An Insight- and Task-based Methodology for Evaluating Spatiotemporal Visual Analytics

Steven R. Gomez, Hua Guo, Caroline Ziemkiewicz, and David H. Laidlaw

Abstract—We present a method for evaluating visualizations using both tasks and exploration, and demonstrate this method in a study of spatiotemporal network designs for a visual analytics system. The method is well suited for studying visual analytics applications in which users perform both targeted data searches and analyses of broader patterns. In such applications, an effective visualization design is one that helps users complete tasks accurately and efficiently, and supports hypothesis generation during open-ended exploration. To evaluate both of these aims in a single study, we developed an approach called layered insight- and task-based evaluation (LITE) that interposes several prompts for observations about the data model between sequences of predefined search tasks. We demonstrate the evaluation method in a user study of four network visualizations for spatiotemporal data in a visual analytics application. Results include findings that might have been difficult to obtain in a single experiment using a different methodology. For example, with one dataset we studied, we found that on average participants were faster on search tasks using a force-directed layout than using our other designs; at the same time, participants found this design least helpful in understanding the data. Our contributions include a novel evaluation method that combines well-defined tasks with exploration and observation, an evaluation of network visualization designs for spatiotemporal visual analytics, and guidelines for using this evaluation method.

Index Terms—Evaluation methodology, insight-based evaluation, visual analytics, network visualization, information visualization.

1 Introduction

Evaluating visual analytics systems is challenging because users need to know that the system supports both basic information retrieval tasks as well as complex reasoning and exploration. A system that is good for looking up specific data is not always good for building insights and testing hypotheses, and vice versa. At the same time, practical applications frequently demand that the same tool be used for both purposes. Despite visual analytics' focus on reasoning, many studies evaluate tools using task-based protocols that measure only user performance on low-level tasks. By contrast, insight-based methodologies aim to measure how well visualizations promote insight generation, using characteristics like the domain value of observations users make about the data model. However, these methodologies can be difficult to follow, and it is not clear how best to capture insight characteristics alongside users' task performance, as is relevant in visual analytics applications that support both targeted data searches and analysis of broader patterns.

Here we present a method for evaluating visualizations using both tasks and exploration, and demonstrate this method in a study of four spatiotemporal network designs for a visual analytics system. We call the approach *layered insight- and task-based evaluation* (LITE) because it interposes several prompts for observations about the data model between sequences of predefined search tasks. Our evaluation demonstrates the feasibility of a lightweight, within-subjects insight-based evaluation. We reflect on the relationship between users' task performance with a visualization and how well it promotes insights in assessing the best choice among four visualization designs for a spatiotemporal visual analytics system.

Our contributions here include: 1) a novel method of evaluating both task performance and insight characteristics of visualizations in a single study using a mixed design; 2) a demonstration of the method in a case study of four network-layout designs for spatiotemporal visual analytics, and 3) guidelines for using the evaluation method in future studies. While our case study focuses on a spatiotemporal visual analytics application where both exploration and routine search tasks might be performed, the evaluation method can be applied to other visualization types.

- *Steven R. Gomez is with Brown University. E-mail: steveg@cs.brown.edu.*
- *Hua Guo is with Brown University. E-mail: huag@cs.brown.edu.*
- *Caroline Ziemkiewicz is with Aptima, Inc. E-mail: cziemkiewicz@aptima.com.*
- *David H. Laidlaw is with Brown University. E-mail: dhl@cs.brown.edu.*

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

2 Related Work

Many evaluation methods have been demonstrated in empirical visualization research. Carpendale reviews evaluation approaches for information visualization [10] and describes challenges outlined in earlier works by Plaisant [22] and others. Another overview of approaches aimed at visual analytics appears in the VisMaster consortium book [15]. The biennial BELIV workshop (Beyond Time and Errors: Novel Evaluation Methods for Visualization) has significantly added to the discussion of challenges in visualization evaluation. The research contributions in its proceedings have focused on developing more effective evaluation methods that avoid the pitfalls of traditional methodologies. Taxonomies of past studies have also been helpful in constructing guidelines for evaluating new visualizations [16, 13, 20].

In the remainder of this section, we describe methods relevant to a combined insight- and task-based evaluation, as well as to evaluations of information layouts for visual analytics.

2.1 Task-based Evaluations

Controlled laboratory studies with predefined tasks are commonplace in visualization research. In general, these studies aim to produce measurable outputs that are comparable among participants, design conditions, or other independent variables. Accuracy and response time for tasks are typical measures, with accuracy sometimes being used to filter task executions from the response-time analysis (e.g., [12]). In such studies the objective is to demonstrate differences in task efficiency. The evaluation approach described here collects user efficiency and accuracy measures for tasks selected using a typology covering the basic analysis questions one might ask of a spatiotemporal data model. These tasks represent analysis pieces that could be composed into a larger-scale, exploratory analysis. We acknowledge that there are tradeoffs in the realism of tasks performed in order to gain precise, quantitative results [16]. Our study uses non-experts rather than professional data analysts, and tasks have been abstracted to remove any dependence on domain knowledge.

2.2 Insight-based Evaluations

Unlike task-based evaluation methods, insight-based methodologies are motivated by the realization that the goal of a visualization tool is

usually to enhance understanding of the underlying data, not to improve task accuracy and efficiency [9, 17, 23]. Saraiya et al. presented an insight-based approach for evaluating bioinformatics tools [26] and later used it in a longitudinal study where insights were developed over months [27]. Characteristics of insights include the number of distinct data observations, the time needed to reach each insight, the domain value of each insight, breadth-versus-depth labeling, and other characteristics. Quantifying some of these attributes requires domain experts to participate as response coders in the evaluation. Even with this scheme, eliminating all subjectivity from the evaluation is difficult; for instance, the cutoff between a depth insight and a breadth insight might vary depending on the expert coder.

Other studies have applied similar methods to measure insight characteristics between visualization conditions. It is worth noting that insight characteristics have been adapted from those proposed by Saraiya et al. in order to fit the hypotheses of other studies. For instance, O'Brien et al. made an insight-based evaluation of two tools for visualizing genomic rearrangements using a reduced set of insight characteristics [19]: researchers counted the instances of three categories of insights as well as the total number of insights, total "hypothesis-driving" insights, and the insights per minute of analysis. Our method also uses a simplified set of insight characteristics and collects these with a single study protocol alongside task performance.

North et al. found that the results of an insight-based evaluation can both support and contradict findings of studies using benchmark tasks with the same visualizations [18]. It is possible that evaluators who use only one of these methods will miss details visible using the other. We aim to combine the two in a single, practical protocol while minimizing interactions or biases in the results. Our method differs in time scale from longitudinal studies in visualization, such as multidimensional in-depth long-term case studies (MILCS) [30]. Unlike previous insight-based evaluations, the evaluation we present uses non-expert participants. Using non-experts lets us achieve a larger sample size than would otherwise be possible, enabling us to test hypotheses about task performance and quantified insight characteristics more precisely. There are drawbacks in using non-experts; e.g., asking participants for initial analysis questions might be unreliable; however, even if domain experts were used, they would not necessarily have experience with the analysis tools in the study, as in [26]. Furthermore, we expect that combining tasks with exploration provides extra training and motivation for participants. Previous studies [11] and models [25, 24] have demonstrated how predefined tasks enhance exploratory learning of computer interfaces. While the insights themselves are likely to be less deep for non-experts than for domain experts, it is possible to compare insight-promoting characteristics between visualizations using non-experts.

2.3 Spatiotemporal Tasks and Visual Designs

An indispensable part of designing a visual analytics tool is considering the set of analytical tasks to be supported. The visualizations evaluated here are grounded in previous work on visual analysis of spatiotemporal data. In [21], Peuquet distinguished three components in spatiotemporal data and queries about those components: space (*where*), time (*when*), and objects (*what*). Users can complete queries when two of the three components are known and the other is the search target. Andrienko et al., drawing on Peuquet's work as well as other task typologies, proposed a typology for visual analytical tasks with the dimensions of search target, search level, and cognitive operation [4]. Others [6, 2, 28] have proposed more general task typologies that also apply to spatiotemporal data.

Many visual analytics designs for spatiotemporal data exist, as reviewed comprehensively in [4]. Notably, maps and timelines, the most common representations for spatial and temporal data, have been combined in previous design studies. Slingsby et al. showed that these representations can be configured as levels of a tree map in order to support different queries [31]. More recently, Andrienko and Andrienko proposed the cartographic map display and time-series display as the two visualization components in their visual analytics framework for spatiotemporal analysis [3].

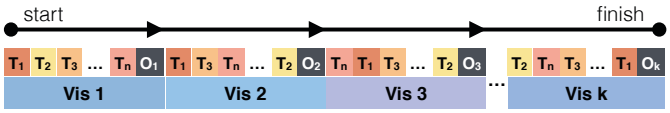


Fig. 1: Example ordering of k visualization conditions and n task types in LITE. After each block of tasks with a visualization (labeled $T_1 \ldots T_n$), the participant is prompted for exploration and observation about the data (labeled $O_1 \ldots O_k$). Task ordering within a visualization condition is randomized using a balanced Latin square, and visualization orders are randomized between participants using a balanced Latin square. In our case study, $k = 4$ and $n = 4$.

3 Layered Insight- and Task-based Evaluation

We propose combining a lightweight insight-based evaluation adapted from Saraiya et al. [26] with a traditional task-based evaluation. We call this approach **l**ayered **i**nsight- and **t**ask-based **e**valuation, or LITE, because it interposes several prompts for observations about the data model between sequences of predefined search tasks or queries.

3.1 Motivation

Two main goals for this method are: 1) to measure the accuracy and efficiency of common tasks alongside insight characteristics without compromising task measurements; and 2) to measure insight characteristics while sidestepping some of the difficulties of performing the insight-based method, such as:

- **D1** Users must be intrinsically motivated to look for insights during a session that might be open-ended.
- **D2** Training new users on visualization interfaces can be challenging. Training can fatigue users and make them try less hard in the actual study [26].
- **D3** After the user study, coding observations for measurable insight characteristics like domain value is difficult and requires domain experts.

Even when these difficulties are managed in an insight-based evaluation, challenges arise when performing such an evaluation separately from a task-based evaluation so as to collect measures of both task performance and insight generation. If these studies use different participants it can be difficult to draw conclusions about relationships between tasks and exploration. Individual differences or differing sample sizes must be considered.

Performing separate task- and insight-based evaluations back to back creates other challenges. If a full insight-based evaluation is performed before a task-based evaluation, open-ended exploration may fatigue users to the point that they perform poorly on the follow-up study. If a full task-based evaluation is performed before an insight-based evaluation, users may have less motivation to explore the data model: they might satisfice and report only shallow insights in order to finish the study.

3.2 Steps

The initial stages in a LITE evaluation are similar to those in previous insight-based methodologies. As a study session proceeds, sets of predefined tasks are interleaved with exploration periods letting participants find and record insights. The steps are:

1. *Background about the dataset* is provided, then participants are prompted for *initial analysis questions*. Alternatively, initial analysis questions can be provided by the evaluators.
2. Participants are then *trained on each task type* for different visualization conditions. Participants are not trained on exploration, as in [26].
3. When the study begins, participants complete *blocks of tasks* with each visualization condition.

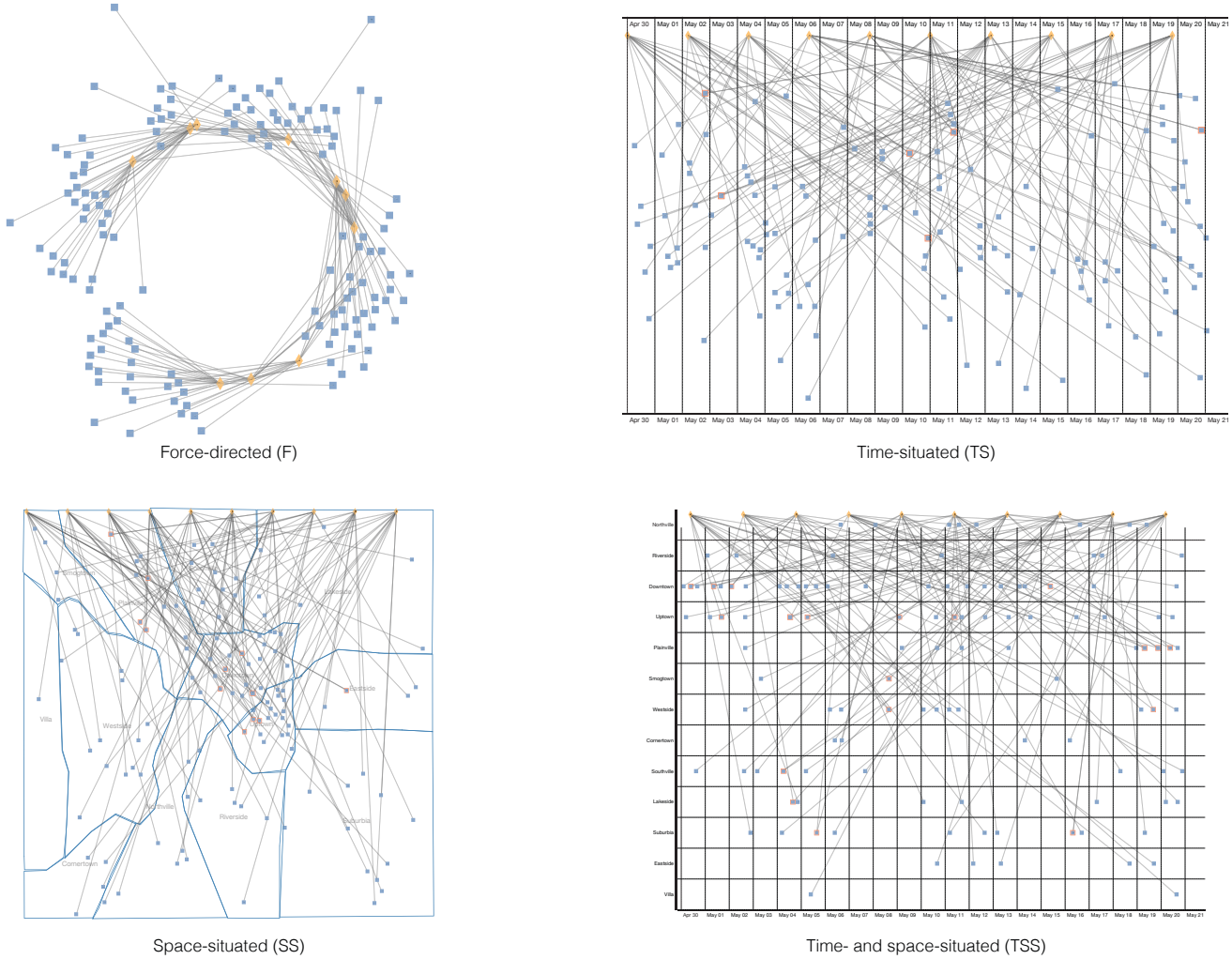


Fig. 2: Four visualization designs were evaluated using a layered insight- and task-based evaluation: *force-directed* (F), *time-situated* (TS), *space-situated* (SS), and *time- and space-situated* (TSS). These visualizations depict microblog messages and their authors, and the designs differ in how attributes of the nodes, like timestamp and location, are used to lay out the diagram.

4. After each block, participants *explore the data* freely using the visualization and *record insights*. Each exploration period is open-ended. In order to keep participants from skipping these periods, a minimum time requirement may be enforced before they can move to the next visualization and block of tasks. Figure 1 shows an example ordering of tasks and visualization conditions in which each participant completes each task type once using each visualization.
5. Finally, a *post-test questionnaire or interview* may be used after all tasks and exploration periods are finished. Subjective feedback about the insightfulness of visualizations may be used to explore findings from insight characteristics measured during exploration periods.

The proposed method addresses some difficulties of the traditional insight-based evaluation listed earlier. Study participants in LITE may feel more motivated because the session makes concrete progress through task completions rather than asking for open-ended exploration alone (D1). Tasks may improve participants' confidence with the visualizations and provide extra experience that promotes exploration and insight generation (D2). In our case study, we developed and used a scoring system without domain experts to code the value of insights (D3), but this system is not specific to LITE and could be applied to other insight-based methods.

4 Case Study

We evaluated four node-link diagram layout designs for an interactive visual analytics system that uses a graph-based model of real-world entities, like documents and people. We chose node-link diagrams here because of their flexibility in representing arbitrary node and edge types in the model. That said, we expect most nodes to have spatiotemporal attributes that describe when and where events happen. Based on this, we developed designs that differ in how location and time attributes are used to lay out nodes with these attributes in the diagram. Specifically, we looked at ways to project location and time attributes onto the drawing-plane axes. This is conceptually similar to previous work in which generic quantitative attributes are mapped onto axes to guide node placement [7]. In this study, we restricted ourselves to designing a layout for a single display. We considered four distinct node-link diagram layouts for the network model:

F Force-directed: A force-directed layout plots marks based on a physical simulation and has the effect of reducing visual density in the node-link diagram. Force-directed layouts are widely used and well understood. We consider this a control condition in an evaluation of visualization designs that position nodes using spatial or temporal attributes.

SS Space-situated: The space-situated layout overlays document

marks on a map of the city based on documents' geotags. Nodes without geotags are placed at the top of the visualization and distributed evenly.

TS Time-situated: The time-situated layout aligns document marks with a horizontal timeline. The vertical positions of document marks are determined using a force-directed layout to reduce visual density in the diagram. Nodes without timestamps are placed at the top of the visualization and distributed evenly.

TSS Time- and space-situated: The time- and space-situated layout plots document marks according to both geotags and timestamps. Nodes without geotags and timestamps are placed at the top of the visualization and distributed evenly. In TSS the horizontal axis is a timeline, as in TS. In our prototype, the vertical axis is divided into categories corresponding to neighborhoods in the data model. Categories on the vertical axis can be ordered in different ways, for instance from top to bottom based on an ordering of neighborhood locations from northernmost to southernmost. In this case, boundaries between categories could reflect some information about the geographic boundaries between neighborhoods.

Figure 2 shows each of these layout designs. All visualizations were prototyped using D3 [5] and JavaScript, and share some visual encodings. The entity type of each node is double-encoded by shape and color. Marks representing documents are blue squares and marks representing people are gold diamonds; these two sorts of marks have roughly the same size in the browser. A detailed description of each node appears in a scrollable tooltip when the user hovers over the node. For documents, this description includes the author, timestamp, location, and content. In general, document content is limited to 140 characters, since documents in our data model are microblog formats like Twitter messages that enforce a content-length limit.

A simple aggregation scheme is built into each prototype so that node marks that would otherwise overlap cannot become inaccessible to the user. When marks of the same entity type overlap, both are removed from the diagram and a single aggregated mark is added. Only marks representing entities of the same type can be aggregated: thus, documents can be aggregated only with other documents. Aggregated marks retain the same entity-type encoding (shape and color) but are distinguished by a red border and increased size. Because multiple marks might overlap, the size of aggregated marks is used to encode the number of individual entities it represents.

We considered several approaches to aggregating nodes in node-link diagrams. A common approach is to aggregate a primary entity node and nodes representing its attributes into a compound node [29, 8]. This approach does not work for our case, however, as the mapping between two types of entities in our data model might be many-to-many. In our prototypes, when a node is aggregated into a different mark, each edge mark connected to that node is replaced by another that is connected to the aggregated node. The underlying data model is not changed by this process. Two nodes connected by an edge cannot be aggregated together.

Hovering over a node mark highlights all edges connected to that entity. For example, hovering over a person node highlights edges to all document nodes connected to that person by an "authored-by" relationship. Hovering over a document highlights the edge to its author node. Highlighting is implemented by restyling edges from transparent gray to opaque red. A selection interaction is also included to allow persistent highlighting during user exploration. Users can toggle selection on node marks by clicking them with the cursor.

5 Experiment Design

After a small pilot study, we performed an experiment to evaluate a set of hypotheses about task performance and insight characteristics for participants using four visualization designs. A $2 \times (4 \times 4)$ mixed design was used to examine the independent variables of dataset size between subjects, and visualization design and task type within subjects.

5.1 Hypotheses

In general, we expect that layouts that position nodes by projecting their attributes onto the axes will improve task performance and promote insight generation. Below are specific hypotheses about the effect of independent variables on task performance (H1, H2), subjective ratings from participants (H3, H4, H5), and insight characteristics (H6–H10):

H1 For all tasks, participants will be fastest using TSS, which uses both spatial and temporal attributes to lay out nodes. For all tasks, participants will be the slowest using F.
H2 Visualization type will have a significant effect on task accuracy.
H3 Participants will report feeling most confident in their task responses when using TSS and least confident when using F.
H4 Participants will report that TSS is the most helpful visualization type for understanding the data and that F is the least insightful in this way.
H5 Participants will report that TSS is the easiest visualization type to use and that F is the hardest.
H6 Total domain value for observation prompts will be highest for the TSS condition and least for the F condition.
H7 Visualization type will have an effect on the total domain value during observation prompts.
H8 Dataset size will have an effect on both total time and total domain value during observation prompts. Both characteristics will be higher in the large dataset than in the small one.
H9 The order of observation prompts will have an effect on the total domain value during those prompts.
H10 The order of observation prompts will have an effect on the total response time during those prompts.

5.2 Visualization Types

The four visualization types in our study are described in Section 4 and shown in Figure 2. In addition to the visualization layouts, the user interface included controls to filter document nodes by publication time and location. Data-filter controls are common in visual analytics applications, and it is important that the test interface match realistic usage scenarios. The time filter is a slider that can be moved on both ends in increments of one day. Node and edge marks related to documents published outside the chosen range are invisible. The location filter contains checkboxes that correspond to all neighborhood locations in the data model and can be toggled to filter marks related to documents published outside selected neighborhoods. This filter also provides "Select all" and "Deselect all" interactions.

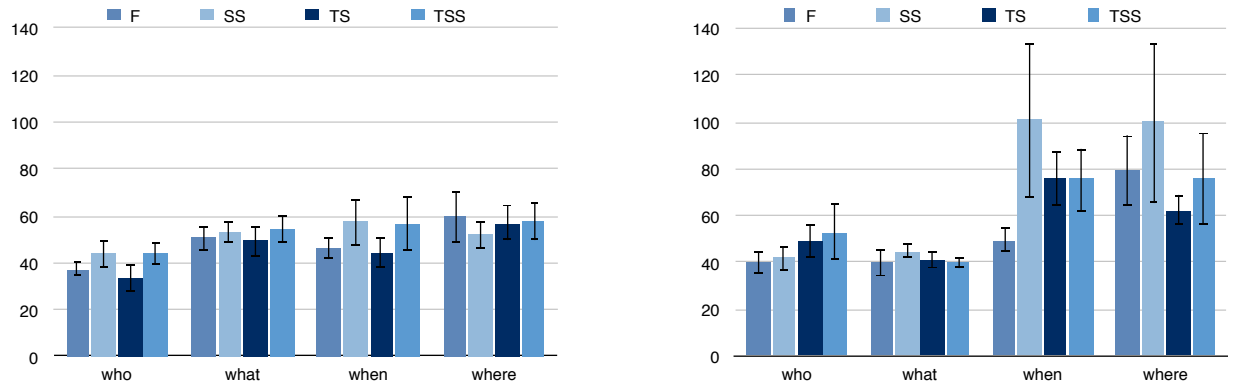
5.3 Datasets

Dataset size is an important consideration in designing network visualizations. In general, larger data models add complexity and visual density that can expose scalability problems in different designs. For our experiment, two graph-based datasets of different sizes were compiled using data from the 2011 VAST Challenge Mini-Challenge #1 (MC1) [1]. Both are subsets of a synthetic dataset containing timestamped, geotagged microblog messages from residents in a city experiencing a health epidemic.

- Small – includes 10 person nodes and 139 document nodes. There are 139 "authored by" edges that connect documents to their authors. Documents were published from 13 different neighborhoods over a span of 22 days.

- Large – includes 74 person nodes and 999 document nodes. There are 999 "authored by" edges that connect documents to their authors. Documents were published from 13 different neighborhoods, and some lacked a neighborhood-specific location (i.e., location is "Vastopolis", the city name). They were published over a span of 22 days.

Both datasets were created by sampling the Challenges full-size dataset, and both contain evidence of the health epidemic in the microblogs. These 'evidence' microblogs appear in similar proportions



(a) Small dataset

(b) Large dataset

Fig. 3: Response times grouped by task type for each visualization type. Each participant completed each task type with each visualization type. Columns show the mean of total time spent (sec) across participants (n=12 in both (a) and (b) dataset groups) and error bars show ±1 standard error. Response times corresponding to incorrect task answers are not shown.

in both datasets. We note that, while larger data models are common in real analysis scenarios, we limited the size in order to keep tasks and exploration manageable for non-expert participants during single study sessions.

5.4 Tasks

Based on the spatiotemporal network data model, tasks were selected using a simple typology based on *when*, *where*, *what*, and *who* queries. This task typology is similar to ones used in previous studies [4, 28]. We note that in the training and task instructions, the word "tweet" was used as a colloquialism for a microblog message. No data or services from Twitter were used in the study. The four task types are:

- *who* + *when* + *where* → *what*: Given a microblog's author, date, and location, summarize the content in a few words. For example, *"Cara Guthrie published a tweet in Plainville on May 20. Summarize the content of that tweet in a few words."*
- *what* + *who* + *when* → *where*: Given a brief summary of the microblog's content, author, and date, find where it was published. For example, *"Angela Barnett published a tweet about stylish watches on May 5. Where was that tweet published?"*
- *where* + *what* + *who* → *when*: Given a microblog's location, a summary of its content, and its author, find when it was published. For example, *"Bradley Church published a tweet about loss of appetite in Plainville. When was that tweet published?"*
- *when* + *where* + *what* → *who*: Given a microblog's date, location, and a summary of its content, find its author. For example, *"Someone published a tweet about Sham Wow in Uptown on May 11. Who is the author of that tweet?"*

An answer key for all task instances was created in order to score responses as accurate or inaccurate.

5.4.1 Prompts for Exploration and Observation

After each block of tasks, participants were prompted to explore the data using the visualization and record observations relevant to the epidemic in the data model. The instructions are:

> Explore the data using the visualization, then write down your observations about the data below. You should record observations about the data that are relevant to the following questions: "Do you find evidence in the data of an outbreak?"; "If so, when and where do you think it started? And how might the infection be transmitted, and is it contained?" Please number each observation.

These specific questions were taken from the instructions for MC1 [1]; they are the questions MC1 participants were asked to answer by exploring and observing a superset of the data we used. We provided these as replacements for the initial analysis questions asked as part of the insight-based methodology [26].

In response to findings from our pilot study, we added a *minimum time* for the observation prompt before each participant could move ahead to the next block of visualization tasks. During this time, participants could not access the "Next" button. When an onscreen timer showing the amount of time remaining (sec) reached 0, the "Next" button became available. At that point, participants could either continue exploring and making observations about the data or move onto the next block of tasks.

5.5 Participants

We recruited 24 participants for the study, 10 men and 14 women. Participants were primarily graduate and undergraduate students whose ages ranged from 19 to 30 years (M=24.4, SD=2.6). We assigned participants randomly to the small and large dataset groups so that each had 12 people. Participant prior experience with node-link diagrams was similar in both groups. In follow-up questions after the experiment, about half the participants in each group (5 out of 12 in the small dataset and 7 out of 12 in the large dataset) responded that they 'somewhat agree' to 'strongly agree' with the statement "I have experience using visualizations of nodes and edges," using a 7-point Likert scale. The remaining participants responded that they 'somewhat disagree' to 'strongly disagree' with that statement. No participants gave a neutral response.

5.6 Protocol

Participants were given background information about the data model and were trained for approximately 20 minutes on the four visualization designs, including the time and location filter controls in the user interface. During this training, participants performed practice trials for each task type. With the informed consent of participants, all tasks and exploration following the training were video-recorded for later analysis.

Each participant in the study performed four blocks of tasks, one per visualization. Each block contained one instance of each of the four task types. Participants performed different task instances between blocks. For each task, responses were recorded and timed for later analysis. At the end of each task block, participants explored the data using the visualization for at least three minutes and recorded insights by typing into an on-screen text field. In total, each participant performed 16 tasks and four observation prompts. This part of the study session lasted 40–60 minutes on average. Figure 1 shows an example workflow for this part of the study. Ordering effects for both

visualization types and task types are mitigated by counterbalancing. The order of visualization types is chosen between participants using a balanced Latin square, as is the order of task types within each visualization block for each person.

Participants were asked in a post-test questionnaire to report their preferred visualization type for each of the four task types. They were also asked to rate each visualization type for *ease of use*, *confidence* in task responses, and how well the visualization helped them *understand* what is happening in the data model. Ratings were on a 7-point Likert scale from "strongly disagree" to "strongly agree" for statements corresponding to these properties.

5.6.1 Insight Characteristics

Two insight characteristics were measured during each observation prompt in the study: total time spent and total domain value of observations. Total time spent had a lower bound because of the three-minute minimum time before participants could move to the next task block, as described in Section 5.4.1.

Scoring Domain Value We developed a simple scoring system to assess the domain value of individual observations. From a four-user pilot study, we identified two main parts of each observation about the data model: a *general claim* about the data (e.g., "It looks like the outbreak started in Downtown"), and 0 or more specific data points that are *evidence* for the observation (e.g., "John Doe tweeted about feeling sick – from Downtown on April 19"). In the scoring system, each recorded observation has a starting score based on whether or not it makes a new claim that was not previously reported by the user during an earlier observation prompt. Because participants explore the same data model repeatedly, it is important not to double-count observations that were arrived at earlier. For our purposes, a claim is a general hypothesis, question, or remark about the data model that is potentially synthesized from multiple observations.

On top of the starting score, points are added to observations that include specific references to data points in the model as evidence for the claim. The total points awarded during an observation prompt is equal to the sum of scores of individual observations i in the set of observations I:

$$base(i) = \begin{cases} 0 & \text{if } i \text{ makes no new claim} \\ 2 & \text{if } i \text{ makes new claim} \end{cases} \quad (1)$$

$$bonus(i) = \begin{cases} +0 & \text{if } i \text{ includes no new, supporting data points} \\ +1 & \text{if 1 new, supporting data point in } i \\ +n & \text{if } n \text{ new, supporting data points in } i \end{cases} \quad (2)$$

$$score(i) = base(i) + bonus(i) \quad (3)$$

$$total(I) = \sum_{i \in I} score(i) \quad (4)$$

In this system, we expect individual observations to range from 2 (e.g., a new claim provided without details) to 5 points (e.g., a new claim with a few supporting data points). Previous insight-based evaluations scored domain values for individual insights in a similar range and also awarded points to insights based on depth [19, 26].

Two authors of this paper independently coded all insights from the experiment using this system. Both coders were doctoral candidates studying visualization and had experience with the datasets and visualization designs. Scores for the total domain value of each observation prompt from both coders were averaged for later analyses.

6 Results

All statistical tests described in this section were performed using SPSS. The results include support for some hypotheses from Section 5.1 but not others: we accept H4, H9, and H10; we find partial support for H3, H5, and H8; and we reject H1, H2, H6, and H7.

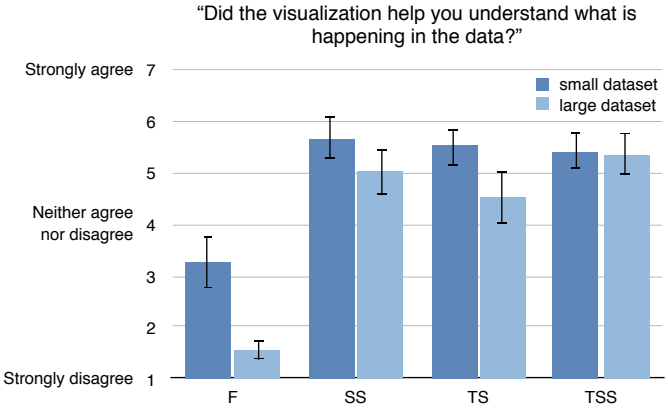


Fig. 4: Subjective ratings of visualization insightfulness on a 7-point Likert scale collected on the follow-up questionnaire. Columns show the mean response (n=12 in both groups) and error bars show ±1 standard error.

6.1 Task Performance

Overall, participants were very accurate during the study: accuracy across all participants and tasks is 96% and did not differ significantly between visualization types. Therefore, we reject hypothesis H2.

We used a mixed ANOVA to analyze how the response time varied across visualization types and tasks. Average response times for all task and visualization types are shown in Figure 3. We performed the ANOVA analysis on the log-transformed time data, as is typical in response-time analysis. Times corresponding to incorrect task answers were replaced with the mean response time for all correct responses under the same condition. Otherwise, the repeated measures analysis would exclude data from correct tasks by participants who gave one or more incorrect answer.

The results showed that task type had a main effect on response time ($p < .001$, $F_{3,50.743} = 13.109$, with Greenhouse-Geisser correction). Pairwise comparisons were made using Bonferroni-corrected p-values by SPSS. These comparisons showed that participants were significantly faster on the *who* task than on the *when* ($p < .001$) and *where* ($p < .001$) tasks. Participants were also significantly faster on the *what* than on the *where* task ($p = .025$).

We did not find support for hypothesis H1 and reject it. In fact, as shown in Figure 3, we found that the mean response time using TSS is greater than the mean response time using F for most task types in both the small and large dataset size conditions. We did not observe a main effect of visualization type on response time ($p = .147$, $F_{3,66} = 1.848$) or an effect of dataset size on response time ($p = .179$, $F_{1,22} = 1.931$).

6.2 Insight Characteristics

Insight characteristics measured during the study are shown in Figure 6 and Figure 7. We first analyzed the insight scores together with time spent on each insight task (using a log-transformation on times) with a multivariate mixed ANOVA with visualization type as the within-subject independent variable. The results showed that visualization type did not have a main effect on either the insight value score or the exploration time. We did not find evidence for H6 or H7 and reject both. We found partial support for H8: dataset size had a main effect on time ($p = .041$, $F_{1,22} = 4.702$), but not on the total domain values of insights ($F_{1,22} = 0.092$, n.s.). There was also an interaction effect between visualization type and dataset size on the total domain values ($p = .035$, $F_{3,66} = 3.031$) but not on time ($F_{3,66} = 0.347$, n.s.).

We then performed a similar analysis with presentation order of the visualizations as the independent variable. This time we observed a strong main effect of presentation order on both the total domain value scores of insights ($p < .001$, $F_{3,66} = 7.488$) and the exploration time ($p < .001$, $F_{3,38.256} = 11.621$, with Greenhouse-Geisser correction). We

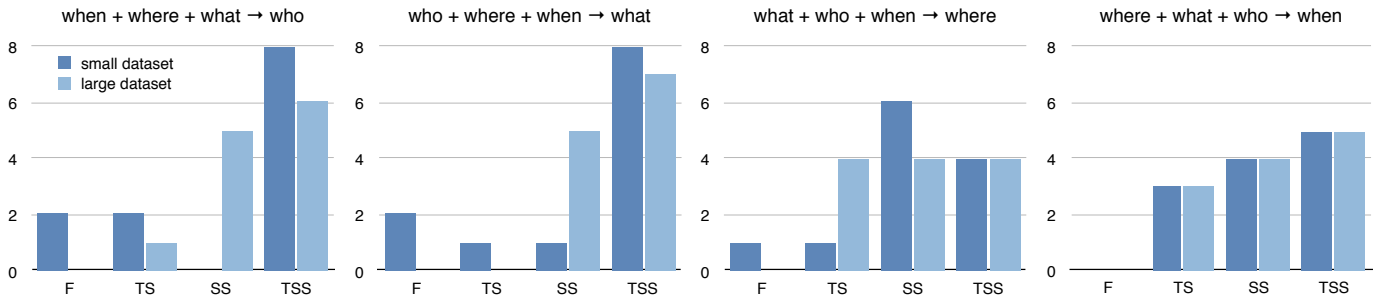


Fig. 5: Preferences for visualization type based on task type. From left to right, the tasks shown are *who*, *what*, *where*, and *when* queries. Columns show the number of participants (n=12 in both groups) who preferred each visualization for the task.

thus found support for H9 and H10. Participants spent significantly more time on the visualization that was presented first than on the following three (p = .033, p = .011, and p = .002 respectively), and also spent more time on the second visualization than on the last one (p = .02). Participants also had higher insight scores on the first visualization than on the third ($p < .001$) or the last (p = .005) visualization.

6.3 Subjective Ratings

Figure 5 shows the numbers of participants who preferred each visualization type for each task type. No participants who interacted with the large dataset preferred the force visualization for any task. TSS was preferred by more participants than any other visualization for both datasets, except on the *where* task. In that task, participants using the small dataset preferred SS more than TSS, and participants using the large dataset preferred TS, SS, and TSS in equal numbers.

We analyzed the subjective Likert-scale ratings of the four visualizations using a multivariate mixed ANOVA. Visualization type had strong main effects on all three measures (*understanding*: $p < .001$, $F_{3,51} = 18.374$; *ease of use*: $p < .001$, $F_{3,51} = 9.117$; *confidence*: $p < .001$, $F_{3,38.955} = 10.386$, with Greenhouse-Geisser correction). Dataset size had a main effect on *understanding* (p = .049, $F_{1,17} = 4.512$) and *ease of use* (p = .014, $F_{1,17} = 7.557$), but not on *confidence* ($F_{1,17} = 0.705$, n.s.).

Pairwise comparisons of the visualization types showed that participants found the force visualization the least useful in helping them understand the dataset; on average F was rated significantly lower than TS, SS, and TSS ($p < .001$ in all cases). TSS was rated as the most helpful for understanding the data, although it was only significantly higher than F. Thus, we find support for H4. F was rated as the most difficult to use (lower than TS, p = .003; lower than SS, p = .004; lower than TSS, p = .013). Participants rated SS easiest (not significantly higher than TS or TSS). Therefore, we found partial support for H5. Participants also felt the least confident with the F (lower than TS, p = .006; lower than SS, p = .001; lower than TSS, p = .007). They were most confident with TS (not significantly higher than SS or TSS). Therefore, we found partial support for H3. Pairwise comparisons for the two dataset sizes showed that participants generally felt that they had a better understanding of the small dataset and also found the visualizations easier to use with the smaller dataset.

6.4 What Is the Best Design?

We expected that visualization designs using spatiotemporal attributes of nodes in the layout (SS, TS, and TSS) would have better task performance than F, but this was not the case. A possible explanation is that the process of using node positions along with guide marks on axes (e.g., in TS and TSS) to solve search tasks is less efficient than using the data filters for time and location. In fact, the features of these spatiotemporal layouts might have distracted participants from using filters as much as they did in the F layout. Task-execution videos showed that most participants used filtering often, even with the spatiotemporal layouts, so other factors may be involved. For instance, participants might have taken extra time to verify their answers using guide marks, and tasks in our typology might have been easy enough that this verification step added time without significantly improving accuracy.

The efficiency of filtering might also account for the significant differences in average response time between task types. Overall, participants were faster on *who* and *what* tasks, which gave both location and time components in the task description. In these tasks, participants can use both location and time filters before inspecting any nodes in the visualization. In the other tasks, participants had only enough information to use one filter – location or time – based on the task description.

Looking at task performance alongside user feedback, it is difficult to choose a best layout for the data model studied. The same layout with the fastest overall task performance (F) was also the one that participants felt least confident with overall and found the hardest to use overall. F was rated significantly less helpful in understanding the data than the other types. In such cases, a visual analytics designer must choose a layout by weighing competing objectives for the tool, including efficient task performance and subjective user preferences that might impact adoption rates and indicate insightfulness. When task efficiency is prioritized, F is a good layout choice in a visual analytics system with interactive, spatiotemporal data filtering. If we prioritize user preferences and subjective feedback about usability and insightfulness, SS or TSS might be a better layout.

7 Discussion

Here we discuss what we learned about LITE through our case study and present open challenges and guidelines for using the methodology.

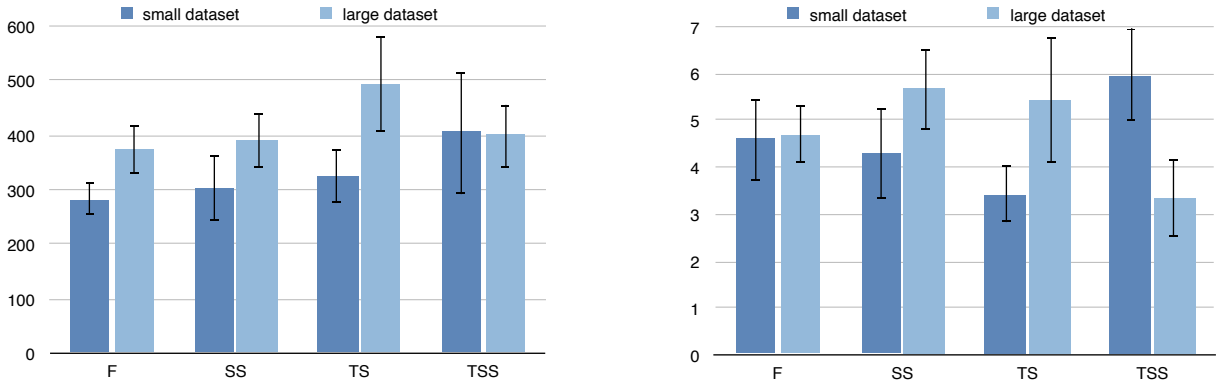
7.1 Limitations

We set out to develop a practical visualization evaluation method that combines components of task-based and insight-based evaluations. In doing so, we attempted to explore and mitigate the interactions or biases that North et al. warn about when combining these approaches [18]. Other limitations exist as well.

A practical consideration in most user studies is the time needed to run each participant, and LITE – like insight-based methods – has an open-ended exploration component that makes it difficult to estimate how long a single participant will take. In our case study, sessions lasted from 30 to 90 minutes. This uncertainty must also be considered when designing the tasks and repetitions in the task-based portion of LITE. Conducting a pilot study is a reasonable way to discover whether the task portion is feasible alongside the insight component. LITE studies with many tasks or visualization conditions might be prohibitively lengthy for participants.

A second limitation that follows from splitting time between tasks and exploration is related to the power of the results. The task-based portion of a LITE study design might have fewer trials than a dedicated task-based study design. Therefore, hypotheses could exist about task performance that can be tested in a task-only study but not in a LITE study.

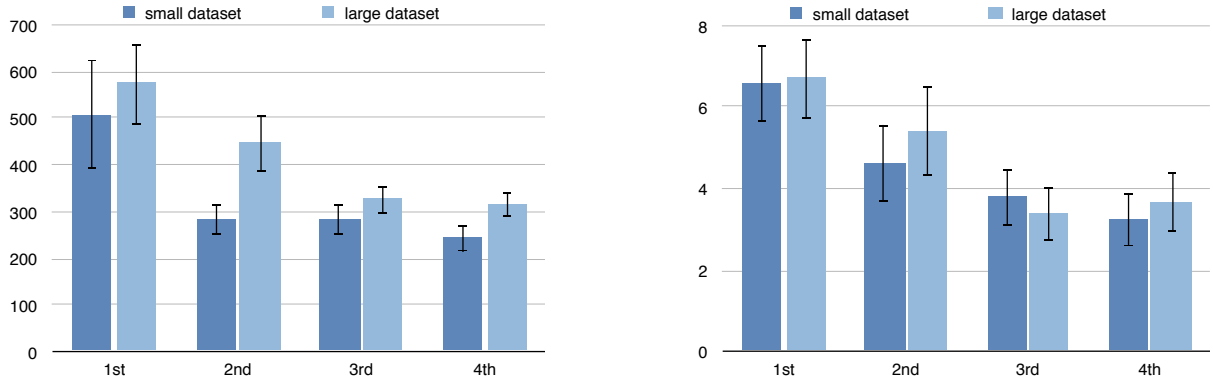
Third, participants in a LITE study alternate between blocks of tasks and exploration, and that context-switching might negatively impact how people perform these activities. On the other hand, it is also



(a) Total time spent per visualization type

(b) Total domain value of observations per visualization type

Fig. 6: Insight characteristics organized by the visualization type given to participants, each of whom was prompted for observations once per visualization type. The orderings of visualization types were counterbalanced across participants. Columns show the mean of total time spent (sec) (a) and the mean of total domain value (b) across participants (n=12 in both groups) and error bars show ± 1 standard error.



(a) Total time spent per prompt

(b) Total domain value of observations per prompt

Fig. 7: Insight characteristics organized by the order in which observation prompts were given to participants, each of whom was prompted once per visualization type. The orderings of visualization types were counterbalanced across participants. Columns show the mean of total time spent (sec) (a) and the mean of total domain value (b) across participants (n=12 in both groups) and error bars show ± 1 standard error.

possible that these switches keep participants engaged and give them a sense of making concrete progress, as mentioned in Section 3. Further study is needed to understand how these context switches affect analysis behaviors with visualizations.

Having evaluations of both insight characteristics and task performance is useful for the visual analytics application in our case study; the tool is intended both to promote insights about events and support routine data queries. Other visualizations might be aimed at only one of these purposes, and would be better evaluated using either benchmark tasks or an insight-based evaluation. Evaluators with both aims could opt to run separate studies with those methods, which is more time-consuming than running a single LITE study but might give more powerful results. These tradeoffs should be considered carefully.

7.2 Lessons from the Case Study

We encountered a variety of choices and challenges during our study that suggest guidelines for using the method.

7.2.1 Reinforcing Instructions for Different Portions of LITE

Some participants either did not understand the instructions or forgot background information on the data provided during the training period. For instance, one participant commented during her fourth observation prompt that the outbreak "Seems more over the place this time", even though participants were told that they would explore the same data set multiple times using different visualizations. This detail could be easy to forget since the network layouts changed between blocks of tasks. Participants who investigated the small dataset made no such observations, possibly because they were able to revisit and recognize microblogs between visualization conditions.

Other participants answered the initial questions given in the prompt directly rather than providing observations about the data that confirm or disconfirm those questions. For instance, some participants began their list of observations like "1. Yes. 2. ...". In a few cases, observations appeared to be numbered according to the three questions in the prompt (i.e., observations specifically for those questions, with no more than three separate observations) rather than being numbered by separate insights about any of the initial questions. We interpreted comments like "yes" as a belief that the corresponding initial question was true.

Guideline *Be explicit about how participants should record insights. Since participants switch between different types of responses during the task and insight portions of the study, these instructions should be reinforced.*

7.2.2 Coder Agreement for Insights

Overall, the two coders were fairly consistent in applying the scoring scheme to assess the domain value of insights for each prompt; their scores were within 2 points of each other for 81 out of 96 prompts, or 84.4% of the time. The coding scores are positively correlated, with Pearson's $r = 0.87$. That said, the coders agreed exactly on a score only in 36 of the 96 prompts, or 37.5% of the time. Evaluating the scoring system in future studies could help improve the scoring rules

and coder manual and therefore improve consistency in assessing the domain-value insight characteristic. As far as we know, coder consistency has not been explored in depth in the literature for insight-based evaluations. In some cases, it is unclear whether multiple coders were used to assign domain values, how well they agreed, how coding conflicts were resolved, and what expertise the coders had. We believe that the practice of reporting details of the coding process will generally benefit the development and standardization of insight-based evaluation methodologies.

Guideline *In the results of a LITE or insight-based evaluation, provide information about the process of coding the domain value of insights.*

7.2.3 Reduced Set of Insight Characteristics

Our study measured a subset of insight characteristics adapted from previous studies [26, 19]. Some insight characteristics are difficult to measure using LITE. For instance, we did not measure the time needed to reach each insight, which could be misleading in a within-subjects design that lets participants analyze the same data model over multiple iterations. Instead, the total time for exploration in each visualization condition was used, as in [19].

We also found it difficult to count the number of individual insights without using a think-aloud protocol. Our LITE case study used an on-screen text field that lets participants record observations in a manner similar to recording task responses. We relied on participants to input observations as a numbered list, but participants had different styles for doing this. One alternative is to guide users in constructing insights and evidence through a user-interface feature. For instance, Jianu and Laidlaw let users click nodes in a protein-signaling visualization to construct visual hypotheses about potential pathways, rather than having them provide unstructured text input [14]. Another possible solution that we did not test formally is using a think-aloud protocol during the insight portions of LITE.

We did not divide insights into categories or label them as breadth versus depth. Instead, the scoring system for domain value distinguishes between claims and supporting evidence. In the datasets used in this study, the range in the types of comments participants made was small and hence we saw no need to impose categories. Distinguishing between insights might be more practical with a dataset that contains more initial questions or in a domain with complicated relationships among data points, like systems biology.

Finally, providing the initial questions about the data model rather than asking participants for their initial questions makes it possible that participants had other unreported insights that seemed irrelevant to the specific initial questions but ultimately showed evidence of insight. Because the participants in our study were non-experts, it is a reasonable assumption that the initial questions encapsulated most of what they were able to analyze and observe. With domain experts as participants, however, there might be questions worth analyzing that would be difficult for us to predict and hard-code into the evaluation. In such cases, starting the evaluation by gathering initial questions from participants makes more sense.

Guideline *Consider the complexity of the data and participant expertise when choosing insight characteristics to measure. With a non-expert study population, provide initial analysis questions rather than requesting them from participants.*

7.2.4 Task and Workflow Considerations

We faced several workflow-related considerations during the design of the case study. First, there is a relationship among the training participants get, the specific tasks they perform, and the types of insights they are likely to report. It is possible that tasks or training direct users toward certain types of exploration activities. We deliberately tried to avoid this scenario in our case study by choosing low-level tasks that were unlikely to lead to insights on their own. An alternative approach used by North et al. is to give more complex tasks that can be classified into the same categories as the insights, in order to directly compare the activities that each visualization supports and promotes [18]. However, in that study, task performance and insights were measured using two separate experiments with different participants, and 'insightful' tasks could significantly interact with exploration and insights in a within-subjects design like LITE.

Second, we recognized that the results in LITE would be impossible to interpret correctly if the order of visualization conditions was not counterbalanced. Because the same data is explored by each participant repeatedly with different visualizations, an ordering effect on the measured insight characteristics should be expected. In our case study, we found evidence that participants spent more time and reported more valuable observations during the earlier observation prompts than during later ones (see Figure 7). Counterbalancing the orderings of visualization conditions, as we did in the case study, can mitigate the effect of order on the results.

Finally, based on our experience in our pilot study, which let participants effectively skip the exploration portion of LITE, we decided to require in our case study a minimum time during each observation prompt. This seemed to motivate participants to explore the data; we did not find that participants sat idly while the clock counted down, or that they ended their exploration as soon as the minimum time was finished. Participants were given as much time as needed to record and explore observations, so this approach does not affect the results as it would in an insight-based study with fixed length. That said, other ways to motivate participants during the insight portion of LITE might be more effective than a time requirement.

Guideline *In LITE, choose low-level tasks that will not steer participants toward insights, and be sure to counterbalance the ordering of visualizations.*

8 Conclusion

We present and demonstrate a method for evaluating visualizations called layered insight- and task-based evaluation (LITE) that combines predefined tasks and exploration. The method, which measures both task performance and insight characteristics, was applied in a case study of four different designs for a spatiotemporal network visualization in a visual analytics system. The results of our case study helped us assess which design best fit different objectives for the visual analytics system, including optimizing for task efficiency or promoting insights.

We also identified several guidelines for using LITE based on the study.

- Choose low-level tasks that are components of realistic analysis scenarios but will not steer participants toward insights.
- Counterbalance the ordering of visualizations to mitigate ordering effects in the insight component of LITE.
- Consider the complexity of the data and participant expertise when choosing insight characteristics to measure.
- Report details of the process of coding insights: who are the coders, how well did they agree, and how were disagreements resolved into one score?

Opportunities exist to address challenges we encountered using LITE. We are interested in better understanding how to run lightweight, insight-based evaluations of visualizations using the non-experts who are often recruited for task-based visualization studies. This work is a step toward more diverse evaluations of visualization tools and ones that evaluate multiple objectives for tools in a controlled setting.

Acknowledgments

This work was supported in part by Aptima, Inc., and NSF award IIS-10-16623. All opinions, findings, conclusions, or recommendations expressed in this document are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

References

[1] IEEE VAST Challenge 2011. http://hcil.cs.umd.edu/localphp/hcil/vast11/. Accessed: 2014-03-22.

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 111–117, Oct 2005.

[3] N. Andrienko and G. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Min. Knowl. Discov.*, 27(1):55–83, July 2013.

[4] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, Dec 2011.

[6] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2376–2385, Dec 2013.

[7] M. Cammarano, X. Dong, B. Chan, J. Klingner, J. Talbot, A. Halevy, and P. Hanrahan. Visualization of heterogeneous data. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1200–1207, 2007.

[8] N. Cao, D. H. Gotz, and J. Sun. Multi-faceted visualization of rich text corpora, Aug. 31 2010. US Patent App. 12/872,794.

[9] S. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization – Using Vision to Think*. Morgan Kaufmann, San Francisco, CA, USA, 1999.

[10] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors, *Information Visualization*, pages 19–45. Springer-Verlag, Berlin, Heidelberg, 2008.

[11] D. Charney, L. Reder, and G. Kusbit. Goal setting and procedure selection in acquiring computer skills: A comparison of tutorials, problem solving, and learner exploration. *Cognition and Instruction*, 7(4):323–342, 1990.

[12] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2402–2410, Dec 2012.

[13] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Moller. A systematic review on the practice of evaluating visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2818–2827, Dec 2013.

[14] R. Jianu and D. Laidlaw. An evaluation of how small user interface changes can improve scientists' analytic strategies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2953–2962, New York, NY, USA, 2012. ACM.

[15] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, Goslar, Germany, 2010.

[16] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1520–1536, Sept 2012.

[17] C. North. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE*, 26(3):6–9, May 2006.

[18] C. North, P. Saraiya, and K. Duca. A comparison of benchmark task and insight evaluation methods for information visualization. *Information Visualization*, 10(3):162–181, July 2011.

[19] T. O'Brien, A. Ritz, B. Raphael, and D. Laidlaw. Gremlin: An interactive visualization model for analyzing genomic rearrangements. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):918–926, Nov 2010.

[20] A. Perer and B. Shneiderman. Integrating statistics and visualization for exploratory power: From long-term case studies to design guidelines. *Computer Graphics and Applications, IEEE*, 29(3):39–51, May 2009.

[21] D. J. Peuquet. It's about time: a conceptual framework for represetation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994.

[22] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 109–116, New York, NY, USA, 2004. ACM.

[23] C. Plaisant, J.-D. Fekete, and G. Grinstein. Promoting insight-based evaluation of visualizations: from contest to benchmark repository. *Visualization and Computer Graphics, IEEE Transactions on*, 14(1):120–134, Jan 2008.

[24] J. Rieman. A field study of exploratory learning strategies. *ACM Trans. Comput.-Hum. Interact.*, 3(3):189–218, Sept. 1996.

[25] J. Rieman, R. M. Young, and A. Howes. A dual-space model of iteratively deepening exploratory learning. *Int. J. Hum.-Comput. Stud.*, 44(6):743–775, June 1996.

[26] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 11(4):443–456, July 2005.

[27] P. Saraiya, C. North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1511–1522, Nov 2006.

[28] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2366–2375, Dec 2013.

[29] Z. Shen, K.-L. Ma, and T. Eliassi-Rad. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1427–1439, 2006.

[30] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, pages 1–7, New York, NY, USA, 2006. ACM.

[31] A. Slingsby, J. Dykes, and J. Wood. Configuring hierarchical layouts to address research questions. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):977–984, Nov 2009.