reVISit: Looking Under the Hood of Interactive Visualization Studies

Carolina Nobre
cnobre@g.harvard.edu
Harvard University

Dylan Wootton
dwootton@mit.edu
Microsoft

Zach Cutler
zcutler@sci.utah.edu
University of Utah

Lane Harrison
ltharrison@wpi.edu
Worcester Polytechnic Institute

Hanspeter Pfister
pfister@g.harvard.edu
Harvard University

Alexander Lex
alex@sci.utah.edu
University of Utah

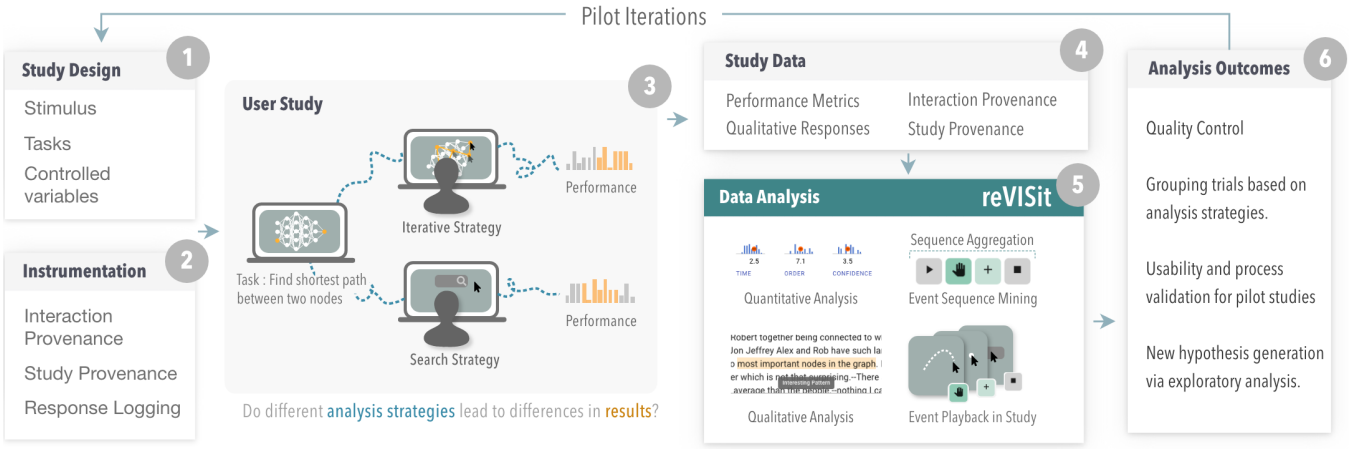

Figure 1: Analysis pipeline for instrumenting empirical user studies with provenance data for flexible analysis of the collected data. During study design (1), the visualizations, tasks, and any controlled variables are defined. The visualization technique is (2) instrumented using provenance and response logging. The (3) user study produces the (4) study data, which is then analyzed in (5) reVISit using a suite of data analysis methods. (6) The outcomes of the analysis process include quality control, detecting participant analysis strategies, and generating new hypotheses. The process can be also be used to refine pilot studies.

ABSTRACT

Quantifying user performance with metrics such as time and accuracy does not show the whole picture when researchers evaluate complex, interactive visualization tools. In such systems, performance is often influenced by different analysis strategies that statistical analysis methods cannot account for. To remedy this lack of nuance, we propose a novel analysis methodology for evaluating complex interactive visualizations at scale. We implement our analysis methods in reVISit, which enables analysts to explore participant interaction performance metrics and responses in the context of users' analysis strategies. Replays of participant sessions can aid in identifying usability problems during pilot studies and make individual analysis processes salient. To demonstrate the applicability of reVISit to visualization studies, we analyze participant data from two published crowdsourced studies. Our findings show that reVISit can be used to reveal and describe novel interaction patterns, to analyze performance differences between different analysis strategies, and to validate or challenge design decisions.

CHI '21, May 08-13, 2021, Yokohama, Japan

CCS CONCEPTS

• **Human-centered computing** → **Visualization design and evaluation methods**.

KEYWORDS

Visualization, evaluation methodology, user studies, provenance, event sequences.

ACM Reference Format:
Carolina Nobre, Dylan Wootton, Zach Cutler, Lane Harrison, Hanspeter Pfister, and Alexander Lex. 2021. reVISit: Looking Under the Hood of Interactive Visualization Studies. In *Proceedings of CHI '21: ACM Conference on Human Factors in Computing Systems (CHI '21).* ACM, New York, NY, USA, 13 pages.

1 INTRODUCTION

The diversity of methods for evaluating visualization techniques have expanded significantly over the last decade, but controlled empirical studies remain an important tool. Crowdsourcing, in particular, has made it possible to efficiently collect large amounts of data from empirical experiments [4]. Studies that evaluate user performance are either geared toward understanding human perceptual and cognitive limits or focused on evaluating an interactive

system or technique [25]. Although crowdsourced studies have been widely applied to perceptual experiments and simple visualizations, we recently proposed methods for using crowdsourced empirical studies for sophisticated interactive visualizations [29]. Evaluation of such complex systems, however, comes with new challenges, including the need to train participants on using these visualizations as well as controlling for the effect of different interaction affordances on participant performance.

Accuracy and time are standard metrics used to quantify user performance in empirical studies, but metrics can also be more diverse, including insights [15] and open-ended responses [25]. Capturing time and accuracy as performance metrics is a well-established practice, but numerous visualization researchers and cognitive scientists have highlighted problems with relying solely on these two metrics [17, 18]. For example, speed and accuracy often lack the required precision to measure cognitive effort, and different analysis strategies can produce vastly different accuracy and speed responses [30]. This problem is compounded when evaluating complex visualizations, since different levels of participant expertise, diverging but equally valid analysis strategies, and familiarity with the technique can influence task performance. It is therefore essential in empirical studies to capture and evaluate more comprehensively how participants interact with complex visualizations [29] and what effect their different analysis strategies have on task performance. Understanding different analysis strategies and how they affect performance requires both data collection and a human-in-the-loop approach, as user analysis strategies are hard to objectively quantify. For example, knowledge about the goals of the study and the properties of the interface being evaluated can play a crucial role in interpreting participant actions.

The first contribution of this paper is **a novel analysis methodology for evaluating complex interactive visualizations at scale**. This methodology complements existing empirical study approaches in that it provides a set of guidelines along all phases of study design and is applicable to both quantitative and qualitative large-scale studies. The proposed workflow covers steps from the initial study design to outcomes of the data analysis stage (Figure 1). A pillar of this methodology is fully instrumenting interactive visualizations using provenance tracking [8]. We distinguish provenance tracking [32] from logging since provenance data makes it possible to fully reconstruct an application's state, and hence retrace and analyze all of a participant's actions. The provenance and response data captured during the study is then used for data analysis. Actionable outcomes of this analysis include characterizing user analysis strategies, process and design validation of visualizations during pilot studies, and hypothesis generation for downstream statistical analysis. However, analyzing the results of provenance-tracked user studies is complex because it relies on two orthogonal data streams — interaction provenance and user responses to study prompts — which, when combined, support a richer understanding of the types of analysis different visualization techniques support.

To address this challenge, we developed **the open-source reVISit system**[1], which enables analysts to capture and characterize different interaction approaches and their influence on task performance (Figure 1–5). ReVISit provides support for both exploratory and query-driven analysis of empirical study data. Overview-first summaries support open-ended exploration to highlight high-level patterns. Analysts can also take a bottom-up approach, querying for specific interactions, participants, or performance metrics. A 'playback' feature recreates a participant's analysis path within the study, allowing the analyst to see the sequence of actions that the participant executed. The reVISit workflow complements statistical analysis tools by allowing analysts to look more closely at *how* participants solve tasks, and externalize their findings with tags directly in the data. ReVISit then interfaces with statistical analysis tools by supporting the export of the study data along with analyst metadata such as tags and annotations.

We validate our approach and our tools in **two detailed case studies using previously published crowdsourced user studies [12, 29]**. Our findings show that reVISit can be used to reveal and describe different analysis approaches, to analyze performance differences between various strategies, and to validate or challenge design decisions in the interactive visualizations.

2 RELATED WORK

Current analysis workflows for collecting and analyzing data from empirical studies can vary depending on: (1) the types of data collected, (2) the goals of the study, and (3) the data analysis approaches used on the study output. Here, we discuss existing approaches to each of these categories, as outlined in Figure 2.

2.1 Data Collection

The data that is collected during empirical user studies is dictated by the intended study goal and ensuing analysis. Carpendale et al. [6] classify empirical studies into two broad camps: those that strive to capture 'realism', i.e., observational studies, and those that prioritize 'precision', or controlled studies.

Observational studies capture participant insights during the analysis process with think-aloud sessions and video recordings of participants. These studies are well suited for capturing visual analysis and reasoning processes [24, 26]. However, given the resources required to observe participants and extract insights from the captured qualitative data, this approach does not scale well to large numbers of participants. Conversely, controlled studies capture information in a controlled environment via tasks with measurable performance metrics. These studies prioritize precision over realism, and provide valuable empirical evidence on how the evaluated visualizations perform under different conditions [23]. Although they can be deployed to large groups via crowdsourcing, these approaches rarely capture *why* certain stimuli perform differently or *how* participants interact with the stimuli.

The reVISit workflow builds on existing approaches to data collection by integrating data types captured in observational studies to add a 'situated' component to controlled studies. To this end, we capture stateful interaction provenance (Figure 1–2), which allows analysts to play back participant actions as performed in the original study. This simulates a video playback of the participant's interactions, and gives analysts insitu context that cannot be captured by quantitative metrics alone. By applying event sequence mining algorithms to the captured provenance data, we make analysis of interaction sequences scalable to large studies.

[1] https://vdl.sci.utah.edu/reVISit/

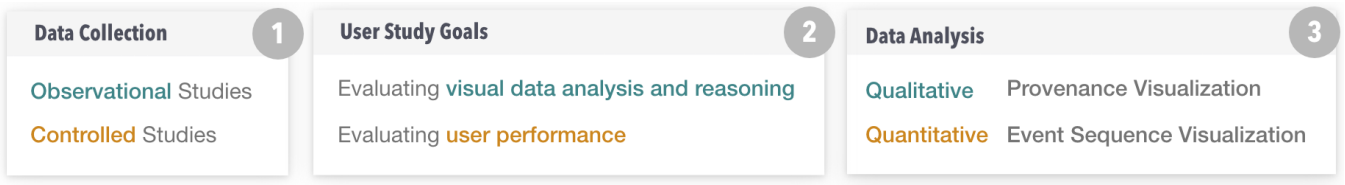

Figure 2: Existing approaches for different aspects of collecting and analyzing user study data on interactive visualizations. (1) Observational studies capture participant insights and reasoning. Controlled studies collect performance metrics for each task. (2) Visual data analysis and reasoning studies aim to capture the analysis process, whereas user performance studies analyze the study outcome. (3) Qualitative approaches derive themes and capture insights from user responses. Quantitative methods analyze measurable user performance metrics such as accuracy and time. Provenance and event sequence visualization methods support investigation of the analysis process. ReVISit and our workflow build on these existing approaches by *capturing multiple sources of data to support insights about how user analysis approaches influence performance metrics.*

2.2 User Study Goals

Lam et al. [25] organize the space of empirical studies in visualization according to the study goals and types of research questions. Of direct relevance to user studies that assess interactive visualization are those that evaluate visual data analysis and reasoning (VDAR) and those that capture user performance (UP).

The main goal of VDAR evaluations is to assess a visualization tool's ability to support visual analysis and reasoning about data. Study outputs include quantifiable metrics such as the number of insights obtained during analysis, as well as subjective feedback such as opinions on the quality of the data analysis experience. User performance studies are those that measure participant performance on a set of task in terms of metrics such as time and accuracy. Outputs in these studies are generally quantitative and are analyzed using descriptive statistics such as means, standard deviations, confidence intervals, and p-values.

In this work, we contribute a novel methodology that supports goals from both VDAR and UP user studies. Detailed provenance data supports inquiries on how participants perform visual analysis to solve each task. Additionally, qualitative participant responses shed light on the types of insights they achieve while solving the tasks throughout the study. Quantitative metrics captured during the study provide information on performance. One of the key strengths of the proposed methodology is that in capturing multiple sources of data that support both VDAR and UP goals, analysts can now investigate *how* data analysis informs user performance.

2.3 Data Analysis

To support the investigation of how data analysis patterns affect user performance, reVISit leverages data analysis approaches in the areas of provenance visualization, event sequence visualization, and general qualitative and quantitative methods. Provenance data in the context of data visualization refers to the history of interactions, visualization states, insights, or the reasoning an analyst traverses during analysis [14, 32]. Provenance visualization research investigates how this data can inform user engagement, insight generation, analysis replication, and general analysis strategies [35]. ReVISit builds on existing provenance visualization research by using provenance to support playback of participant analysis sessions directly in the original study stimulus. This playback gives analysts context for understanding interaction patterns that can greatly aid in assigning semantic meaning to different approaches.

Interaction provenance is often captured as a series of events. Event sequence exploration methods can be grouped into two main categories: exploration through overview or through pattern searching [5]. ReVISit supports both overview and query-centric methods to analyze the event sequences in the provenance data. Visual analytics systems focused on event sequences allow analysts to interactively explore and derive insights from large amounts of interaction provenance data [3, 13, 16, 34]. For example, Blascheck et al. [3] introduce VA^2, an evaluation approach for VA applications. The system is focused on gaining insights into how users perceive and interpret a new visualization approach through interaction logs and eye tracking data. To this end, they focus on identifying interaction patterns to derive common participant strategies. ReVISit expands on VA^2 and related efforts by also supporting linked analysis of interaction provenance with participant performances as captured by both quantitative and qualitative metrics.

Qualitative data captured in user studies can be analyzed using several existing methods [1, 27, 33]. Many of these methods are types of thematic analysis, where themes are extracted from the underlying data and then used to guide the coding of user responses [19]. Coding is the process of labeling raw data, and then using the collected codes to form a theory [7]. ReVISit supports the thematic analysis of qualitative study results, allowing analysts to create and store codes at several levels of granularity, from specific segments of text data, to individual participants, to entire groups of participants as defined by analyst-driven faceting of the data. These codes can then be exported with the final dataset and used in further analysis outside reVISit.

Although many empirical studies in visualization and HCI analyze quantitative results with null hypothesis significance testing (NHST), this practice has come under increased criticism [22]. Instead, researchers are advocating for *transparent statistics* [9], which, among other things, proposes straightforward graphical communication of the results, including a representation of the associated uncertainties [9, 22]. In reVISit, we follow these recommendations for rendering quantitative metrics, showing 95% bootstrapped confidence intervals as well as a histogram of the underlying distribution of data. Analysts can also export the data and any computed metrics for more in-depth statistical treatment outside reVISit. Ultimately,

reVISit builds on existing data analysis methods for empirical studies by leveraging the analysis of exploration and reasoning data to inform the interpretation of user performance metrics.

3 GOALS AND TASK ANALYSIS

We designed reVISit and our workflow to address the analysis needs of researchers conducting empirical studies to evaluate interactive visualizations at scale. As discussed in Section 2, analysis of user studies that investigate interactive systems focuses on either understanding the visual analysis process or user performance. ReVISit expands on the types of analysis tasks that are possible by also supporting inquiries on *how visual analysis strategies influence user performance.*

- **T1 Discover analysis strategies.** Although performance metrics can indicate whether participants are able to solve a task successfully, they do not reveal *how* users solve the task. Understanding the interaction sequences used by participants can either confirm or challenge visualization designers' assumptions about how a technique supports a given task. Additionally, more than one approach often exists to performing a task on interactive visualization. Understanding the different, and sometimes unexpected, ways in which participants solve tasks can provide valuable insight into how interactions affordances are being used. Investigating performance metrics on these different strategies can highlight the strengths and weaknesses of each approach.
- **T2 Disambiguate variations in performance.** Statistical analysis of performance metrics such as accuracy and time cannot discern between the different approaches used to achieve those metrics. However, an understanding of different analysis strategies can shed light on which approaches were more successful, as well as possible reasons for unsuccessful interactions.
- **T3 Validate or challenge visualization and interaction design decisions.** Designing interactive visualizations requires making decisions about which interactions and which visual encodings to provide and how to introduce them to novice users. Analysis of interaction patterns can help researchers investigate whether the design decisions were appropriate and led to both engagement from participants and successful task responses.
- **T4 Quality control of study data.** Running large studies has the benefit of statistical robustness, but with the drawback of making it challenging to control for the quality of the trials. Provenance data can be useful for quality control, as it, for example, can reveal participants who disengage with the tasks for longer periods of time, thereby skewing the concept of 'time to complete' for that task.
- **T5 Process validation for pilot studies.** Pilot studies are a common device to identify problems in the study process, find bugs and usability issues in the stimuli, and validate the assumptions made when designing a user study. Joint analysis of interaction patterns and results at the pilot phase can reveal whether participants understand the technique well enough to perform the tasks in a satisfactory way, or whether usability problems hinder them in completing a task. Additionally, analysis of pilot data can highlight whether the captured provenance appropriately supports the questions the analyst wishes to investigate with the deployed study data.
- **T6 Explore new hypothesis.** Although user studies are often designed to test a predefined set of hypothesis, freely exploring the resulting data can lead to unexpected findings. This is particularly true for more complex systems, where our assumptions on how users engage with unfamiliar tools may be inadequate.

To address these goals, we developed a workflow that outfits user studies with detailed provenance tracking and a set of analysis methods for mining insights from the resulting data. Next, we describe the workflow along with considerations of how analysts can implement it in their own empirical studies.

4 WORKFLOW

Figure 1 outlines the steps in the workflow, along with how information flows between steps. During the (1) *study design* phase , the stimulus (visualization techniques), tasks, and any controlled variables are defined. Although the initial study design is not influenced by the reVISit workflow when using this methodology to run and analyze pilot studies, the study design can be iteratively refined with the results of the analysis. The study is then (2) *instrumented* with detailed provenance and response logging, which includes tracking interaction provenance, study provenance, and all responses from participants. *Interaction provenance* captures the sequence of interactions a user performs with the visualization, often defined as a series of time-stamped actions. ReVISit supports several analysis tasks on interaction provenance data, including event sequence mining, filtering, and grouping sequences of interest. These operations allow analysts to abstract lower level interactions into higher level 'analysis approaches'.

Study provenance refers to larger scale events that capture a participant's progression through the study phases. Unlike interaction provenance, these events do not log interactions with the visualization, but capture the start and end time of semantically meaningful events within the context of the user study. These events include logging when participants browse away from the study, how long they spend on a training video, time spent on postsurvey questions, etc. Overall, study provenance can provide valuable information on how participants engage with the study that goes beyond their performance and interactions in individual tasks.

Study *responses* contain the two main types of data collected during the study itself: *controlled responses* that can then, for example, be used to compute accuracy, and *open-ended responses.* Qualitative responses can be particularly useful for capturing higher level participant goals and insights during exploration. ReVISit supports analysis of both types of result data with different affordances for each one, as discussed in Section 5.

Once the study has been designed and instrumented, it can be (3) deployed to participants . Given the automatic logging of user interactions and input, these studies can scale to large numbers of participants. The (4) data captured in the study contains computed

performance metrics, qualitative responses, and both interaction and study provenance. This data, together with the study parameters, feed into a suite of data analysis methods, which we implement in the (5) reVISit system. These methods support the dual analysis of provenance and results, providing insights into tasks such as how different analysis strategies influence study results. ReVISit includes support for analyzing results, provenance, and capturing analyst insights during the analysis process with qualitative tags.

The possible (6) outcomes of this workflow include performing quality control, grouping trials based on analysis strategies, and generating new hypotheses on the strengths and weaknesses of the visualization techniques being evaluated. Additionally, this workflow can provide valuable insights into usability and process validation for pilot studies. For example, study provenance can highlight participants who take too long on training and trial sections and are struggling to understand the visualization technique. If problems are detected, the workflow can be restarted, and adjustments to any of the phases can be made, by, for example, improving training or fixing usability problems. Detecting such problems at a pilot stage can greatly enhance the quality of the final study.

5 VISUALIZATION AND INTERACTION DESIGN

ReVISit supports the joint analysis of provenance and result data in an interactive web-based tool. Here we report on the design decisions that went into realizing this tool and how they support the tasks outlined in section 3. ReVISit has five views: (1) The *task overview* provides a summary of each task, including the stimulus given for each condition, the top ten interaction patterns, and a summary of qualitative and quantitative results (**T1**, **T2**, **T6**). (2) The *participant timeline* view contains an annotated timeline for each participant's progression through the study (**T4**, **T5**). (3) The *task analysis* view displays a table of participant provenance and performance data that allows for flexible sorting, faceting, and comparison of groups along any provenance or performance metric (**T2**, **T3**, **T4**, **T6**). (4) The *event manager* allows the analysts to group, hide, and create sequences of events to best suit their analysis needs (**T1**, **T6**). Finally, (5) the *playback* view enables analysts to replay interactions of selected participants directly in the original study stimulus (**T2**, **T5**).

5.1 Task Overview

The task overview is the starting point for an analyst using reVISit. It provides a summary for tasks and conditions in the study. The faceting order and rules (conditions first or tasks first) can be customized. This view includes both *study design* aspects such as stimulus and task prompt and summaries of *study provenance* and *study results*. The view and interactions in the task overview are designed to support top-down analysis tasks, such as *Which tasks had greater variations in analysis strategies?* (**T1**) or *Which tasks showed the biggest difference in performance between conditions?* (**T6**).

Figure 3 shows the task overview for a task in one of the crowdsourced studies we explore in Section 8. From left to right, the summary includes *study design*, *study provenance*, and *study results*. The study design information includes the task prompt, answer, and stimulus, giving context to the study provenance and results. Clicking on any of the study stimuli takes the analyst directly to the live study, allowing for more detailed exploration of the interaction affordances for that task.

Inquiries about the main analysis strategies used by participants (**T1**) are supported by the interaction provenance summary (Figure 3–B), which displays the ten most common event sequences for the task in that condition. Each interaction is represented by a glyph that contains either an icon or a two-character abbreviation for that action. The top ten patterns are computed using the PrefixSpan algorithm [20]. We also tested and considered several algorithms in the SPMF data mining library [11], but found that PrefixSpan performed best for the scale and level of iterations supported by reVISit. Whereas the output of different event sequence mining algorithms has variations, our design for displaying these summaries is agnostic to the choice of algorithm. The participant count column uses unit visualizations [31] where each participant is represented as a small glyph. If a participant performed a sequence, they are shown in dark gray. We chose unit visualizations since they are an intuitive encoding for showing intersection of sets. Analysts can inspect the degree of overlap between participants with certain analysis strategies by hovering over a row (**T1**), which highlights in orange common participants between the hovered row and each other row in the table. Clicking on a row highlights participants who do *not* intersect with the clicked row in blue. Hovering over a row updates the results section to show distribution plots and qualitative results for only those participants (**T2**). Analysts can inspect the effect of interactions sequence length as well as how many participants used each sequence by sorting on the respective columns in the table.

The study results section of the task overview provides a summary of both quantitative and qualitative results (**T6**). Quantitative results are presented as histograms, with means and bootstrapped 95% confidence intervals superimposed. We explored non-aggregated representations such as beeswarm plots in order to show the raw data, but those approaches did not scale well for large studies with several quantitative metrics for each participant. Summarizing qualitative data is not as straightforward since extracting meaning from this type of data often involves input from a human coder, which we support in the *task analysis* view in reVISit. For the task overview, we show frequent words (excluding stopwords) so that analysts can get an initial idea about the content. The 20 most frequent words are displayed as a bar chart below the quantitative results. These study result summaries allow analysts to detect high-level patterns that can prompt a more detailed inspection for particular tasks. Analysts can sort all tasks in reVISit based on any metric and condition.

One of the key strengths of reVISit is the integrated analysis of provenance and performance (**T2**). In the task overview, this analysis is supported by linked highlighting between the interaction provenance and results views. Showing these relationships supports preliminary hypothesis generation as to possible effects of analysis strategies on participant performance.

5.2 Event Manager

The interactions captured with provenance can vary greatly in both level of abstraction and semantic meaning — from individual

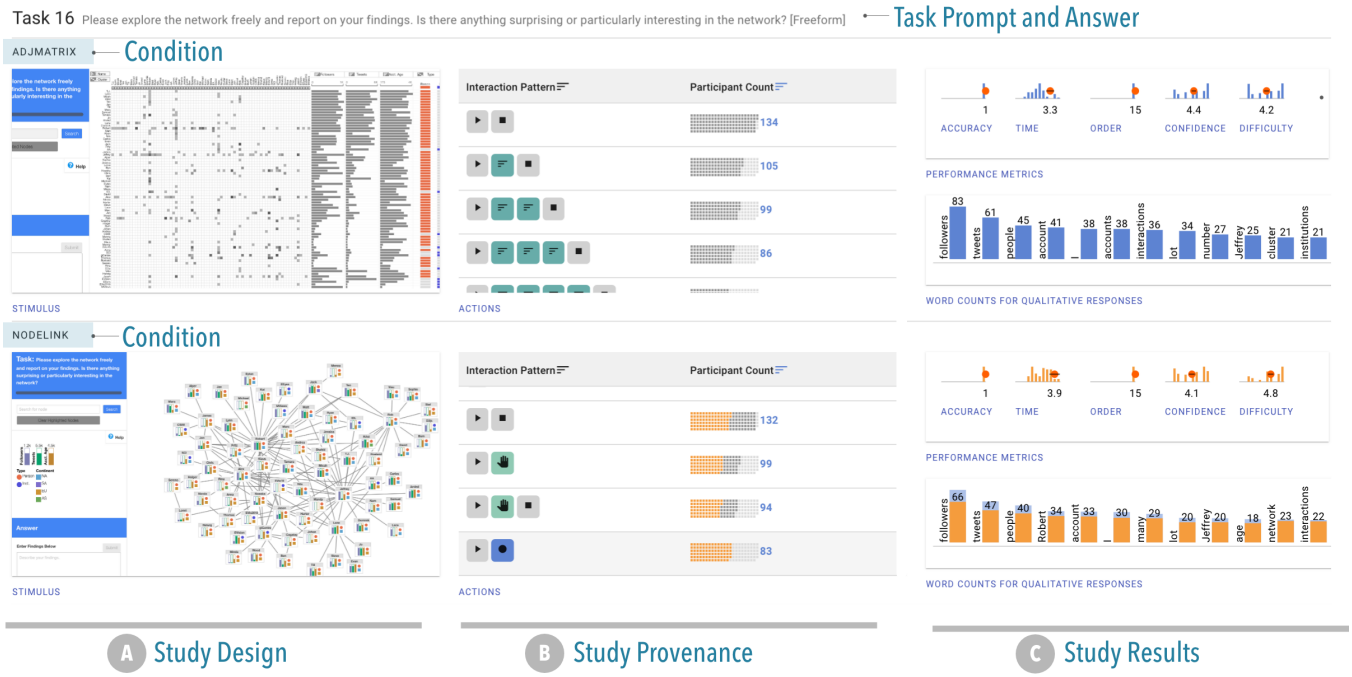

Figure 3: Task overview interface. (A) The condition, task description, and a screenshot of the stimulus are shown in the study design panel. (B) The study provenance panel shows common interaction patterns and how frequently they were used. (C) The results panel shows performance data and heuristics for qualitative data, such as word counts.

mouse clicks to high-level intents [36]. Whereas reVISit supports analysis of provenance data at various degrees of abstraction, our experience collecting and analyzing provenance revealed that (a) capturing fine-grained data is essential, but (b) post hoc analysis often requires moving up the 'ladder of abstraction' in order to extract semantically meaningful findings. The concept of different semantic levels is described by Gotz et al. [13], who characterize analytic behavior at different levels of granularity based on the semantic richness of the activity. In order to support analysis at different semantic levels, the event manager in reVISit (Figure 4) allows analysts to dynamically increase the level of abstraction of the original data (**T6**). This can be done in one of two ways: (1) event grouping and (2) sequence abstraction.

With event grouping, a new event is created as a proxy for a set of lower level events based on 'or' logic. An example of such grouping is to group 'click', 'hover', and 'drag' events into a 'mouse action' event. Alternatively, users can create sequences from lower level events. Grouping is particularly useful to capture entire analysis strategies in a single element. Figure 4 shows a 'Node Select' group that joins different ways of selecting a node (attribute row, row label, or simple selection) into a single 'Node Select' action. In creating sequences and groups, analysts can easily compare the performance metrics from participants with different strategies to determine the effect of each approach. (**T2**).

Analogously, sequence abstractions group events but uses 'and' logic while also considering the order of events. Figure 4 shows an example of a sequence event and a group being created: the 'Quick Answer' sequence is composed of starting the task, selecting an answer, and ending the task.

The event manager also supports configuring the color and label for each event glyph. Analysts can use these encodings to visually highlight events of interest in analysis across all views in reVISit.

5.3 Task Analysis

The *task analysis* view, shown in Figure 5, is used for the joint exploration of provenance and performance data (**T2**). To support this exploration, the data is shown in an interactive table, which enables faceting, grouping, and sorting on any attribute. The initial display of the table represents all participants who completed a given task. The columns in the table include attributes on study design, such as the condition, performance metrics such as accuracy

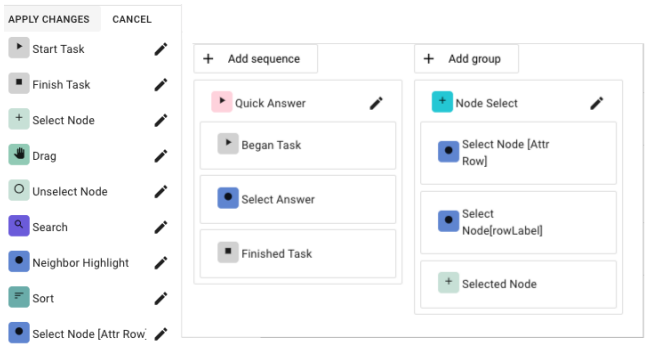

Figure 4: Event Manager view in reVISit. The interactive panel allows analysts to edit the icon, color, and label for each action in the dataset. Additionally, analysts can define groups and sequences of interest.

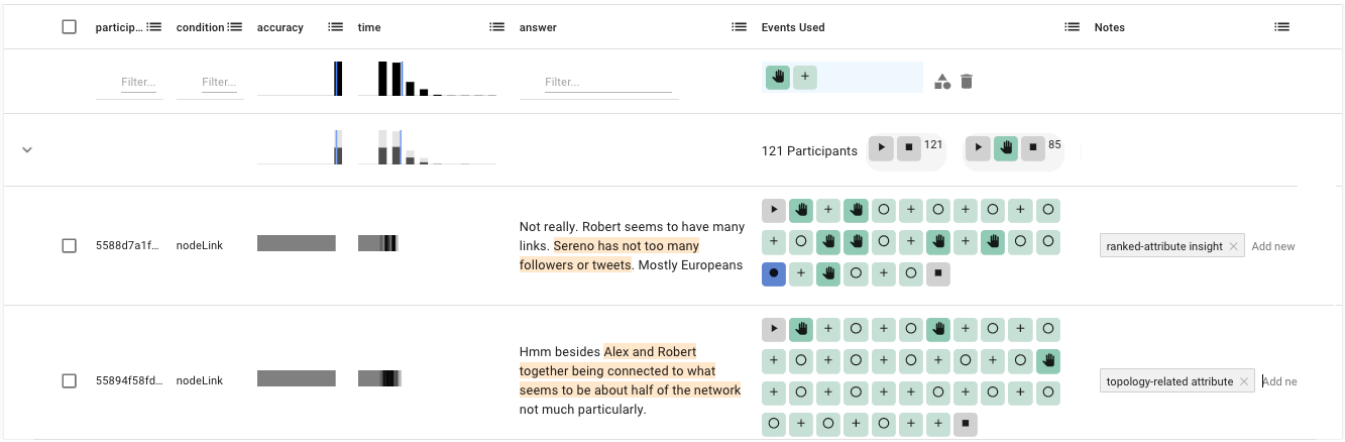

Figure 5: The task analysis panel allows analysts to select which attributes to visualize and then filter and group rows by any relevant attributes. Dragging a column header to the top row groups the table by that attribute, creating aggregate rows that can then be expanded to show individuals in that group. For free-form text fields, analysts can highlight and tag specific sections, adding qualitative codes and tags to capture analyst insight.

and time, provenance information with sequences of glyphs for each interaction, and a notes column to capture analyst insights during the analysis. A menu on the far right of the table controls which columns to show and hide. With these columns, analysts can leverage sorting, filtering, and grouping to perform their analysis. For example, investigating whether using a search feature in a study improved participant performance can be done with a filter-then-group operation (**T2**). The analysts can first use the filter header in the 'Events Used' column to filter rows to only those that contain a 'search' interaction. Dragging the event column header over to the grouping area at the top of the table groups all rows into two groups: those that contain the 'search' interaction and those that do not. These aggregated rows display a summary distribution of values across all columns. The analyst can now inspect and compare the average accuracy, time, confidence, and any other collected attributes for each of the groups in the table. Grouping can be done on multiple attributes. New groups are added in a hierarchy. In the current example, grouping by 'condition' then creates additional nested groups inside the 'contains search' and 'does not contain search' rows, one for each 'condition'.

Insights achieved with the task analysis can be captured in the 'Notes' column via 'tags' or 'codes', added directly in the row of interest (**T3**). Another useful application of tags is for qualitative coding of free-form text responses. The task analysis view supports highlighting specific segments of text in participant answers, and adding qualitative codes to those segments. This coding allows analysts to perform thematic analysis directly in the table, capturing dimensions of participant responses that require human interpretation to do so. These tags can then be exported with the accompanying data to perform additional analysis outside reVISit (**T4**).

5.4 Study Playback

Analysis tasks that rely on interaction provenance can be well served by the action glyphs used in the *task overview* and *task analysis* views. However, as a summary representation, these glyphs cannot capture the surrounding context in which the actions were performed. This context includes information such as the data element interacted with and the state of the surrounding visual components. To address the need for context, reVISit supports the playback of a participant analysis sequence, directly in the original study stimulus (Figure 9). This playback results in a video-like experience, similar to footage captured with observational user studies. The analyst can choose to either auto play all the interactions sequentially, or navigate to specific actions with the navigation strip. Visualizing participant actions in the study stimulus can help analysts to disambiguate similar analysis strategies (**T2**). For example, in the network user study we analyzed, participants were asked to find the shortest path between two nodes in the graph. Although several analysis strategies involved selecting and de-selecting nodes, playback of the interaction sequences allowed analysts to see exactly which nodes were being interacted with and thereby identify the search strategies used to find the shortest path.

The study playback can also help analysts inspect and understand unsuccessful strategies for solving tasks. This analysis is particularly important with complex techniques, where participants who are not yet familiar with a visualization may struggle to properly leverage its affordances to solve the task (**T5**). In the same network user study, for example, study playback for incorrect responses to the path task in the adjacency matrix revealed gaps in participants' understanding of how neighbors are represented in the matrix.

5.5 Participant Timeline

Study provenance data captures how long participants spend on each section of the study. This type of data is particularly important for understanding phases of the study that can influence performance on the tasks themselves, such as participant training. When evaluating complex interactive techniques, particularly in a crowdsourced setting, training is a key element of ensuring appropriate participant expertise [29]. Tracking how long participants spend on training and how well they interact with trials before starting the study can be fundamental in interpreting the results and

Figure 6: The participant timeline view shows the time participants spent on each portion of the study. It also highlights events such as when users browsed away from the study window, seen here as the blue segments overlaid on the gray bars.

can account for effects such as participants who did not properly understand the technique (**T5**).

The participant timeline view displays the study provenance data in a temporal context for each participant (Figure 6). Analysts can assign different colors to events of interest, such as browsing away from the study window, in order to highlight their duration. For events that also contain participant input, such as tasks and trials, the participant timeline shows a summary of analyst-selected metrics in labels positioned around the timeline. This information can give analysts an overview of how long participants took on given tasks, as well as how well they performed on each task.

One of the main tasks this view supports is quality control of participants (**T4**). Low average accuracies, as well as long times browsed away from the study window, for example, can indicate a low-effort participant who can be tagged for removal.

6 STUDY DESIGN GUIDELINES

The reVISit workflow and system can be used with new studies that leverage all stages of the proposed workflow, or with data from existing user studies. In this section, we outline considerations for both cases.

Using reVISit with new studies allows analysts to carefully consider the types and granularity of data captured during the study (Steps 1 and 2 in Figure 1). Although instrumenting a visualization system with detailed provenance tracking imposes a technical burden, the quality of insights analysts can expect with the reVISit workflow also increases. Furthermore, existing provenance tracking libraries can significantly reduce that burden [8]. We recommend capturing interactions at the highest possible semantic level, but also recording information on the user interface element interacted with. For example, if a user can sort in multiple ways in the visualization, logging them all as 'sort' is logical in that they all have the same outcome. However, also storing which UI elements were interacted with allows for insights into whether certain features are being leveraged as expected. Additionally, logging the data element associated with an interaction is valuable when interpreting task-based analysis strategies. With this information, analysts can discern, for example, interactions with a specific target element from another one and can reason about why an element was interacted with. Finally, using provenace tracking and reVISit from the beginning of the design phase facilitates pilots and testing of data collection modalities so that surprises after a study can be minimized.

When using reVISit in studies that are already completed, it is most useful for the data analysis capabilities described in Step 5 of Figure 1. Analysts can upload the collected provenance, log, and response data in tabular form to the reVISit system, with a column for every variable collected or computed by the analyst. Custom-derived metrics, such as the number of insights achieved during exploration, can be stored in additional columns and analyzed in conjunction with the captured variables. ReVISit uploads these files to an SQL database, and uses unique participant ids as foreign keys to link the different tables. Analysts can customize the variables along which the study data is faceted and aggregated (e.g., conditions, tasks) in a configuration file. An example study dataset is available at https://github.com/visdesignlab/reVISit.

7 IMPLEMENTATION

ReVISit is implemented as a React web application using Typescript and D3 on the client and Python and Flask on the server. The study data is stored in a MySQL database on the server, running in a separate Docker container on designated ports. Sequence mining queries as well as access to the raw study data are exposed through a REST API. The full stateful provenance data is stored in a firestore database. ReVISit is open source and uses the permissive BSD license. The reVISit tool can be accessed at https://vdl.sci.utah.edu/reVISit/, the source code is available at https://github.com/visdesignlab/reVISit.

8 CASE STUDIES

We evaluate the workflow and the reVISit tool with two case studies conducted on data from two published crowdsourced studies from our lab with over 300 participants and 100 participants, respectively. We asked the authors of each publication (who partially overlap with the authors of this paper) to use reVISit with their study data and report on their insights.

8.1 Multivariate Network Evaluation Study

Our team previously ran a crowdsourced user study to evaluate how the two main forms of visualizing multivariate graphs — the node-link diagram and the adjacency matrix — supported different types of exploration tasks [29]. We collected provenance data with the Trrack library [8], which captures both the state of the application and the series of interactions that each participant performs in a given task.

Each network visualization technique has its own strengths and weaknesses. For example, node-link diagrams are well suited to finding neighbors and paths but can become cluttered when displaying multiple attributes. Conversely, finding paths in adjacency matrices is known to be difficult. However, in dense networks or

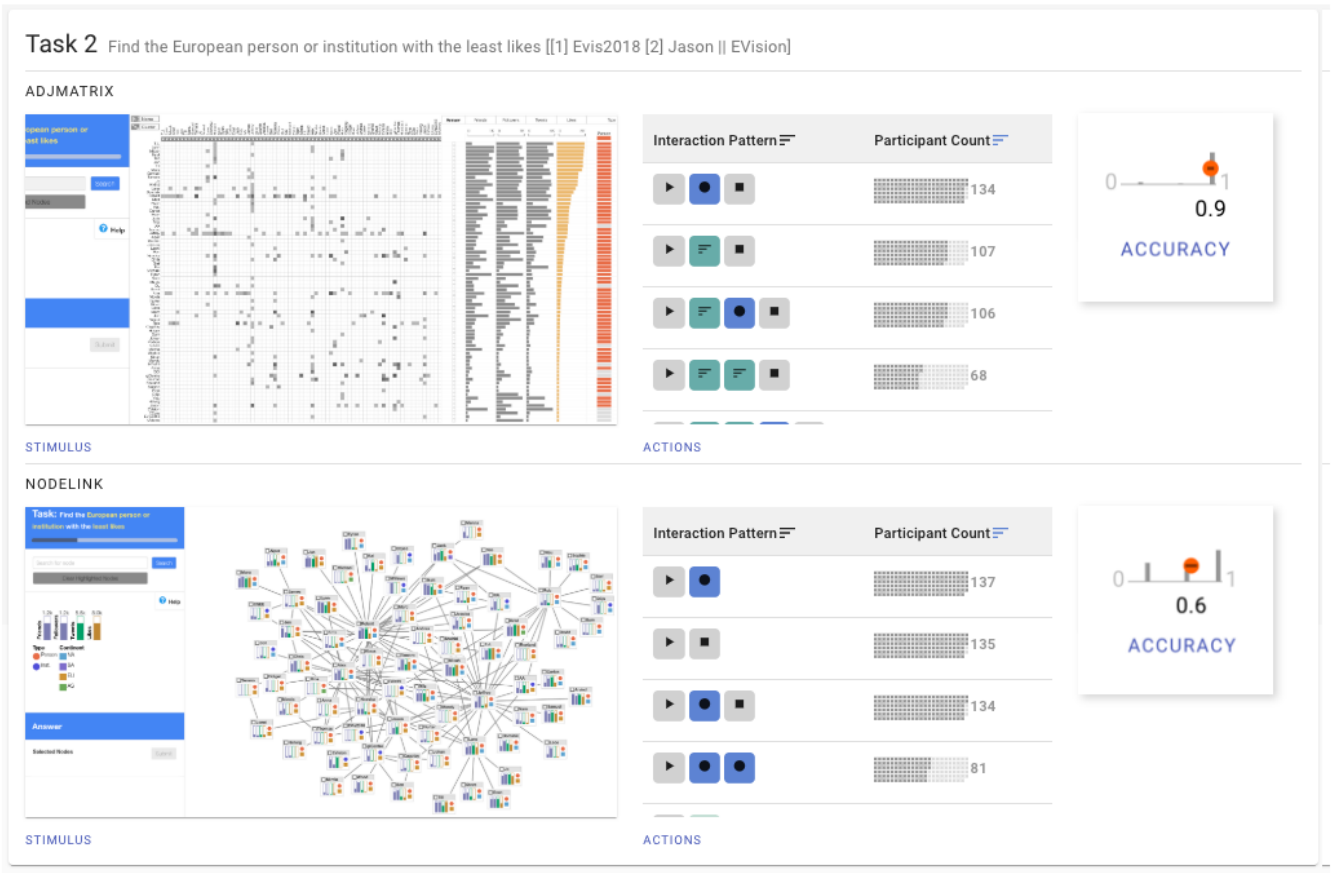

Figure 7: Overview showing stimulus and results for a task that investigates whether encoding additional attributes that are not essential to the task ('distractors') affects performance in either condition. The results indicate that participants who used the adjacency matrix had an average score of 0.9, compared to the node-link participants who scored 0.6. The node-link distribution reveals a bimodal trend with some participants performing well and others poorly.

large multivariate datasets, adjacency matrices lead to better performance [28].

In one of the study tasks, analysts were looking to investigate whether encoding additional attributes that were not essential to the task ('distractors') would affect performance in either condition. Participants were asked to 'Find the European person or institution with the least likes' in network visualizations that had several 'distractor' attributes encoded. The stimulus and results for each condition are shown in the task overview (Figure 7). The study results showed that adjacency matrix participants performed significantly better and faster, with an average accuracy of 0.9 (on a scale of 0–1) in the matrix vs the 0.6 in the node-link condition. Superior performance for the adjacency matrix for this task was expected, since encoding multiple attributes directly on the nodes in the node-link diagram can lead to visual clutter, and make solving the task harder. This task aimed to confirm the analyst's hypothesis that the sorting affordances of the adjacency matrix made it much more suitable to performing tasks on multiple attributes. The distribution plot of accuracies for the node-link diagram, however, showed a bimodal distribution (Figure 7), with some participants doing well and others poorly. Hovering over the top interaction patterns did not disambiguate the two user groups, so the analyst drilled down into the data in the task analysis section to further inspect the results.

The analyst grouped all participants by condition to separate the adjacency matrix trials from the node-link ones. Hovering over the histograms for accuracy and time confirmed the performance differences between the two conditions. The analyst then grouped on accuracy, creating one group for participants who got the answer correct (N=67) and one for those who did not (N=64). The 'Events Used' column displayed the top interaction sequences for each aggregate row, revealing that for 43 of the 67 participants who scored perfectly on this task used the *drag operation* one or more times. Conversely, the drag operation does not show up at all in the top five interactions for participants who got the task wrong. This finding was somewhat surprising since the task required participants to answer based on the attributes encoded in the node, and not on the structure of the graph. To more precisely assess the impact of dragging on successful interaction strategies for this task, the analyst created groups of participants based on whether or not they

Figure 8: Analysis of interaction strategies to solve the task that inspected whether encoding non-task-essential attributes hindered performance in either condition. The table is grouped by participants in the node-link diagram who dragged multiple times, and shows that the average accuracy for the 'multiDrag' approach is 74% whereas it is 51% for non-draggers.

had dragged nodes, shown in Figure 8. Iterating over this strategy led to the finding that participants who had dragged nodes multiple times (three or more times) had an average accuracy 25 percentage points higher than those who did not. This 'multidrag' interaction strategy also impacted other participant metrics. Notably, users of this interaction strategy took nearly a minute longer to complete the task and had more confidence in their answers.

To understand why dragging enabled higher task accuracy, analysts selected participants from the 'dragged' group and watched the playback of their original study. The playbacks showed that the dragging operation was used to visually sort the nodes, shown in Figure 9. Participants scanned each node to assess whether it fit the criteria in the task and dragged it left or right accordingly. *This analysis revealed an unexpected way in which participants were*

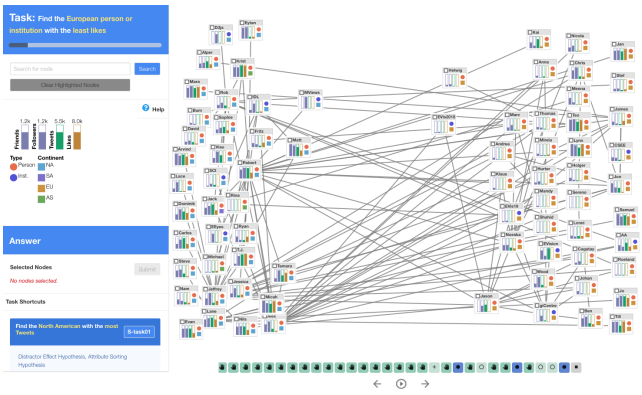

Figure 9: The playback view enables analysts to re-run an individual analysis session step by step. The figure shows a task for a participant who used the 'multidrag' approach to solve a task in the node-link condition. The playback reveals an unexpected analysis strategy to solve a task with a visualization that does not naturally support sorting.

using the node-link diagram to solve the task using a spatial arrangement strategy [2]. Although this finding confirmed the analysts' hypothesis that the node-link diagram does not easily support tasks that involve sorting on multiple attributes, it also provided novel information that can guide future implementations of interactive graph visualizations.

Another task in the study asked participants to freely explore the network and report on their insights. Analysts then used reVISit to perform a qualitative coding of participant responses, categorizing insights into whether they were based on the network's topology, attributes, or both. Grouping participants based on their assigned network visualization revealed that the adjacency matrix predominantly led to overview and ranked-attribute insights whereas the node-link diagram resulted in many topology-only, topology-attribute, and within-node-attribute comparison insights.

Overall, in a one-hour analysis session, **our analyst found eight interesting patterns** between interaction strategies and performance metrics of time, accuracy, confidence, and perceived difficulty. Analysts commented that viewing participants' analysis strategies also served to validate the design decisions they made when developing the study. For example, the adjacency matrix allowed users to group neighbors by clicking on a node label. Analyzing strategies for solving neighborhood tasks in reVISit showed that participants who used the grouping neighbors feature (N=82) exhibited greater accuracy (90%) and a faster task completion time (1 minute) when compared to those who did not (N=47, accuracy= 68%, time=1.25 minutes). Using this finding, the analysts could validate that their design decision (enabling users to group neighbors) had a demonstrable impact on participant performance.

A valuable outcome that became apparent from this session is that reVISit can provide analysts with names for different analysis strategies. Our analysts used terms like 'the multidrag' approach, or the 'sort and select' strategy, when discussing which approaches worked well and which did not. Naming analysis strategies serves to give semantic meaning to similar sequences of events, which greatly facilitates discussion of interactive features among visualization designers. Using specific terms to discuss analysis approaches has

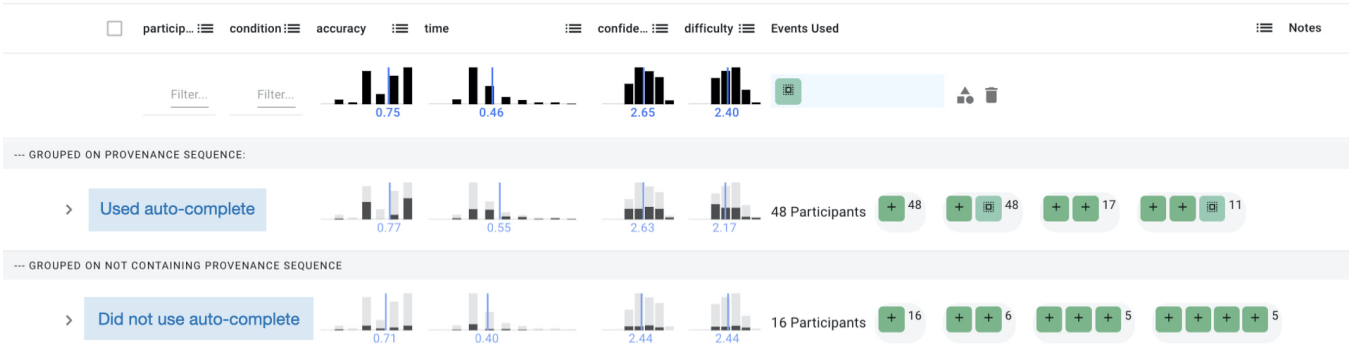

Figure 10: Participants grouped by whether or not they used the auto-complete feature available in the supported condition. Results show that using the auto-complete (middle row) led to improved accuracy and higher confidence than not using auto-complete (bottom row).

also been done in observational studies [24], but such studies lacked the scale necessary to make judgments on effectiveness.

8.2 Predicting Intent

Gadhave et al. [12] introduced a method to infer analyst intent for selections and brushes in scatterplots. To evaluate their approach, they ran a crowdsourced user study with 130 participants, where each participant performed tasks in two conditions: (1) manual selection of points in a scatterplot with no computer assistance, and (2) selection with an 'auto-complete' feature that suggests selections once the participant starts interacting with the visualization. They collected detailed provenance logs as well as performance metrics during the study. Statistical analysis of the performance metrics revealed some surprising findings, such as longer completion times for the 'computer supported' than the 'manual' condition, and less difference in accuracies between the two conditions than they had anticipated. To investigate how participants were using the + selections and auto-complete features, two authors from the Gadhave et al. [12] study used reVISit to analyse the data from the study.

The analysts first grouped all participants by condition, and then created additional groups separating participants based on whether they had used the 'auto-complete' feature available in the supported condition.

Separating participants in the 'supported' condition showed that those who used auto-complete took 10% longer to complete the task than those who did not. Comparing trials that did not use auto-complete in the supported condition with those that did not have that option in the manual condition showed similar time and accuracy metrics, but much higher self-reported confidence in the manual condition. This finding led the analysts to infer that one of the side effects of showing an auto-complete value that was not used by participant lowered the confidence in the responses (Figure 10).

Analysis of a separate task that asked participants to select points in a cluster revealed specific strategies for highly successful participants. Manual participants who used a + select followed by an de-select operation tended to have greater accuracy (90%, N=7) when compared to those who did not use a de-select (73%, N = 54). This 'select and refine' analysis strategy took nearly double the amount of time (0.63 vs 0.37 minutes) and was seen only in tasks with greater difficulty.

A post-analysis reflection revealed that analysts gained insights into which strategies worked best for solving tasks. Additionally, the analysts reported that using reVISit highlighted aspects of provenance that they did not record (such as hovering over nodes) that would have been useful in disambiguating people who explored the auto-select option versus those who simply did not engage with the prompt. The authors plan on running a follow-up study, and will use these insights to plan their visualization and the granularity of provenance to collect.

9 DISCUSSION AND CONCLUSION

When evaluating complex interactive visualizations with empirical studies, traditional analysis methods cannot account for variations in participant analysis strategies. We present a workflow and associated suite of methods for capturing and analyzing detailed provenance data to shed light on how these strategies affect study results. We believe this work is just a first step toward supporting user studies that evaluate complex visualization techniques. Future work can build on this approach by expanding on the types of provenance to include, for example, audio data from think-aloud protocols or eye-tracking data. Views that visualize mouse movement and hovers could provide additional insights into a participant's attention during analysis. Additionally, expanding on the types of event sequence mining algorithms, and enabling more complex, regular-expression-like event grouping mechanisms, could give analysts more flexibility in finding relevant analysis strategies.

A limitation of the current implementation of reVISit is that we do not consider timing or duration of events in our event sequence mining or visualization. This data could be integrated as an additional layer in the participant timeline. For the event sequence mining, both visualization and querying/filtering based on temporal information would make additional ways of grouping/visualizing participant trials possible.

The statistics-enabled workflow in reVISit relates to the broader discussion in visualization and human-computer interaction concerning statistical standards and tools that support them. ReVISit,

through features like faceting, filtering, and visualizations of statistical measures, provides users with a range of capabilities for analyzing user studies beyond aggregate measures — like means and confidence intervals — that form the basis of a large portion of task-based empirical studies in visualization. As illustrated in recent works [10], systems that support multiple comparisons in datasets must take care to avoid leading users toward "p-hacking", which involves exploring and manipulating data and making statistical comparisons until a desired result is found. ReVISit was designed with these considerations in mind. For example, no statistical tests are run in reVISit. Instead, the bootstrapped 95% confidence intervals aim to align workflows in reVISit with statistical standards recommended in methods-focused proposals [9]. Furthermore, given reVISit's explicit focus on using interactive visualization to disambiguate variance in participants' performance in user studies, future iterations may align efforts to use statistical approaches more robust to variance such as the Bayesian methods proposed by Kay et al. [21] and others.

Another area for future work is data integration: we plan on developing guidelines on how to store provenance data, so that it can easily be ingested by a tool like reVISit, without the need for pre-processing. In this way, studies could monitor pilots in real-time, and re-play and analyze data as soon as a participant has completed a task, and flexibly adjust the study design or data collection modalities if problems become apparent.

The visualization community has significant knowledge about how to design static and simple interactive visualizations to support data exploration. However, more complex interactive visualizations are only now being studied more closely, and efforts such as reVISit can provide valuable insights to inform the design of these interactive visualizations. Analyzing *how* participants engage with interactive visualizations can either validate or challenge our assumptions as visualization designers. In either case, they inform our efforts in this direction, and pave the way to creating interactive visualizations that cater to the users it aims to support.

ACKNOWLEDGMENTS

We want to thank Kiran Gadhave for making his study data available and for his help with the provenance tracking library. We gratefully acknowledge funding by the National Science Foundation (IIS 1751238, IIS 1815587, and OAC 1835904).

REFERENCES

[1] Anne Adams, Peter Lunt, and Paul Cairns. 2008. A Qualititative Approach to HCI Research. In *Research Methods for Human-Computer Interaction*, Paul Cairns and Anna Cox (Eds.). Cambridge University Press, Cambridge, UK, 138–157.

[2] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to Think: Large High-Resolution Displays for Sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 55–64. https://doi.org/10.1145/1753326.1753336

[3] Tanja Blascheck, Markus John, Kuno Kurzhals, Steffen Koch, and Thomas Ertl. 2016. VA2: A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 61–70. https://doi.org/10.1109/TVCG.2015.2467871

[4] Rita Borgo, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, Fintan McGee, Luana Micallef, Tatiana von Landesberger, Katrin Ballweg, Stephan Diehl, Paolo Simonetto, Michelle Zhou, Stephan Diehl, Paolo Simonetto, and Michelle Zhou. 2017. Crowdsourcing for Information Visualization: Promises and Pitfalls. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Vol. 10264. Springer, Cham, 96–138. https://doi.org/10.1007/978-3-319-66435-4_5

[5] Bram C.M. Cappers and Jarke J. van Wijk. 2018. Exploring Multivariate Event Sequences Using Rules, Aggregations, and Selections. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 532–541. https://doi.org/10.1109/TVCG.2017.2745278

[6] Sheelagh Carpendale. 2008. Evaluating Information Visualizations. In *Information Visualization: Human-Centered Issues and Perspectives*, John T. Stasko, Jean-Daniel Fekete, Chris North, and Chris North (Eds.). Springer, 19–45. https://doi.org/10.1007/978-3-540-70956-5_2

[7] Juliet Corbin and Anselm Strauss. 2014. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications.

[8] Zachary T Cutler, Kiran Gadhave, and Alexander Lex. 2020. Trrack: A Library for Provenance Tracking in Web-Based Visualizations. In *Proceedings of IEEE VIS Short Papers*. IEEE. https://doi.org/10.31219/osf.io/wnctb

[9] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13

[10] Philipp Eichmann, Emanuel Zgraggen, Zheguang Zhao, Carsten Binnig, and Tim Kraska. 2016. Towards a Benchmark for Interactive Data Exploration. *IEEE Data Eng. Bull.* 39, 4 (2016), 50–61.

[11] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. 2016. The SPMF Open-Source Data Mining Library Version 2. In *Machine Learning and Knowledge Discovery in Databases*, Bettina Berendt, Björn Bringmann, Élisa Fromont, Gemma Garriga, Pauli Miettinen, Nikolaj Tatti, and Volker Tresp (Eds.). Vol. 9853. Springer International Publishing, Cham, 36–40. https://doi.org/10.1007/978-3-319-46131-1_8

[12] Kiran Gadhave, Jochen Görtler, Zach Cutler, Carolina Nobre, Oliver Deussen, Miriah Meyer, Jeff Phillips, and Alexander Lex. 2020. Capturing User Intent When Brushing in Scatterplots. *Preprint* (2020). https://doi.org/10.31219/osf.io/mq2rk

[13] David Gotz and Michelle X. Zhou. 2009. Characterizing Users' Visual Analytic Activity for Insight Provenance. *Information Visualization* 8, 1 (2009), 42–55. https://doi.org/10.1057/ivs.2008.31

[14] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Nicola Cosgrove, and Marc Streit. 2016. From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum* 35, 3 (2016), 491–500. https://doi.org/10.1111/cgf.12925

[15] Hua Guo, Stephen R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. 2016. A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 51–60. https://doi.org/10.1109/TVCG.2015.2467613

[16] Yi Han, Gregory D Abowd, and John Stasko. 2016. Flexible Organization, Exploration, and Analysis of Visualization Application Interaction Events Using Visual Analytics. In *IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis*. IEEE, 4.

[17] Weidong Huang, Peter Eades, and Seok-Hee Hong. 2009. Measuring Effectiveness of Graph Visualizations: A Cognitive Load Perspective. *Information Visualization* 8, 3 (2009), 139–152. https://doi.org/10.1057/ivs.2009.10

[18] J. Hullman, E. Adar, and P. Shah. 2011. Benefitting InfoVis with Visual Difficulties. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2213–2222. https://doi.org/10.1109/TVCG.2011.175

[19] Petra Isenberg, Torre Zuk, Christopher Collins, and Sheelagh Carpendale. 2008. Grounded Evaluation of Information Visualizations. In *Proceedings of the 2008 Conference on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*. ACM, Florence, Italy, 1–8. https://doi.org/10.1145/1377966.1377974

[20] Jian Pei, Jiawei Han, B. Mortazavi-Asl, Jianyong Wang, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. 2004. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. *IEEE Transactions on Knowledge and Data Engineering* 16, 11 (Nov. 2004), 1424–1440. https://doi.org/10.1109/TKDE.2004.77

[21] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300432

[22] Maurits Kaptein and Judy Robertson. 2012. Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*. ACM Press, Austin, Texas, USA, 1105. https://doi.org/10.1145/2207676.2208557

[23] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1375–1386. https://doi.org/10.1145/3025453.3025592

[24] Brittany Kondo and Christopher Collins. 2014. DimpVis: Exploring Time-Varying Information Visualizations by Direct Manipulation. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)* 20, 12 (2014), 2003–2012. https://doi.org/10.1109/TVCG.2014.2346250

[25] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536. https://doi.org/10.1109/TVCG.2011.279

[26] Miro Mannino and Azza Abouzied. 2018. Qetch: Time Series Querying with Expressive Sketches. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*. Association for Computing Machinery, New York, NY, USA, 1741–1744. https://doi.org/10.1145/3183713.3193547

[27] A McAlister, D Lee, KM Ehlert, RL Kajfez, CJ Faber, and MS Kennedy. 2017. Qualitative Coding: An Approach to Assess Inter-Rater Reliability. In *ASEE Annual Conference & Exposition*. ASEE.

[28] Carolina Nobre, Miriah Meyer, Marc Streit, and Alexander Lex. 2019. The State of the Art in Visualizing Multivariate Networks. *Computer Graphics Forum (EuroVis)* 38, 3 (2019), 807–832. https://doi.org/10.1111/cgf.13728

[29] Carolina Nobre, Dylan Wootton, Lane Harrison, and Alexander Lex. 2020. Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3313831.3376381

[30] L. M. K. Padilla, S. C. Castro, P. S. Quinan, I. T. Ruginski, and S. H. Creem-Regehr. 2020. Toward Objective Evaluation of Working Memory in Visualizations: A Case Study Using Pupillometry and a Dual-Task Paradigm. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 332–342. https://doi.org/10.1109/TVCG.2019.2934286

[31] D. Park, S. M. Drucker, R. Fernandez, and N. Elmqvist. 2018. Atom: A Grammar for Unit Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 24, 12 (2018), 3032–3043. https://doi.org/10.1109/TVCG.2017.2785807

[32] E.D. Ragan, A. Endert, J. Sanyal, and J. Chen. 2016. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics (VAST '15)* 22, 1 (2016), 31–40. https://doi.org/10.1109/TVCG.2015.2467551

[33] Anselm Strauss and Juliet Corbin. 1994. Grounded Theory Methodology. *Handbook of qualitative research* 17 (1994), 273–85.

[34] K. Wongsuphasawat and D. Gotz. 2012. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2659–2668. https://doi.org/10.1109/TVCG.2012.225

[35] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. 2020. Survey on the Analysis of User Interactions and Visualization Provenance. *Computer Graphics Forum* 39, 3 (2020), 757–783. https://doi.org/10.1111/cgf.14035

[36] Ji Soo Yi, Youn ah Kang, John Stasko, and J.A. Jacko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '07)* 13, 6 (2007), 1224–1231. https://doi.org/10.1109/TVCG.2007.70515