw tool ecology: Lessons from the greatest international scavenger hunt the world has ever seen. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, p. 946–957. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2675133.2675240

[23] C. E. Gonzalez, N. Brito-Dellan, S. R. Banala, D. Rubio, M. Ait Aiss, T. W. Rice, K. Chen, D. C. Bodurka, and C. P. Escalante. Handoff Tool Enabling Standardized Transitions Between the Emergency Department and the Hospitalist Inpatient Service at a Major Cancer Center. *American Journal of Medical Quality*, 33(6):629–636, Nov. 2018. doi: 10.1177/1062860618776096

[24] R. Gove. Automatic narrative summarization for visualizing cyber security logs and incident reports. *IEEE Trans. Vis. Comput. Graphics*, 28(1):1182–1190, jan 2022. doi: 10.1109/TVCG.2021.3114843

[25] N. Goyal and S. R. Fussell. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, p. 288–302. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2818048.2820071

[26] G. Grinstein, S. Konecni, C. Plaisant, J. Scholtz, and M. Whiting. Vast 2010 challenge: arms dealings and pandemics. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 263–264. IEEE, 2010. doi: 10.1109/VAST.2010.5649054

[27] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE Trans. Vis. Comput. Graphics*, 22(1):51–60, Jan 2016. doi: 10.1109/TVCG.2015.2467613

[28] J. Heer, J. D. Mackinlay, C. Stolte, and M. Agrawala. Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1189–1196, Nov. 2008. doi: 10.1109/TVCG.2008.137

[29] M. Herschel, R. Diestelkämper, and H. Ben Lahmar. A survey on provenance: What for? What form? What from? *The VLDB Journal*, 26(6):881–906, Dec. 2017. doi: 10.1007/s00778-017-0486-1

[30] M. K. Hong, U. Lakshmi, T. A. Olson, and L. Wilcox. Visual odls: Co-designing patient-generated observations of daily living to support data-driven conversations in pediatric care. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1—–13. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174050

[31] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, p. 3363–3372. Association for Computing Machinery, May 2011. doi: 10.1145/1978942.1979444

[32] Y.-a. Kang, C. Gorg, and J. Stasko. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, p. 139–146. IEEE, 2009. doi: 10.1109/VAST.2009.5333878

[33] P. E. Keel. Collaborative visual analytics: Inferring from the spatial organization and collaborative use of information. In *2006 IEEE Symposium On Visual Analytics Science And Technology*, p. 137–144, Oct 2006. doi: 10.1109/VAST.2006.261415

[34] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso. A data-frame theory of sensemaking. In *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, pp. 113–155. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, Jan. 2007.

[35] J. K. Li, S. Xu, Y. C. Ye, and K.-L. Ma. Resolving Conflicting Insights in Asynchronous Collaborative Visual Analysis. *Computer Graphics Forum*, 39(3):497–509, 2020. doi: 10.1111/cgf.13997

[36] J. Lung and S. Easterbrook. Inflo: Collaborative reasoning via open calculation graphs. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, p. 1199–1202. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2145204.2145384

[37] K. Madanagopal, E. D. Ragan, and P. Benjamin. Analytic Provenance in Practice: The Role of Provenance in Real-World Visualization and Data Analysis Environments. *IEEE Computer Graphics and Applications*, 39(6):30–45, Nov. 2019. doi: 10.1109/MCG.2019.2933419

[38] A. Mathisen, T. Horak, C. N. Klokmose, K. Grønbæk, and N. Elmqvist. Insideinsights: Integrating data-driven reporting in collaborative visual analytics. *Computer Graphics Forum*, 38(3):649–661, Jun 2019. doi: 10.1111/cgf.13717

[39] L. Murray, D. Gopinath, M. Agrawal, S. Horng, D. Sontag, and D. R. Karger. Medknowts: unified documentation and information retrieval

for electronic health records. *arXiv:2109.11451 [cs]*, Sep 2021. arXiv: 2109.11451. doi: 10.1145/3472749.3474814

[40] A. Narechania, A. Coscia, E. Wall, and A. Endert. Lumos: Increasing awareness of analytic behavior during visual data analysis. *IEEE Trans. Vis. Comput. Graphics*, p. 1–1, 2021. doi: 10.1109/TVCG.2021.3114827

[41] 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 *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 91–100, Oct. 2016. doi: 10.1109/VAST.2016.7883515

[42] P. H. Nguyen, K. Xu, A. Wheat, B. W. Wong, S. Attfield, and B. Fields. SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Trans. Vis. Comput. Graphics*, 22(1):41–50, Jan. 2016. doi: 10.1109/TVCG.2015.2467611

[43] J. Nielsen, T. Clemmensen, and C. Yssing. Getting access to what goes on in people's heads? reflections on the think-aloud technique. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, NordiCHI '02, p. 101–110. Association for Computing Machinery, New York, NY, USA, 2002. doi: 10.1145/572020.572033

[44] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink. Analytic provenance: process+interaction+insight. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, p. 33–36. Association for Computing Machinery, May 2011. doi: 10.1145/1979742.1979570

[45] W. Oliveira, D. D. Oliveira, and V. Braganholo. Provenance Analytics for Workflow-Based Computational Experiments: A Survey. *ACM Computing Surveys*, 51(3):53:1–53:25, May 2018. doi: 10.1145/3184900

[46] D. Park, M. Suhail, M. Zheng, C. Dunne, E. Ragan, and N. Elmqvist. StoryFacets: A design study on storytelling with visualizations for collaborative data analysis. *Information Visualization*, Aug. 2021. doi: 10.1177/14738716211032653

[47] T. Pasquier, M. K. Lau, A. Trisovic, E. R. Boose, B. Couturier, M. Crosas, A. M. Ellison, V. Gibson, C. R. Jones, and M. Seltzer. If these data could talk. *Scientific Data*, 4(1):170114, Sept. 2017. doi: 10.1038/sdata.2017.114

[48] P. Pirolli and S. Card. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *International Conference on Intelligence Analysis*, Jan. 2005.

[49] N. Rachatasumrit, G. Ramos, J. Suh, R. Ng, and C. Meek. ForSense: Accelerating Online Research Through Sensemaking Integration and Machine Research Support. In *26th International Conference on Intelligent User Interfaces*, pp. 608–618. ACM, College Station TX USA, Apr. 2021. doi: 10.1145/3397481.3450649

[50] 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. Graphics*, 22(1):31–40, Jan 2016. doi: 10.1109/TVCG.2015.2467551

[51] E. D. Ragan and J. R. Goodall. Evaluation methodology for comparing memory and communication of analytic processes in visual analytics. In *Proceedings of the Fifth Workshop on Beyond Time and Errors Novel Evaluation Methods for Visualization - BELIV '14*, p. 27–34. ACM Press, 2014. doi: 10.1145/2669557.2669563

[52] E. D. Ragan, J. R. Goodall, and A. Tung. Evaluating how level of detail of visual history affects process memory. In *Proc. of the 33rd Ann. ACM Conf. on Human Factors in Comput. Syst.*, CHI '15, p. 2711–2720. Association for Computing Machinery, Apr 2015. doi: 10.1145/2702123.2702376

[53] A. C. Robinson. Collaborative synthesis of visual analytic results. In *2008 IEEE Symposium on Visual Analytics Science and Technology*, p. 67–74, Oct 2008. doi: 10.1109/VAST.2008.4677358

[54] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, and D. A. Keim. Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE Trans. Vis. Comput. Graphics*, 24(1):120–130, Jan 2018. doi: 10.1109/TVCG.2017.2744805

[55] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Trans. Vis. Comput. Graphics*, 22(1):240–249, Jan. 2016. doi: 10.1109/TVCG.2015.2467591

[56] A. Sarvghad and M. Tory. Exploiting analysis history to support collaborative data analysis. In *Proceedings of the 41st Graphics Interface Conference*, GI '15, p. 123–130. Canadian Information Processing Society, CAN, June 2015. doi: 10.5555/2788890.2788913

[57] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing Dimension Coverage to Support Exploratory Analysis. *IEEE Trans. Vis. Comput. Graphics*, 23(1):21–30, Jan. 2017. doi: 10.1109/TVCG.2016.2598466

[58] N. Sharma and G. Furnas. Artifact usefulness and usage in sensemaking handoffs. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–19, 2009. doi: 10.1002/meet.2009.1450460219

[59] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Trans. Vis. Comput. Graphics*, 27(2):453–463, Feb 2021. doi: 10.1109/TVCG.2020.3030403

[60] H. Stitz, S. Gratzl, H. Piringer, T. Zichner, and M. Streit. Knowledgepearls: Provenance-based visualization retrieval. *IEEE Trans. Vis. Comput. Graphics*, 25(1):120–130, Jan 2019. doi: 10.1109/TVCG.2018.2865024

[61] L. A. Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Nov 1987.

[62] P. E. Tetlock. Accountability and the perseverance of first impressions. *Social psychology quarterly*, pp. 285–292, 1983. doi: 10.2307/3033716

[63] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124

[64] A. C. Valdez, M. Ziefle, and M. Sedlmair. Priming and anchoring effects in visualization. *IEEE Trans. Vis. Comput. Graphics*, 24(1):584–594, 2018. doi: 10.1109/TVCG.2017.2744138

[65] A. Walch, M. Schwärzler, C. Luksch, E. Eisemann, and T. Gschwandtner. Lightguider: Guiding interactive lighting design using suggestions, provenance, and quality visualization. *IEEE Trans. Vis. Comput. Graphics*, 26(1):569–578, 2020. doi: 10.1109/TVCG.2019.2934658

[66] C. Walchshofer, A. Hinterreiter, K. Xu, H. Stitz, and M. Streit. Provectories: Embedding-based analysis of interaction provenance data. *IEEE Trans. Vis. Comput. Graphics*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3135697

[67] R. Walker, A. Slingsby, J. Dykes, K. Xu, J. Wood, P. H. Nguyen, D. Stephens, B. W. Wong, and Y. Zheng. An extensible framework for provenance in human terrain visual analytics. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2139–2148, 2013. doi: 10.1109/TVCG.2013.132

[68] E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE Trans. Vis. Comput. Graphics*, 28(1):966–975, Jan 2022. doi: 10.1109/TVCG.2021.3114862

[69] K. E. Weick, K. M. Sutcliffe, and D. Obstfeld. Organizing and the Process of Sensemaking. *Organization Science*, 16(4):409–421, Aug. 2005. doi: 10.1287/orsc.1050.0133

[70] W. Willett, S. Ginosar, A. Steinitz, B. Hartmann, and M. Agrawala. Identifying redundancy and exposing provenance in crowdsourced data analysis. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2198–2206, Dec 2013. doi: 10.1109/TVCG.2013.164

[71] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch. Survey on the analysis of user interactions and visualization provenance. *Computer Graphics Forum*, 39(3):757–783, Jun 2020. doi: 10.1111/cgf.14035

[72] S. Xu, C. Bryan, J. K. Li, J. Zhao, and K.-L. Ma. Chart constellations: Effective chart summarization for collaborative and multi-user analyses. *Computer Graphics Forum*, 37(3):75–86, 2018. doi: 10.1111/cgf.13402

[73] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2080–2089, 2013. doi: 10.1109/TVCG.2013.167

[74] J. Zhao, M. Fan, and M. Feng. Chartseer: Interactive steering exploratory visual analysis with machine intelligence. *IEEE Trans. Vis. Comput. Graphics*, 28(3):1500–1513, Mar 2022. doi: 10.1109/TVCG.2020.3018724

[75] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan. Supporting Handoff in Asynchronous Collaborative Sensemaking Using Knowledge-Transfer Graphs. *IEEE Trans. Vis. Comput. Graphics*, 24(1):340–350, Jan. 2018. doi: 10.1109/TVCG.2017.2745279