Exploiting Analysis History to Support Collaborative Data Analysis

Ali Sarvghad* and Melanie Tory*
University of Victoria

ABSTRACT

Coordination is critical in distributed collaborative analysis of multidimensional data. Collaborating analysts need to understand what each person has done and what avenues of analysis remain uninvestigated in order to effectively coordinate their efforts. Although visualization history has the potential to communicate such information, common history representations typically show sequential lists of past work, making it difficult to understand the analytic coverage of the data dimension space (i.e. which data dimensions have been investigated and in what combinations). This makes it difficult for collaborating analysts to plan their next steps, particularly when the number of dimensions is large and team members are distributed. We introduce the notion of representing past analysis history from a *dimension coverage* perspective to enable analysts to see which data dimensions have been explored in which combinations. Through two user studies, we investigated whether 1) a dimension oriented view improves understanding of past coverage information, and 2) the addition of dimension coverage information aids coordination. Our findings demonstrate that a representation of dimension coverage reduces the time required to identify and investigate unexplored regions and increases the accuracy of this understanding. In addition, it results in a larger overall coverage of the dimension space, one element of effective team coordination.

Keywords: Visualization, History, Distributed collaboration, Coordination, Dimension coverage, Tabular data.

Index Terms: Human-centered computing: Collaborative and social computing systems and tools, Visualization

1 INTRODUCTION

We introduce and evaluate a *dimension coverage* perspective on data analysis history to support the visual data analysis of tabular data. A dimension coverage view reveals which dimensions (i.e. attributes or variables in a tabular dataset) have been explored in past analysis and in which combinations. We demonstrate that revealing this information can help collaborating analysts to better coordinate their work by improving their understanding of each other's activities.

We focus on exploratory data analysis tasks by collaborators who are distributed in both time and space. According to Munzner's task abstraction [20], *exploration* is searching the data space for unknown targets in unknown locations. An analyst doing exploratory analysis constantly needs to formulate new goals/questions (targets) and decide on subsets of data (locations) to investigate. With distributed collaboration, there is a further challenge of the 'hand-off', where work started by one collaborator is continued by another; here the second person needs to learn what has been already done and then choose which aspects to investigate next. Gaining a good understanding of this past work is critical to ensuring effective coordination and minimizing duplicate effort. For example, imagine that Mary has begun evaluating business performance by exploring sales data. She has looked at dimensions 'Sales', 'Profit', 'Margin' and 'Product Category' for interesting patterns and/or outliers. Following this initial analysis, she has passed the task to her colleague Joe to continue. To avoid duplicating Mary's work and to evaluate business performance from all possible angles, Joe needs to know what Mary has investigated and what other avenues remain. Joe's task would be difficult with existing visual history representations: they would show the set of charts Mary created but would not make it easy to figure out which dimensions she investigated. In contrast, our approach directly reveals dimension coverage information to support such review tasks.

Visualization history modules track past visualization states throughout a data analysis session and therefore track which data dimensions have been investigated. However, this information is not easily accessible in most common visual representations of history (typically a list or graph of past visualization states or actions). Thus, existing history designs provide very limited support for understanding dimension coverage. We propose an alternative perspective on this information: a *dimension view* that makes dimension coverage information explicit.

We speculate that providing a dimension coverage view for collaborative data exploration will 1) improve speed and accuracy of discovering information about underexplored dimensions, in comparison to a typical linear history view, and 2) improve coordination by enabling an analyst to focus on aspects that were less investigated by others. Since (2) is only plausible if (1) holds, we first compared the effects of adding dimension coverage information to a typical sequential history view (Study 1). Results showed that people with access to Dimension view acquired more detailed information about the analytic coverage of dimension space in less time. Study 2 then assessed coordination. It showed that participants with access to Dimension view were much more likely to focus on data that were previously underexplored. They performed a more thorough overall investigation of the problem space, indicating better coordination with a collaborator. These findings demonstrate the value of representing analysis history from a dimension coverage perspective.

2 RELATED WORK

Because our work offers a new perspective on data analysis history, we first review prior work on visualization histories. We then describe research that investigates the use of history to support collaborative data analysis, our ultimate objective.

2.1 Visual History for Data Analysis

Visualization histories automatically record past work of an analyst, enabling them to easily revisit earlier states of the analysis. There are two main models of visualization history:

* asarv@uvic.ca, mtory@cs.uvic.ca

state-based and action-based [11]. History tools with an action-based model capture single or groups of user interactions; these interactions typically result in a transformation of the system and/or visualization. In contrast, state-based history tools record information about the state of the system and/or visualization at specific times; these records can be used to duplicate that system state at a later time. State-based history tools may also include analyst externalizations such as notes and annotations.

At the system level, history tools commonly utilize node-link data structures to store history internally [11]. Often this same structure is used at the interface level to visually represent history. Depending on the underlying history model, nodes of the graph may represent either actions or states, and connections may show dependencies or precedence. GRASPARC [3], ExPlates [15], GraphTrails [10], VisTrails [1] and CzSaw [16] are example history tools that employ a node-link graph to visually represent the analysis history. Revisiting the graph nodes helps an analyst to recall/review/reuse previous states or actions. Other visualization techniques such as tree maps [8] and tag clouds [6][22] have also been used to represent information about the analysis history.

State-based history models intrinsically contain information about dimension space coverage. Yet the common visual representations (usually a sequential list or tree) mainly facilitate linear review and reuse of states. Existing history tools therefore overlook the potential to provide insight into the explored dimensions and data values. We exploit this potential by adding a dimension coverage view that supports understanding the analytic coverage of the dimension space; that is, which dimensions have been examined in which combinations (and which have not).

2.2 Visual History in a Collaborative Context

Visualization history enables an analyst to review the work of a collaborator situated in a different time and place (i.e. asynchronous, distributed collaboration) to come up to speed on the state of the analysis process. Gaining an understanding of what has been examined by others helps an analyst to recognize what is still left to investigate. In synchronous collaborative work, real-time shared views and instant-communication modalities can help in building common ground. For instance, CoMotion [7] enables sharing of personal views across the group. Similarly, Cambiera [14] enables an analyst to maintain an awareness of a collaborator's search queries and reviewed documents for co-located analysis of document collections. However, in an asynchronous context, a collaborator must rely on trails of information left behind by the previous analyst.

Asynchronous collaboration tools most commonly capture and share externalizations (recorded findings, hypotheses, etc.). Sense.us [12], CommentSpace [25], Analytic Trails [18] and ManyEyes [23] are example visualization tools that use externalizations to provide awareness. Wattenberg et al. [24] suggested using information scent (i.e. attention pointers that assist a person in navigating the information space) to provide visual cues into the past exploration of time series data. In their prototype, uninvestigated time series were highlighted to help people discover uninvestigated data. However, this design fell short of fully exposing the investigation of dimension space. Similarly, Willet et al. [26], incorporated visual cues into common interface widgets to help collaborators identify under-explored data values. This approach is limited to data values and does not provide information about dimension space coverage, which is important for exploring multidimensional data sets.

3 Visually Representing Dimension Coverage

Here we describe the most salient features of the history prototypes used in our two studies, including our representations of dimension coverage. Due to space limitations, the complete descriptions of these prototypes, their functionality, and the control (baseline) tools are presented as supplementary material.

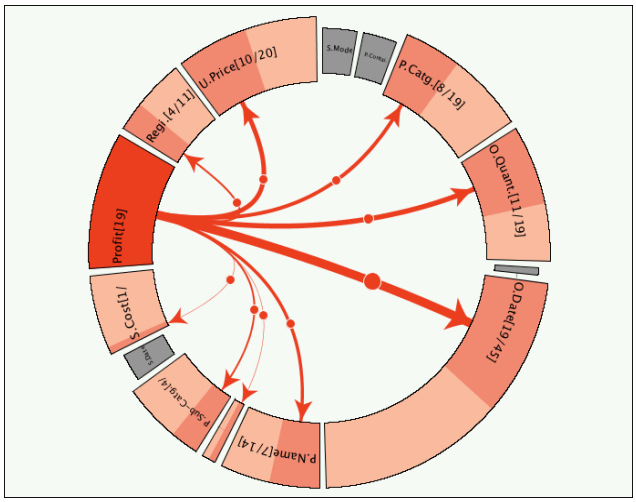

Figure 1: Circular design of Dimension View. Selecting a dimension (e.g. Profit, in solid red) shows co-mapping information by arrows stemming from the selection and ending at co-mapped dimensions (shown in lighter shades of red). The darker shade of red encodes the frequency of co-mapping with the selected dimension (e.g. out of 45 total mappings, O.Date was co-mapped with Profit 19 times).

3.1 Dimension view

Inspired by the notion of information scent [21], we propose visually representing the analysis history from the perspective of dimension space coverage, a perspective we call *Dimension view*. We aimed to facilitate quickly understanding analytic coverage of the dimension space, specifically, which dimensions had been explored and in what combinations and frequencies. (E.g., "Has my collaborator investigated all the dimensions? Are there unexplored topics that I should focus on?").

The primary focus at this stage of our research was to assess the value of a dimension centric perspective. While it seemed promising, no previous research had reported such a view or validated its value for collaborative analysis. Thus our primary contributions are the studies reported in sections 4 and 5; they explore the value of a dimension coverage perspective. To complete these studies, however, we needed a visual representation that could expose dimension coverage. Rather than attempt to explore the (immense) possible design space, we followed an iterative design process and attempted to make good design choices based on current perceptual knowledge. The resulting designs, described in this section and further in the supplemental material, are imperfect but were sufficient to explore the dimension coverage idea in our studies. Future work can further explore the design space of dimension coverage views.

We explored two different designs, used in Studies 1 and 2. The initial design (Figure 1, used in Study 1) closely resembled a Circos plot [17]. Each investigated dimension was represented as a curved trapezoidal segment. Collectively, segments formed a doughnut. Labels on segments display the name of the dimension represented by the segment and the total number of charts that

dimension appeared within. Length of each arc was proportional to the total number of times the dimension was included in a visualization. Relationships were represented as curves connecting the segments, with width relative to the total co-mapping frequency of the pair (i.e. how often the pair was included together in a chart).

After Study 1, we changed the design to a treemap with a squarified layout (Figure 2, used in Study 2) where cells represented data dimensions. This change addressed the most important complaints/suggestions from both participants in study 1 (enhance legibility of labels) and several visualization experts (facilitate discovery of trinary or higher order co-mapping relationships between dimensions, improve scalability and make better use of space).

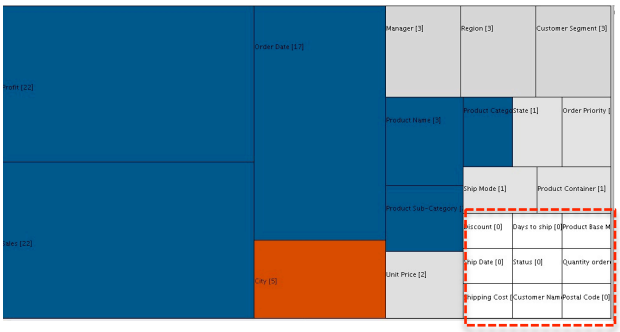

Figure 2: Treemap design of Dimension view. Uninvestigated dimensions are rendered with white background (see red border, added for illustration purposes). Investigated dimensions are non-white. Selecting one or more dimensions (e.g. City, in orange) shows co-mapped dimensions in blue.

Treemaps are typically used to visualize hierarchical data. Although we only had two categories ('investigated' and 'uninvestigated'), we found that the space-filling nature and scalability of the treemap made it a suitable choice for this view. The initial rendering of the view (uninvestigated dimensions in white, investigated ones in grey) revealed the relative frequency of dimensions in the prior analysis and therefore the focus of prior work. Interactions enabled users to discover co-mapping dependencies. When a user clicked on a dimension, the selected dimension's cell became orange and any dimensions that had been included in a visualization along with this dimension (i.e. co-mapping) became blue. Other cells remained unaffected. A comprehensive description of both versions and the supported interactions can be found in the supplementary material.

3.2 History tool prototypes

We built two prototype history tools, each incorporating a version of Dimension view (as described above). Both prototypes also included the most common (linear) representation of history, which we call Sequence view and which revealed the temporal progression of the previous analysis. Like Dimension view, Sequence view also underwent two major design revisions. Initially, this view contained a list of visualization thumbnails (Figure 3, used in Study 1), ordered by time of creation from top to bottom. Each thumbnail was labeled with a timestamp and information about dimensions involved in the visualization (e.g. 13:50:40, Margin, City, Lines). Double clicking on a thumbnail opened the full size chart in an external window.

We redesigned Sequence view following feedback from Study 1. As shown in Figure 4, our second design used a non-cyclic directed graph to represent the branching structure of the analysis process. Each node in the graph represented a visualization. Directed links depicted the progression of analysis over time. Each branch indicated a line of inquiry. Revisiting a previous state during analysis marked the beginning of a new line of inquiry and added a new branch to the graph stemming from the revisited visualization. This visual representation of Sequence view is very similar to common representations of chronological progress of history (e.g., [4], [16], [22]). This design was used in Study 2.

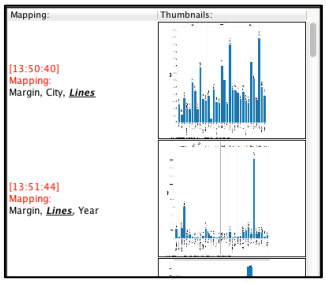

Figure 3: Initial design of Sequence View, used in Study 1.

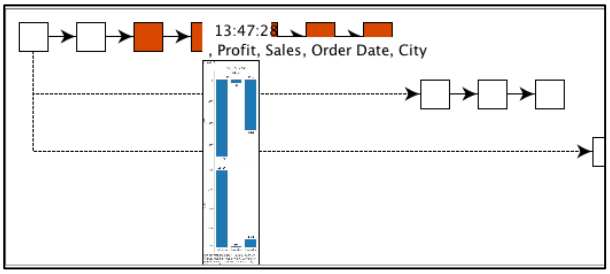

Figure 4: Sequence view used in Study 2. Hovering the mouse over a node shows a thumbnail image of the visualization and a list of dimensions included in that visualization.

Hovering the mouse over a node in the graph popped open a thumbnail view of the visualization represented by the node and information regarding the mapped dimensions (Figure 4). Similar to the first design, clicking opened a full size view.

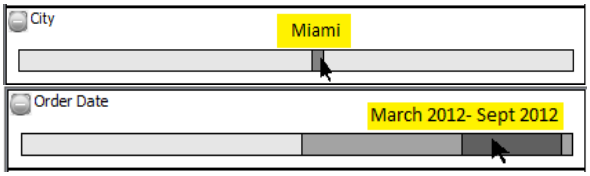

Figure 5: Mouse hover interaction in Data view. In the upper part, the user learns that the city 'Miami' has been investigated more than other cities in the data set. In the bottom part, the user discovers that a specific date range (Mar.–Sept. 2012) has been investigated more than the earlier dates.

In addition to Sequence view, the Study 2 prototype also included a Data view (based on the recommendation of visualization experts who reviewed our earlier design). Data view provided information about the coverage of data space (this goes beyond Dimension view, which only revealed coverage of dimension space). As shown in Figure 5, unique values in each dimension (e.g., all the unique city names under the dimension City), in ascending order from left to right, were represented in a bar. Darker shades indicated data values that were included in more charts within the previous analysis. For example in Figure 5, all date values of the Order Date dimension are represented in the lower bar, ordered from oldest to most recent. Each box marks a specific date range and luminance encodes the magnitude of

investigation (darker = more). Hovering the mouse over the darkest box shows that the previous analyst focused most on March 2012 to September 2012. In other words, the analyst created more charts with date values filtered to include this range. We use filtering information extracted from the history file to compute this view.

In both prototypes, Sequence view, Dimension view and Data view (when present) worked in coordination. Selecting dimensions in Dimension view filtered or highlighted corresponding items in Sequence view.

4 Study 1: Understanding Past Analysis

Study 1 assessed what value (if any) Dimension view adds to a common linear representation of history for understanding a collaborator's prior coverage of dimension space. The main rationale behind this study was to validate the hypothesized value of a dimension coverage view before delving into further research.

4.1 Study Design

We compared our first prototype (see Figures 1 and 3, plus supplemental material) to a baseline version containing only Sequence view (Figure 3) using a between-subjects experiment. Our rationale for this comparison was (1) the baseline tool was very similar to the most common history tools (e.g., [11], [19], [22]) so it represented current best practice and (2) Dimension view was designed to work in coordination with a sequence view, so investigating Dimension view alone would not make sense. Baseline version users could review history sequentially (by looking at visualizations one by one) or selectively (by searching for visualizations with certain dimensions). Details about a visualization could be viewed by hover query or by opening the full size visualization. By making our baseline version identical to the full version in all ways except for Dimension view, our experimental design enabled us to conclude that any difference in performance was caused by Dimension view.

4.2 Data

In order to prepare the history file, we asked a senior computer science PhD student with considerable management background and visual data analysis experience to work on a problem using Tableau Public. The student was not involved in the design of the prototype and is not an author of this paper. She investigated sales data (tabular data with 25 dimension and 8400 records) to try to explain an unexpected drop in profits against an unchanging number of orders over a four-year period. We employed a think-aloud protocol and asked her to explicitly express questions (e.g., "I wonder if there is a relationship between shipping cost and product container and ship mode"). We also asked her to save all visualizations that she used to answer each question.

We carefully reviewed the captured data and counted a total of 44 questions asked and 47 visualizations saved (3 Bar charts, 41 Line charts). Afterwards, we manually extracted mapping and filtering information for each chart (i.e. dimensions mapped on the X and Y axes, and any filtering of dimensions that returned a subset of values). This information was stored in a spreadsheet. This spreadsheet included a timestamp (i.e. time of creation, extracted from videos), dimensions mapped, filtering of dimensions and chart type. It was used as the analysis history file.

4.3 Participants

We recruited 20 computer science students as our participants (14 graduate, 6 senior undergraduate, 11 male, 9 female, average age of 27.9). They were randomly assigned to use either the full or the baseline tool. All the participants were required to have a basic understanding of visual data analysis and prior experience with tools that enable constructing visualizations or statistical charts based on data (e.g. Microsoft Excel).

4.4 Procedure

At the beginning of each study session, we gave a verbal description of the task. This was followed by an introduction to the tool features (either full or baseline tool). Afterwards, participants practiced using the tool by doing a short warm up task with an example history file. The warm up task required working with all the main features of the system. An experimenter was present during the warm up session and participants could stop and ask questions about the system, task and history file. After the warm up task was completed, participants were asked to read a short document that explained the task in detail. Later on, participants were given a booklet that contained questions about the collaborator's work (explained in Section 4.5). We asked our participants to think-aloud and verbalize all their thoughts while doing the task. On average, the preparation procedure took 40 minutes for full and 25 minutes for the baseline tool. The time difference was due to a larger set of features in the full version that required more time to explain. The time that users were given to practice and learn the systems was equal (~10 minutes). For both tools, we gave each participant a printed list of the tool's features and supported interactions that they could refer to (if needed) during the analysis. None of the baseline tool users referred to this list; five full version users made brief reviews.

4.5 Task

Participants answered questions about a past analyst's work but did not do any new analysis of their own. The task consisted of two parts, designed to test participants' ability to gain insight into the analysis history at different levels of granularity:

Part 1 examined participants' ability to determine which data dimensions were investigated versus left out. Participants selected data dimensions that were explored (i.e. ever mapped in a chart) from a list that contained names of all dimensions in the data set.

Part 2 examined participants' ability to understand which dimensions were investigated, with which frequencies, and in which combinations. Participants were first asked to select a statement, among 5, that correctly showed a list of explored dimensions ordered according to the total mapping frequency of each dimension (e.g. Order Date > Order Quantity > Unit Price). Participants then answered true/false questions regarding investigated combinations of dimensions (e.g. the relationship between Unit Price, Order Date and Ship Cost was investigated).

4.6 Data Capture

Participants were asked to record their answers in the paper-based task booklet. We audio recorded each session to capture participants' monologues while answering questions, and video captured the screen and logged users' interactions with the tool. Participants were asked to explicitly express when they started and finished each task. This helped us to accurately time duration of tasks for each participant using the videos. We also video recorded the short interviews that followed the analysis task.

4.7 Results

Full version users were both faster and more accurate in answering the questions. On average full version users spent 7.43 (SD=4.3) minutes to perform the two tasks, while baseline users spent 13.39 (SD=7.7) minutes. After using a log transformation to improve the fit to a normal distribution (SKW=0.196), we analyzed the time results by 2 (Tool) x 2 (Question Part) ANOVA, with Tool as a between-subjects factor and Question Part as a within-subjects factor. Note that questions were not in

random order, but we do not consider this important, as comparing between questions was not the purpose of our analysis. ANOVA showed a significant main effect of Tool ($F(1)$=9.4, p<0.004, η^2 = 0.687), demonstrating that the full version was significantly faster than the baseline. There was no significant effect of Question and no significant interaction between Question and Tool.

We also compared the accuracy of answers between full and baseline versions (maximum score 11). On average, full version users scored 10.1 (SD=0.88) while baseline version users scored 6.1 (SD=2.6). Results of a nonparametric Mann-Whitney test showed that users of the full version had significantly higher accuracy than users of the baseline version (W=9.5, p<0.003).

Overall, this preliminary evaluation strongly suggested that providing Dimension view facilitated gaining an understanding of the prior coverage of the dimension space. Participants' feedback in the follow-up interview also supported these findings. Due to lack of space, we only mention one example: "…[Dimension view] helped me to easily find out what variables were available…I saw that he [the initial analyst] didn't look at Returns so I looked into this". In summary, full version users attained more accurate relational / quantitative knowledge of analytic coverage in a considerably shorter time. We believe this is mainly due to the affordances of linear visual representations of history (Sequence view). This approach does not visually encode any information about investigated dimensions and combinations. To gain such insight, baseline users had to scroll (even when zoomed out) the list of thumbnails and titles and extract dimension coverage information. In addition, they had to rely on memory or externalization to keep this information.

5 Study 2: Continuing Past Analysis

Upon positive findings from Study 1, we assessed the effects of providing analytic coverage information on collaboration. We hypothesized that providing this information would improve coordination by encouraging analysts to take an investigative path more divergent from the prior work, resulting in a better overall coverage of dimension space. In other words, we predicted that providing an analyst with dimension coverage information that communicated what their previous collaborator "did not do" would result an analysis path more divergent from the prior work.

To test this hypothesis, we used a between-subjects design to compare two versions of a history viewer similar to Study 1: (1) a full version containing dimension, sequence, and data views (Figures 2, 4, and 5) and (2) a baseline version containing only the Sequence view (Figure 4). Similar to Study 1, the rationale behind the design of the baseline version was (1) to emulate current history tools, and (2) to control for design differences between tools, allowing us to directly assess the added value of a dimension centered perspective. The experimental setup was similar to Study 1, including nearly identical procedures and data capture. For brevity, here we only describe differences from the methods used in Study 1.

5.1 Data

Data was similar to 4.2 except that we created a new history file. We asked a business PhD student to perform the initial analysis. We asked him to intentionally neglect investigating some of the dimensions that could logically be investigated (leave room for further exploration) and also keep some of his questions at a higher level (reserve potential for drilling down). We chose to create a new history for this study to ensure that the initial analyst intentionally left out some rational avenues of data exploration involving both new dimensions and drilling in on existing dimensions.

5.2 Participants

We recruited 20 business students (12 graduate, 8 senior undergraduate, 4 male, 16 female, average age of 25). We selected business students to ensure that our participants had the necessary domain knowledge to investigate a finance-related problem. They were randomly assigned to use either the full or the baseline version. We recruited only participants who reported having a strong understanding of business data analysis and experience with tools that enable constructing statistical charts based on data (e.g., Microsoft Excel). None of the participants had used Tableau Public before. None of the participants took part in Study 1.

5.3 Apparatus

We used a PC with two side-by-side 17 inch monitors, each with 1280 x 1024 resolution. An instance of Tableau Public software was open on one monitor and depending on the condition, a full or baseline version of the prototype on the other monitor. The rationale behind this setup was to enable users to easily switch between the analysis task in Tableau and reviewing the collaborator's history using the prototype. When asked at the end of the task, none of the participants found switching between two monitors distracting. Participants were also provided with pen and blank paper for taking notes.

5.4 Task

The analysis task required the participant to continue the exploratory analysis started by their "collaborator". Following is the task given to participants: "You are a business data analyst in a large international company. You are working collaboratively with other analysts in your company to explore sales data for the past 4 years and identify any possible strong and/or poor performance. Your collaborators are at different times/locations and work completed by others is passed around to be built upon. For your own analysis, you should explore the data and try to identify any interesting/unexpected patterns in the data with respect to business performance. In order to efficiently continue your collaborator's work, you first need to review and understand the prior work passed to you. This will also help you to keep the similar work minimized and investigate different plausible performance indicators. While doing your analysis, you can review the collaborator's work if required. "

5.5 Data Analysis

In exploratory analysis, allowing collaborators to realize what work has been done will assist them in knowing where to allocate effort next [13] which implies better coordination[5][9]. We hypothesized that this knowledge would lead to a better overall coverage of the dimension space between the participant and the "collaborator". Therefore, we decided to measure the similarity between each participant's work and the initial analysis as an indicator of coordination. The more an analyst's work is different from the initial work, the greater cumulative coverage of problem space is achieved which in turn indicates better coordination.

The data dimensions included in a chart give strong clues as to the question being asked. Thus, if a chart created by the participant contains the same set of dimensions as a chart created by the initial analyst, it is likely they were investigating the same question(s). Likewise, charts containing some matching dimensions represent more similar investigative queries than charts containing completely different sets of dimensions. This observation formed the basis of our similarity analysis.

To compute similarity between a participant's analysis and the initial analysis, we first used videos and saved visualizations to identify all the unique questions that were asked by that participant. (We consider a question equivalent to a query that

returns a subset of data, e.g. what is the relationship between Sales, Profit and Region?) Then using an alias assigned to each data dimension (e.g. Sales = A, Profit = B, Region = C), we converted each question into a set of letters (e.g. relationship between Sales, Profit and Region == {A, B, C}). Questions in the initial analysis were likewise transformed into sets of letters. Jaccard's Similarity Index (1) computes the similarity between two sets. We used a modified version of Jaccard's Index to compute similarity scores (S) between each of the participant's questions and each of the initial analyst's questions.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \quad (1)$$

Our modifications to Jaccard's index took into consideration filtering of dimensions as well as exploring uninvestigated dimensions. Our rationale for these changes was (1) investigating a completely new dimension is conceptually more different than investigating a filtered version of the same dimension and (2) investigating a dimension that nobody has considered before conceptually is more different than investigating new combinations of previously explored dimensions.

Our modifications were as follows. When computing the sets' intersection, if a dimension had different filtering in the two sets (e.g. City: [all cities] in set1 and City: [LA, NY] in set2), the intersection count was increased by 0.5 instead of 1.0. As a result, the similarity score decreased. After careful consideration, we also added a heuristic rule to give weights to dimensions. Dimensions investigated by the previous analyst carried a weight of 1.0 and previously uninvestigated dimensions carried a weight of 1.5. For example, when computing the union of two sets, each dimension was multiplied by its weight as follows: if set2={A, B, C} and set1={A, F, C} where F is the only previously uninvestigated dimension, then the union would be A:(1*1.0)+B:(1*1.0)+F(1*1.5) + C(1*1.0)= 4.5. The resulting score S is a value between 0.0 (no similarity) and 1.0 (identical).

Each question a participant asked should be compared to the most similar question asked by the analyst. Therefore, for each question (i.e. set) of each participant, we computed Jaccard's index in comparison to ALL questions (sets) of the original analyst; the maximum similarity found was assigned as that set's similarity score.

5.6 Results

We use individual questions as our unit of analysis. Figure 6 shows mean similarity score by condition. Since the data was normally distributed, we performed an independent-samples two-tail t-test to check whether there was a statistically significant difference between mean values of full version (mean=0.33, SD=0.11) and baseline version (mean=0.58, SD=0.21) questions. The result ($t(339)$=9.192, p<0.0091) demonstrates that similarity scores for full version questions were significantly lower than for the baseline version.

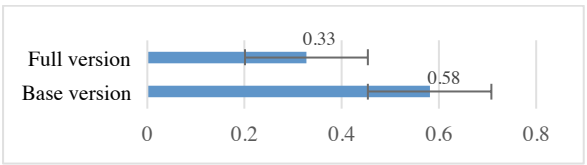

Figure 6: Average similarity scores for full and baseline version questions. Error bars show standard deviation. (1.0 = questions are identical to previous analysis; 0.0 = questions are completely different from previous analysis).

These results demonstrate that participants with access to Dimension view took analysis paths that were more divergent from the initial analysis than participants with only access to Sequence view. One contributor to this phenomenon is the number of unique uninvestigated dimensions considered by each participant. On average, full version users considered 19.6, and baseline version users considered 6.2 uninvestigated dimensions in their analyses. Results of a t-test ($t(18)$=4.98, p<0.001) showed a statistically significant difference between the groups. In addition, six out of ten full version users started their analysis by asking a question involving one or more of the uninvestigated dimensions. On the other hand, only one of the baseline version users did so. Interestingly, only full version users asked questions (total of 19) that were completely different from the questions asked in the initial analysis (i.e., S=0.0). Conversely, baseline version users asked more identical questions (i.e., S=1.0): there were 20 of these for baseline version and only 3 for full version.

To summarize the quantitative results, full version users showed a greater tendency to focus on less explored aspects of the problem while baseline version users placed more effort on drilling in on previous questions. We argue that this is due to full version users' ability to more easily discover what had been focused on and what had been left out in the initial analysis. The Sequence view alone did not make it easy to acquire this information. For baseline version users, gathering information about dimension coverage required multiple passes through items in the Sequence view, which presumably added cognitive costs.

In addition, baseline version users relied on external memory aids such as paper notes for recording their discoveries. There was a substantial difference between the number of notes taken by full version (total of 3) and baseline version users (total of 9). Though this might be a result of personal preference and work style, closer inspection of notes taken by baseline version users revealed five instances of explicitly recording information about dimension space, similar to what was available in the Dimension view for full version users. These five baseline version users manually extracted dimension coverage information by tracing the history and recording different examined combinations; they recorded this information in their notes for later use. For instance, as part of her note a participant recorded that "Profit" was considered with "Product Category", "Product Subcategory", and "Order Date"; 'Sales+Profit' was considered with "Customer Segment", "Product Category", "City", and "Product Subcategory".

We also reviewed captured videos of full version users to understand their Dimension view usage. Eight out of ten participants started their initial review through Dimension view (note that they were not instructed to do so and could use any view at any time). Statements made by participants suggested that they found Dimension view useful and intuitive for gaining an overview of the previous analysis. Following are a few examples from users' alouds that show how Dimension view helped to inform their understanding of prior work: "…[the initial analyst] didn't look at Returns…maybe I should look into this…", "…it seems that he focused greatly on Profit, Sales and Order Date…". During the analysis, users mostly referred to Dimension view to refresh their minds and avoid duplicating work. They also used Dimension view as a visual search aid. Selecting a dimension (or a combination) helped them to easily filter Sequence view. Subsequently they would look at the thumbnail view or open the full size view of the chart.

While Dimension view was clearly important, our findings show that Data view was not used as much. Only half of the full version users (5 out of 10) referred to Data view (total of 22 interactions, Avg = 2.2, SD=3.19). On the other hand, all of them used Dimension view (total of 78 interactions, Avg=8.9, SD = 2.6). Participants mainly used Data view during their initial

analysis to gain an understanding of the focus of prior work in the data space. For example, after opening the darkest region in the Customer Segment in Data view to see the data points in an external table, one participant said "I noticed that in the customer segment, [the initial analyst] was more focused on 'corporate' than the others". Other participants also noticed this trend.

6 Discussion

Our findings from the two studies clearly demonstrate that collaborative visual data analysis can benefit from the addition of dimension coverage representations of history. In Study 1, the full version of the tool enabled analysts to answer questions about the prior analysis more quickly and more accurately than the baseline version. While we cannot fully isolate the reasons for this difference with our experimental design, our qualitative observations and participants' comments strongly suggest that dimension view was the most important factor. In Study 2, full version users showed better coordination with the previous analyst through their focus on uninvestigated aspects of the problem. Our similarity analysis showed that full version users asked questions that were more different from the ones asked by the initial analyst than baseline version users. They investigated the problem from new angles, meaning that overall there was a more comprehensive investigation of the problem. For example, the initial analyst did not investigate the 'Days to Ship' dimension (i.e., days from receiving to shipping of an order). Yet, inefficient order processing times could be responsible for overhead costs and loss of Profit. Six full version users, in contrast to only two baseline version users, examined this possibility. Based on our observations, we attribute the better coordination shown by full version users primarily to the presence of Dimension view — full version users reported that Dimension view helped them to easily and accurately identify underexplored aspects of the dataset. This knowledge in turn influenced the questions that they asked.

On the other hand, we observed greater overlaps between the analysis of baseline version users and the prior analysis. This overlap represents duplicated work and a reduced overall coverage of the problem space, suggesting less well coordinated collaborative work. These users showed a greater inclination towards continuing the prior work and 'drilling down' on questions. This is most likely due to the affordances of Sequence view. At the surface level, this view contains visualizations that each represent a question. Therefore, the immediate messages conveyed by this representation are 'questions'. Gaining an understanding of dimension space coverage requires iteratively reviewing these 'questions' (or manually generating notes) to build a mental map of which dimensions had been covered. This is a rather costly and cumbersome process. Therefore, as anticipated by the principle of least cognitive effort [2], most baseline version users preferred the less costly action of 'picking a question' and drilling down on it rather than identifying the unexplored aspects of the problem and asking new questions.

We make an assumption in this work that better overall coverage of the dimension space is an important aspect of good coordination. This seems reasonable in an exploratory analysis situation where the primary focus is finding as many trends and outliers as possible. This assumption stems from prior research [9][5][13] in the asynchronous collaborative analysis field that suggests an awareness of "what has been covered" can direct current work towards "what has been left out". We believe our findings justify this assumption, in that participants reported that dimension view made it easier for them to understand the prior work, particularly the coverage of dimension space. At the same time, we do not claim that in all exploratory analysis situations broad coverage of all possible questions is the foremost goal. There may be situations in which drilling deeper on previous analysis is preferred. Nonetheless, we speculate that even drilling in on the prior work might benefit from Dimension view. While drilling in (i.e. keeping some dimensions the same while changing others and filtering), analysts can exploit Dimension view to discover what new combinations remain.

There was a substantial difference between the total number of references to Dimension (69 times) and Data (10 times) views by the participants in Study 2. Although we cannot truly isolate that value of Dimension view in Study 2 because it was in the same condition as Data view, it seems that participants found Dimension view to be much more useful. None of the full version users started their initial review of prior work by interacting with Data view. We believe the prominence of Dimension view relates to the specific task of our study, which required participants to find as many trends and outliers as possible. In this sort of exploration, gaining an understanding of 'what has been investigated' in the dimension space comes before the same understanding of the data space (i.e. data values). Yet there might be other exploratory analysis situations where the reverse is true. In [26], the exploratory analysis task involved investigating a constant set of dimensions by manipulating the filtering of values. In such a case, being able to visually understand which data values were explored versus left out would be most valuable.

7 Future Work

To investigate our research questions efficiently with minimal development effort, our prototype history tools were separate from the visual data analysis tool. However, we envision that in practice these two should be integrated. Ideally, a user should be able to access any history view on demand. Sequence view may be more useful when the user is looking for particular instances in the history and is drilling in on a previously asked question. On the other hand, Dimension view could be beneficial when the analyst is formulating new questions.

Additionally, since the primary focus of this research was to examine the speculated value of providing Dimension view, we did not fully explore the rather immense possible design space. We did attempt to make good design choices based on current perceptual knowledge and iterative development; however, future research could further explore this design space and may be able to improve upon our representations. Most notably, our representations of dimension coverage are quite space inefficient. We would like to explore a scented-widget-like [26] version that integrates this information into control widgets for creating new charts in a visualization tool.

Our protoype's design is also specific to tabular data. Extension of the data-centric history idea to other types of data that do not have this discrete tabular nature is not obvious, but the general idea might be applicable with substantial design changes. For example, for network data, interesting attributes to reveal might be network attributes that have been the focus of exploration via filters or visual encodings (e.g., color-based mappings). For a document corpus, it might be useful to highlight entities (e.g., people, places, events, or documents themselves) that have been investigated by the analyst. The Cambiera [14] system did this to some degree, but was intended for in-the-moment awareness of another person's work rather than a detailed post-hoc review.

Dimension view could also be enhanced with capabilities to distinguish among the activities of multiple users in a collaborative team, and even to consider its use during synchronous collaboration. For example, if four people are all exploring the same dataset, it could be helpful to see what aspects each person has worked on. Even if they are working simultaneously, a dimension centric view might serve as a helpful

awareness mechanism, and indicate which parts of the data are being neglected and might be worthy of exploration.

8 CONCLUSION

We examined the value of a dimension coverage perspective on visualization history for supporting asynchronous collaborative exploratory visual data analysis. Our results demonstrate that this novel and valuable perspective can facilitate coordination by helping analysts to understand past work and identify unexplored directions. Representing data analysis history from a dimension centric perspective enables analysts to answer questions such as "What dimensions were investigated, how often, and when?" and "What co-mappings of dimensions were most prevalent?". Experimental results show that incorporating Dimension view can expedite the review process, help analysts to gain a high level understanding of a collaborator's analysis strategy, and encourage them to pay more attention to underexplored or neglected aspects of the problem. Participants who had access to dimension coverage information showed significantly greater divergence from the 'collaborator' in their analysis paths as compared to participants without access to this information.

ACKNOWLEDGMENTS

We thank Veronika Irvine and Anirban Kar for their help preparing the analysis history files. This research was funded by SAP, NSERC, and GRAND.

REFERENCES

[1] Bavoil, Louis, Steven P. Callahan, Patricia J. Crossno, Juliana Freire, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. "Vistrails: Enabling interactive multiple-view visualizations." In *Visualization, 2005. VIS 05. IEEE*, pp. 135-142. IEEE, 2005.

[2] BISGAARD MUNK, Timme, and Kristian Mørk. "Folksonomy, the power law & the significance of the least effort." *Knowledge organization* 34, no. 1 (2007): 16-33.

[3] Brodlie, Ken, Andrew Poon, Helen Wright, Lesley Brankin, Greg Banecki, and Alan Gay. "GRASPARC-a problem solving environment integrating computation and visualization." In *Visualization, 1993. Visualization'93, Proceedings., IEEE Conference on*, pp. 102-109. IEEE, 1993.

[4] Callahan, Steven P., Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. "VisTrails: visualization meets data management." In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 745-747. ACM, 2006.

[5] Carroll, John M., Mary Beth Rosson, Gregorio Convertino, and Craig H. Ganoe. "Awareness and teamwork in computer-supported collaborations." *Interacting with computers* 18, no. 1 (2006): 21-46.

[6] Chen, Yang, Jamal Alsakran, Scott Barlowe, Jing Yang, and Ye Zhao. "Supporting effective common ground construction in asynchronous collaborative visual analytics." In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 101-110. IEEE, 2011.

[7] Chuah, Mei C., and Steven F. Roth. "Visualizing common ground." In *Information Visualization, 2003. IV 2003. Proceedings. Seventh International Conference on*, pp. 365-372. IEEE, 2003.

[8] Dou, Wenwen, Dong Hyun Jeong, Felesia Stukes, William Ribarsky, Heather Richter Lipford, and Remco Chang. "Recovering reasoning process from user interactions." *IEEE Computer Graphics & Applications* (2009).

[9] Dourish, Paul, and Victoria Bellotti. "Awareness and coordination in shared workspaces." In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pp. 107-114. ACM, 1992.

[10] Dunne, Cody, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. "GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1663-1672. ACM, 2012.

[11] Heer, Jeffrey, Jock Mackinlay, Chris Stolte, and Maneesh Agrawala. "Graphical histories for visualization: Supporting analysis, communication, and evaluation." *Visualization and Computer Graphics, IEEE Transactions on* 14, no. 6 (2008): 1189-1196.

[12] Heer, Jeffrey, Fernanda B. Viégas, and Martin Wattenberg. "Voyagers and voyeurs: Supporting asynchronous collaborative visualization." *Communications of the ACM* 52, no. 1 (2009): 87-97.

[13] Heer, Jeffrey, and Maneesh Agrawala. "Design considerations for collaborative visual analytics." *Information visualization* 7, no. 1 (2008): 49-62.

[14] Isenberg, Petra, and Danyel Fisher. "Collaborative Brushing and Linking for Co-located Visual Analytics of Document Collections." In *Computer Graphics Forum*, vol. 28, no. 3, pp. 1031-1038. Blackwell Publishing Ltd, 2009.

[15] Javed, Waqas, and Niklas Elmqvist. "ExPlates: spatializing interactive analysis to scaffold visual exploration." In *Computer Graphics Forum*, vol. 32, no. 3pt4, pp. 441-450. Blackwell Publishing Ltd, 2013.

[16] Kadivar, Nazanin, Victor Chen, Dustin Dunsmuir, Eric Lee, Cheryl Qian, John Dill, Christopher Shaw, and Robert Woodbury. "Capturing and supporting the analysis process." In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 131-138. IEEE, 2009.

[17] Krzywinski, Martin, Jacqueline Schein, İnanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. "Circos: an information aesthetic for comparative genomics." *Genome research* 19, no. 9 (2009): 1639-1645.

[18] Lu, Jie, Zhen Wen, Shimei Pan, and Jennifer Lai. "Analytic trails: supporting provenance, collaboration, and reuse for visual data analysis by business users." In *Human-Computer Interaction–INTERACT 2011*, pp. 256-273. Springer Berlin Heidelberg, 2011.

[19] Mahyar, N; Sarvghad, A; Tory, M. CoSpaces: Workspaces to Support Co-located Collaborative Visual Analytics. In *Workshop on Data Exploration for Interactive Surfaces DEXIS 2011*, p. 36. 2012.

[20] Munzner, T. *Visualization Analysis and Design*. CRC Press, 2014.

[21] Pirolli, Peter, and Stuart Card. "Information foraging." *Psychological review* 106, no. 4 (1999): 643.

[22] Shrinivasan, Yedendra Babu, David Gotz, and Jie Lu. "Connecting the dots in visual analysis." In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 123-130. IEEE, 2009.

[23] Viegas, Fernanda B., Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. "Manyeyes: a site for visualization at internet scale." *Visualization and Computer Graphics, IEEE Transactions on* 13, no. 6 (2007): 1121-1128.

[24] Wattenberg, M., Kriss, J. Designing for social data analysis. *IEEE Transactions on* 12, no. 4 (2006): 549-557.

[25] Willett, Wesley, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. "CommentSpace: structured support for collaborative visual analysis." In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 3131-3140. ACM, 2011.

[26] Willett, Wesley, Jeffrey Heer, and Maneesh Agrawala. "Scented widgets: Improving navigation cues with embedded visualizations." *Visualization and Computer Graphics, IEEE Transactions on* 13, no. 6 (2007): 1129-1136.