Getting the Message? A Study of Explanation Interfaces for Microblog Data Analysis

James Schaffer
Dept. of Computer Science,
University of California,
Santa Barbara, CA 93106
james_schaffer@cs.ucsb.edu

Prasanna Giridhar
Dept. of Computer Science,
University of Illinois at Urbana
Champaign, Urbana, IL 61801
giridha2@illinois.edu

Debra Jones
SA Technologies Inc,
751 Hebron Parkway Suite
310, Lewisville, Texas 75057
debra@satechnologies.com

Tobias Höllerer
Dept. of Computer Science,
University of California,
Santa Barbara, CA 93106
holl@cs.ucsb.edu

Tarek Abdelzaher
Dept. of Computer Science,
University of Illinois at Urbana
Champaign, Urbana, IL 61801
zaher@cs.uiuc.edu

John O'Donovan
Dept. of Computer Science,
University of California,
Santa Barbara, CA 93106
jod@cs.ucsb.edu

ABSTRACT

In many of today's online applications that facilitate data exploration, results from information filters such as recommender systems are displayed alongside traditional search tools. However, the effect of prediction algorithms on users who are performing open-ended data exploration tasks through a search interface is not well understood. This paper describes a study of three interface variations of a tool for analyzing commuter traffic anomalies in the San Francisco Bay Area. The system supports novel interaction between a prediction algorithm and a human analyst, and is designed to explore the boundaries, limitations and synergies of both. The degree of explanation of underlying data and algorithmic process was varied experimentally across each interface. The experiment (N=197) was performed to assess the impact of algorithm transparency/explanation on data analysis tasks in terms of search success, general insight into the underlying data set and user experience. Results show that 1) presence of recommendations in the user interface produced a significant improvement in recall of anomalies, 2) participants were able to detect anomalies in the data that were missed by the algorithm, 3) participants who used the prediction algorithm performed significantly better when estimating quantities in the data, and 4) participants in the most explanatory condition were the least biased by the algorithm's predictions when estimating quantities.

Author Keywords

Anomaly Detection, Intelligent User Interfaces, Data Mining

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IUI 2015, March 29–April 1, 2015, Atlanta, GA, USA.
Copyright © 2015 ACM 978-1-4503-3306-1/15/03 ...$15.00.
http://dx.doi.org/10.1145/2678025.2701406

INTRODUCTION

Intelligent user interfaces have the goal of dynamically adapting to the needs of a user as they interact with an information system. Most people are familiar with personalized product recommendations (e.g. from Amazon) or movie recommendations (e.g. from Netflix). Google's search result lists are another popular example of dynamically adapted content, based on the search history of a target user. Across all of these examples, and most other adaptive information systems, users are typically kept at a distance from the the underlying mechanisms used to generate personalized content or predictions. In this work, we use the term 'recommender' to refer to complex prediction algorithms, data mining algorithms, intelligent systems, or any other algorithms which produce ranked lists of "interesting" data items, but are complex enough that they would not commonly have their mechanisms explained to the user. Note that recommendations from these systems are often presented through an otherwise static interface such as a list of search results or in a grid. Moreover, traditional browsing mechanisms such as text search, data overviews, and sorting mechanisms are often presented separately from, and operate independently of, recommendations (such as Amazon's product catalog).

During an exploratory search session with an interface, users perform iterated cycles of exploration, hypothesis, and discovery - a process often employed in scientific research, statistical analysis, and even catalog browsing. Users may start with very vague parameters, for example: "What are the most interesting movies in this genre?"; "What interesting things do Twitter users say about this topic?". Exploration can yield hypotheses that can then be answered with targeted search, for example, "is this Amazon product cheaper from another seller? Perhaps yes?"; "is this Twitter user familiar with this topic? Probably not."; "is there a higher-rated Netflix movie that is similar to this one? There has to be!". Each answer that the user finds may create new questions, prompt additional exploration, or cause the user to change his or her search strategy. Recommenders can be extremely valuable during these iterated cycles of exploration, but there is not yet a complete understanding of the interaction between recommender

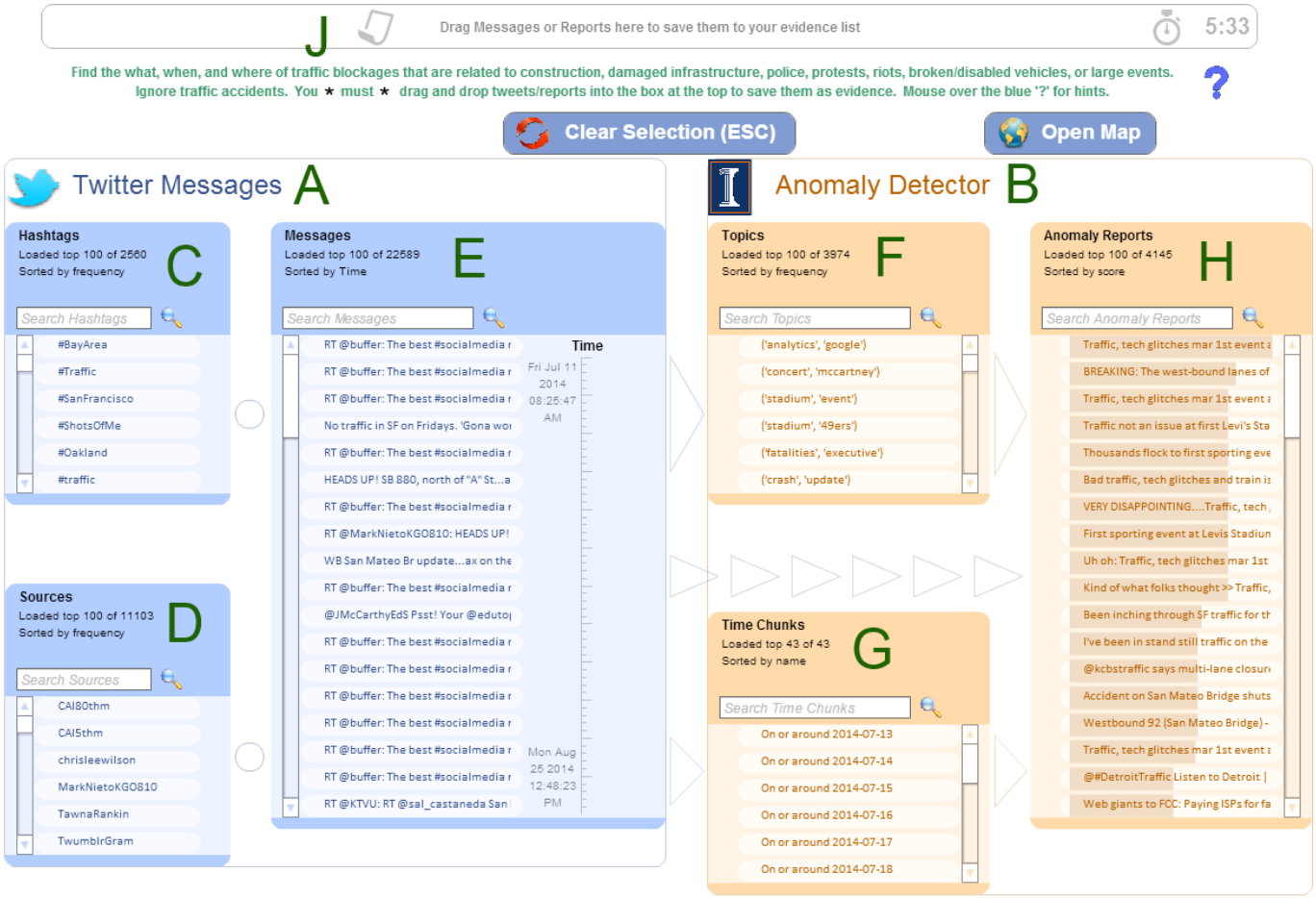

Figure 1. The provenance and data visualization tool, dubbed Fluo, showing A) the Tweet metadata, and B) metadata and results from Clarisense. Metadata from Twitter or Clarisense are organized into separate lists C) Twitter hashtags, D) Twitter sources, E) the tweets themselves, sorted by time, F) topics that Clarisense utilized for anomaly detection, G) intervals of time that Clarisense utilized for anomaly detection, and H) the final anomaly reports, ranked by Clarisense's anomaly score. At the top of the interface (J), the evidence box and remaining task time is displayed. The participant's task prompt (in green) is also shown during the duration of evidence collection as a reminder. When the user mouses over an item, additional details and explanation are displayed in a popup panel (not shown). In the experiment, participants saw a variation of the interface which omitted some metadata availability, see Table 2.

and user search strategies. Recent research in conversational or critiqued recommender systems [16, 29] go some way towards adapting to rapidly changing user needs. Research on explanation interfaces [13, 26, 25] shows that explanations can bias the user towards system predictions, but can also help the user understand why the system is predicting particular content, resulting in a better experience with the system and increased trust in its predictions. However, recommendation algorithms are not perfect as they struggle with noisy and unreliable user-provided content. We believe that providing more transparent and interactive recommenders to users performing exploratory search tasks can result in better consolidation of search strategies and thus improved performance.

Our research results that indicate that varying the degree of available information *about the underlying recommender* significantly affects the trade-offs between 1) information discovery (amount of interesting/useful information found), 2) general insight into the underlying data set, and 3) user experience with the system. By understanding these trade-offs, better interfaces that show the right content to the right user at the right time can be developed. We describe a user experiment ($N = 197$) designed to provide insight on three general research questions: 1) how can an interface be adapted to consolidate user and recommender search and exploration strategies? 2) how do recommendation algorithms change user perception of an underlying data set? 3) what are the positive and negative effects of explaining recommendation algorithms in this context?

As an example task for our experiment, we chose analysis of commuter traffic reports on Twitter in the San Francisco Bay area. This application scenario was chosen because of the large amount of potentially noisy user-provided content (Twitter postings) and associated metadata. The volume of data (22,580 messages) was large enough to make visual scanning of the messages inefficient, necessitating use of visualization and recommendation functionality within our interactive interface, Fluo [23]. An automated anomaly detector and recommendation algorithm [10], Clarisense, was used

to generate recommendations of anomalous messages in the dataset. This system serves as an example of a prediction algorithm - the results from this experiment should reasonably apply to similar systems.

BACKGROUND

This research is about the boundaries, limitations and synergies that exist between a human information analyst and an automated algorithm. Visualization and recommender systems are both key components. While the example anomaly-detection algorithm used (Clarisense) is not a recommender system (RS) in the classic sense (e.g.: no personalization is used), we maintain that the study has particular relevance to recommender systems research because use of interactive interfaces [18, 5, 30], explanations [12, 14, 25] and control [15, 21] in recommenders is increasing, but there is still a limited amount of current research in this specific area. We are specifically interested in the interplay between the RS algorithm and the interactive interface using real world data. Other research in this specific area i includes Parra et al.'s SetFusion/ConferenceNavigator system [21], and Bostandjiev's TasteWeights experiments. [5, 15] We believe that this paper is a small but important step in this direction.

Recommender Systems and Explanations

Over the last 15 years, research has shown that explanation of a recommender system's reasoning can have a positive impact on trust and acceptance of recommendations. Early work by Herlocker [13] studied a variety of explanation mechanisms and their impact on trust, satisfaction and other subjective metrics, concluding that certain styles of explanation can convince a user to adopt recommendations. Bilgic et al. [4] furthered this work and explored explanation from the promotion vs. satisfaction perspective, finding that explanations can actually improve the user's impression of recommendation quality. Later work by Tintarev and Masthoff [25] surveyed literature on recommender explanations and noted several pitfalls to the explanation process, notably including the problem of confounding variables. This remains a difficult challenge for most interactive recommender systems [26], where factors such as user ability, mood and other propensities, experience with the interface, specific interaction pattern and generated recommendations can all impact on the user experience with the system. The importance of system transparency and explanation of recommendation algorithms has also been shown to increase user adoption of recommendations by Knijnenburg in [15].

The broader field of intelligent systems produced research relevant to our study. Gregor et al. [12] provide an excellent summary of the theory of crafting explanations for intelligent systems. User studies that test the effects of explanation typically vary explanation level and quantify concepts such as adherence or knowledge transfer. Key findings show that explanations will be more useful when the user has a goal of learning or when the user lacks knowledge to contribute to problem solving. Explanations have also been shown to improve learning overall and improve decision making. The impact of explanation on both novices and experts has also been extensively studied: novices are much more likely to adhere to the recommender/expert system due to a lack of domain knowledge, and expert users require a strong 'domain-oriented' argument before adhering to advice. Experts are also much more likely to request an explanation if an anomaly or contradiction is perceived. Most of these studies focus on decision making domains (financial analysis, auditing problems) and were conducted before the explosion of data which now characterize typical web databases. When browsing or analyzing data that is too large to be analyzed by hand, decision makers have no choice but to utilize automated filtering techniques as part of their search strategy - this creates new questions about what might change in the dynamics between humans and automated algorithms.

Evaluation of Interactive Interfaces

The visual analytics community has begun to favor open-ended protocols over benchmark tasks for the evaluation of interactive interfaces [7][17][28]. This is partly due to a realization that most visualization systems are overly-specific, and thus not agile or adaptive enough to handle non-deterministic, open-ended data exploration with a higher level goal of decision making or learning [2][1]. Though recent systems have become substantially more expressive [22][11], the question of how to effectively evaluate the usefulness of such systems is still open. Researchers recommend that participants in experimental visualization tasks should be allowed to explore the data in any way they choose, creating as many insights as possible, and then measuring their insight with a think-aloud protocol or qualitative measures, such as quantity estimation or distribution characterization. This contrasts starkly with typically well-defined benchmark tasks, which usually have users do things such as find minimum or maximum values, find an item that meets a specific criterion, etc. North [17] cautions that most benchmark tasks may only evaluate an interface or visualization along a very narrow axis of functionality. We believe that these recommendations on evaluation strategies go beyond the realm of visual analytics, and are applicable to the evaluation of interfaces such as Netflix or Amazon which users return to daily with new insights and understandings of what content or products are available.

Microblogs for Traffic Analysis

Daly et al. [8] also study the domain of commuter traffic, with a view to increasing user understanding of a large corpus of real time Twitter messages. The approach in this case contrasts with our research in that they evaluate a system using a novel combination of Linked Data and Twitter messages to inform users about anomalies, rather than studying how interfaces might best support explanation facilities in this context.

APPROACH

This section describes the the anomaly recommender, Clarisense, and the interactive interface, Fluo, in more detail. The Twitter tweets and related metadata shown to participants were collected between July 12, 2014 and August 24, 2014. Tweets were filtered by looking at the keyword 'traffic' near San Francisco, California, USA. These tweets were then fed to the content recommender for summary, and a provenance view of this operation was shown in the Fluo interface. In

most treatments, participants were also given the original unfiltered dataset alongside the Clarisense recommendation.

Clarisense Architecture

Clarisense is a Twitter-targeted automated anomaly detection algorithm developed at UIUC. To rank microblog items, Clarisense employs a search strategy of examining the frequency of topics over time in the microblog collection. The first step in the Clarisense pipeline is the division of the input data into 24 hour time chunks followed by clustering within each chunk in order to retain only unique tweets and remove any redundant information if present. The 24 hour parameter was determined based on the percentage of delay between the retweets and the original tweet that was observed. In each cluster, only one tweet is chosen as relevant and passed to the next step in the pipeline. The next goal is to find the events within each cluster that stand out from the normal ongoing events during that period of time. We opted for an approach which uses keywords from each tweet as the primary purpose of identification. A pair of keywords, rather than single keywords or n-tuples, showed the highest correspondence between independent events and their keyword signatures. Discriminative keyword pairs were extracted from each time chunk by leveraging spatial and temporal data to determine which keywords were mundane. For each time chunk, the discriminative keyword pairs were used to rank the microblog entries in terms of their information gain. Finally, normalizing the information gain over all time chunks yields the data for the ranked list shown in Figure 1(H). A full description of Clarisense can be found in [10].

Fluo

This section introduces our interactive interface (and experimental platform), Fluo, and briefly describes its methodology.

Design

Fluo (Figure 1) is a provenance visualization that was designed for exploring the top-N results from a ranking algorithm. It is part of ongoing work in the inspectability and control of recommenders and data mining algorithms at UCSB [5]. In the interface, data items or intermediate calculations are represented as nodes and organized into columns, which can be placed serially (creating an upstream/downstream relationship) or in parallel (to represent that multiple sources are weighted together). Each node in the visualization may have a corresponding "score" which is shown as a gauge and can be mapped to any corresponding value in the underlying algorithm (e.g., Pearson Correlation for collaborative filtering). A mapping from interaction techniques to commonly recognized user intents [27] is shown in Table 1.

Fluo simplifies reconfiguration of metadata and visual relationships for each experimental treatment. A breakdown of which metadata is shown in which condition is provided in Table 2. During the experiment, users engaged in an interactive tutorial that explained the interface and all available metadata based on the treatment. The modular design of the interface and consistent interaction techniques across configurations for each treatment allowed for easy interpretation of results.

User Intent	User Action and Response
Select	User selects an item in a list, the system highlights the item and keeps it at the top its list. Details on demand and more explanation are shown on mouse-hover.
Explore	User scrolls a list, the system shows new items along a fixed parameter (time, frequency, relevance to search term)
Reconfigure	For the purpose of evaluation, the types of items represented and the sort parameters were fixed by the experimenters ahead of time.
Encode	For the purpose of evaluation, the color, size, and shape of items was fixed by the experimenters ahead of time.
Abstract /Elaborate	The user mouses over an item in a list, the system provides additional details about the item in a panel.
Filter	The user selects a time bin, only items from that time are shown. The user enters a search term, only items matching that term are shown.
Connect	The user selects an item, connected items (friends) are brought to the top of their respective list. The user can then expand the selection to show even more connected items (friends of friends).

Table 1. User intents supported by the interactive interface for this experiment.

Explanation of Clarisense

Clarisense's search strategy was simplified for participants and presented through the interface, as shown in Figure 1. Intermediate steps of the algorithm and their values (time chunks and topics or keyword pairs) were exposed as provenance metadata. During the training period, applicable participants were given a detailed explanation of each kind of metadata, how they relate, what constitutes a high or low anomaly score, and how to form queries that reveal relationships between the original dataset, the extracted keywords, and their frequency on various time chunks. Participants were also required to answer specific Clarisense-related questions during the training period before they could proceed, and all information about Clarisense remained available to participants even during the task phase.

Experiment Design

We examined how varying levels of explanation from the recommender affect the entire human-recommender system's ability to 1) find relevant, interesting data items and 2) generate an overall understanding or accurate perception of the data, especially when data items are too large to be browsed sequentially. We also measured the effects of various levels of explanation on the user's confidence, perception of the tool, and enjoyment of the task.

In our task protocol, we compromised between a truly open-ended task and a benchmark task by giving users a set of high-level search parameters relating to traffic blockages

Independent Variable	
Treatment	**Description**
Baseline - Tweet Metadata Only (Figure 1, A)	Twitter metadata (source, tweet, hashtags, time) shown. Text search over message body content, filter by time. Different selections of messages, sources, and hashtags unveil different relationships through edges on-demand.
Clarisense Only (Figure 1, H)	Clarisense's summarized reports with text search, filter by time. The 'what' of Clarisense's reports are summarized but not the 'how' (no provenance). Twitter metadata and messages are not present.
Clarisense in Context (Figure 1, A+H)	A combination of the two previous conditions. Additionally, users can see the relationship between the original tweets and the reports, making this a partial provenance view.
Clarisense in Context w/ Explanation (Figure 1, A+B)	Similar to the previous condition, but Clarisense's selected time intervals and topic modeling were exposed to the users, making this tool a full provenance view of Clarisense's anomaly calculation.

Table 2. In this experiment we manipulated the amount of explanation, control, and metadata availability. Above is a description of each experimental treatment.

and allowed them to explore the dataset in any way they chose. Additionally, we measured performance by comparing participant-reported events against a benchmark that was created post-hoc by examining all events discovered by participants. In agreement with [17], we believe that this methodology reduced evaluation biases that occur when users are assigned very specific search parameters. Additionally, this approach also sidestepped some of the major difficulties with longitudinal studies while still being a good representation of many real world exploratory search tasks.

The experimental toolkit was deployed as a web service and the link was made available on Amazon Mechanical Turk (AMT). The AMT web service is attractive for researchers who require large participant pools and low cost overhead for their experiments. However, there is valid concern that data collected online may be of low quality and require robust methods of validation. Numerous experiments have been conducted, notably [6] and [20], that have attempted to show the validity of using the service for the collection of data intended for academic and applied research. These studies have generally found that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled. Previous studies of recommender systems have also sucessfully leveraged AMT as a subject pool [5, 15], however, most AMT workers expect tasks between 60 seconds and 5 minutes on average. Longer tasks may catch users off guard, fatiguing them and increasing tendency for satisficing. While we took detailed timing metrics for all interactions, for some time windows it is difficult to tell if a user is merely thinking or, e.g., went to use the restroom. Additionally, if a participant suffers from a key misconception, we cannot correct or account for it. Fortunately, larger sample sizes and quicker uptake help mitigate some of this noise inherent in AMT experiments.

We carefully followed recommended best practices in our AMT experimental design and procedures [6] and [20]. For filtering AMT workers, we chose to require that participants had successfully completed at least 50 HITs on the system. Participants were paid an average of 4 dollars plus a 1 dollar performance-based incentive. The bonus payment was made to everyone who completed the study. Numerous satisficing checks [19] were placed throughout the pre-study (e.g. what is 4 + 8?), and the training phase (as outlined below) reasonably insured a minimum level of understanding before participants were allowed to proceed. We also collected some basic demographic information from each participant, including information about how frequently they drive a car and their familiarity (if any) with San Francisco's Bay Area.

Experiment Protocol

After accessing the experimental system through AMT, participants were presented with a pre-study questionnaire using the Qualtrics survey tool [1], collecting basic demographic and background information. Next, they were directed to one of the variations of our online tool for a training session. Once training was complete, the open-ended search task was described and participants used the interface to explore the tweets and Clarisense's reports into a list of evidence. Once time was up, the interface was removed and we asked them several questions related to key quantities in the dataset that were related to the exploration they performed. The training and evidence collection protocols are talked about in more detail below.

Training

Since the experimenters could not verbally direct the participants, a complex training module was created which walked the participant through key concepts before the evidence collection portion of the task. The participant was required to answer a series of targeted search questions, the answers to which could only be known once the participant identified which parts of the interface were providing what information. An unlimited number of attempts were given for each question. Easier questions were chosen as multiple choice with fewer than 4 options, while the hardest questions had blank response forms that required the entry of quantities. After informal pilot testing in the lab, it was decided that hints for every question were needed to alleviate participant fatigue during this portion of the study. During data collection, few participants reported difficulties completing the training in any condition. The average failure rate for targeted search questions related to the interface functions and Twitter metadata

[1] www.qualtrics.com

was 95% (which means that, on average, participants submitted a wrong answer for each question on the first attempt). For questions related to Clarisense and the anomaly detection strategy, the failure rate was only 27%.

Evidence Collection

Once training was completed, participants were prompted that the evidence collection phase was about to begin. They were told that the dataset contained numerous traffic blockages and that we were interested in studying blockages related to construction, infrastructure damage, broken or disabled vehicles, police activity (riots, protest), and planned public events. Participants were actively told to ignore traffic accidents, and distinctions were made between planned public events such as sports games or concerts and other events like riots. The active prompt for the task is shown in Figure 1 (J). Participants were told to look for these events and collect evidence in a list (by dragging and dropping either Tweets or anomaly reports), and that they would be paid a bonus for finding more interesting evidence related to blockages. Participants were restricted to 15 minutes for this portion of the task, and a ticking clock (Figure 1) indicated the time remaining.

Metrics

The independent variable in this experiment is detailed in Table 2 (see also Figure 1). An overview of the dependent variables in this experiment are are shown in Table 3.

Event Recall

Once all participant data was collected, an analysis of evidence yielded a list of events, which is shown in Table 4. Each event $e \in E$ was re-constructed by manual inspection from tweets chosen by participants. Non-descriptive tweets that mention traffic but do not mention at least the what or the where were not included in the final benchmark, nor were events that were reported by fewer than 3 participants. After identifying the where, what, and when of the events, we identified the complete set of tweets that described the event. For example, there were 17 tweets in the dataset that unambiguously identified the 'quake' event. During analysis, we said a participant discovered an event if they included at least one message in their evidence, $v \in V_p$, from that event's ground truth in the event benchmark, $g \in G_e$, with the rest of the submitted evidence being classified as noise $n \in N_p$, $N_p \subseteq V_p$. That is, noise is every message in the dataset not related to some event in the post-hoc benchmark. Recall is simply defined as the total number of events detected by a participant:

$$recall = \sum_{e \in E} \begin{cases} 1 & : |V_p \cap G_e| > 0 \\ 0 & : |V_p \cap G_e| = 0 \end{cases} \quad (1)$$

Note that recall is **not** normalized, and can fall between 0 and 22. Recall also indirectly gives us a rate of event discovery since all participants were limited to exactly 15 minutes for the evidence collection phase. However, there was one exception in our benchmark. Due to the large size of the 'Web Traffic' event (see Table 4) and because participants were not explicitly told that this event is interesting at the start of evidence collection, we assessed if participants made this insight using a multiple-choice question after the evidence collection phase.

Dependent Variable	Description
Event Recall	Total number of events the participant discovered through evidence submission during the fixed-time phase of the task.
Estimation Error	Participant's error in estimating the quantity of events related to specific types of blockages (disabled vehicles, damaged infrastructure, police/riot/protest,planned public events). Responses provided during the post-study.
Usability	Participant's confidence, enjoyment, and perceptions of the tool, taken on a Likert scale in the post-study.

Table 3. Description of dependent variables.

We also considered precision, which in this case measures the percentage of noise the participant included in their final evidence list N_p:

$$precision = \frac{|V_p| - |N_p|}{|V_p|} \quad (2)$$

Precision allows us to understand the quality of evidence the participant submitted (e.g., was the participant paying attention or merely grabbing as much evidence as possible?)

Estimation Error

After the evidence collection task ended, we requested that the participant estimate of the number of blockages that were actually represented in the dataset which pertained to a particular type of incident. Disabled vehicles, damaged infrastructure, police/riot/protest, and planned public events were chosen for these questions due to practical limitations on generating ground truth for events that are likely to have two instances occur simultaneously in time (construction, traffic accidents). Participants entered their answer in plain text boxes. These metrics gave us the participant's 'qualitative understanding' of the impact of each type of incident on traffic.

Usability

Participant perceptions of the tool were collected after the evidence collection and estimation tasks. Participants provided answers on a Likert scale (1-7) for each question on a web form. Participants were asked "How confident were you using the tool to complete the task?", "How much did you like the interface tool?", "How much did you like the training portion of the task?", and "How much did you enjoy the evidence collection portion of the task?"

Event	Description	Score	Size
gas leak	On 7/11, a gas leak caused a closure of 7th St and Broadway	0	4
drake/eliseo light	On 7/22, a traffic signal on Eliseo Dr malfunctioned	0.21	4
08/12 oil spill	Oil from a truck was spilled on the San Mateo bridge	0.82	10
08/21 oil spill	Oil from a truck was spilled on Magdalena Ave	0.34	6
cesar light	On 08/18, a traffic light malfunctioned on Cesar Chavez	0.29	4
quake	On 08/24, a major quake damaged multiple roads in Napa, Vallejo, and Sonoma	0.13	17
andy lopez	On 07/12, A protest demanding justice for Andy Lopez blocked highway 101	0.26	6
07/20 market st protest	A protest caused severe congestion on Market St	0	1
07/26 market st protest	A protest caused severe congestion on Market St	0.10	2
ferguson	On, 08/22, a protest of the Ferguson shooting caused traffic to stop near Civic Center Plaza	0.05	3
coliseum	On 07/25, a bomb threat at Coliseum Station caused highway 880 to become blocked	0.46	10
lombard	On 07/12 and 07/19, Lombard St was closed to the public by city officials	0.14	3
49ers	On 08/03, a 49ers game at Levi Stadium caused severe congestion	1.00	30
obama	On 07/23, an Obama visit caused multiple blockages/road closures near downtown	0.48	33
mccartney	On 08/14, a Paul McCartney concert caused severe congestion near Candlestick theater	0.48	30
soccer	On 07/26, a soccer game at UC Berkeley caused severe congestion	0.29	9
marathon	On 07/26, the San Francisco Marathon resulted in multiple road closures	0.18	13
japan	On 07/19, a J-Pop Festival in Japantown resulted in road closures and severe congestion	0.65	12
terminator	On 08/03, the Golden Gate Bridge was closed for the filming of Terminator 5	0.08	2
st francis constr	On 07/16, part of St Francis Dr was closed all day to traffic	0.03	2
slurry seal	on 08/07 and 08/08, construction caused delays and closures near Ralston Ave	0.08	6
web traffic	A significant percentage of the messages in the dataset related to web traffic	0.08	-

Table 4. The post-hoc benchmark - events discovered by participants during the task. 'Score' indicates Clarisense's recommendation for tweets associated with this event. 'Size' indicates the total number of distinct Tweets that identified the what, where, and when of the event.

Hypotheses

Evaluating features of each treatment separately, then in combination enabled systematic assessment of the value of each feature and additionally allows for the identification of synergistic value gained by the combinations. The following hypotheses were evaluated during this experiment:

- H_1: interface type impacts total number of events recalled
- H_2: interface type impacts which events are recalled
- H_3: interface type impacts evidence precision
- H_4: interface type affects estimation error
- H_5: interface type affects the user's confidence
- H_6: interface type affects the degree to which the user liked the tool
- H_7: interface type affects the degree to which the user liked the training session
- H_8: interface type affects the degree to which the user liked the open-ended search task

RESULTS

Participants

AMT participant age ranged from 18 to 65, with an average of 25 and a median of 27. 52% of participants were male while 48% were female. 563 workers completed the pre-study, but only 197 finished the study in its entirety. After visually inspecting the data and plotting the results, we found no strong outliers in the remaining 197.

Analysis

Table 5 shows precision and recall for participants across treatments for the post-hoc benchmark (Table 4). Note that we measure recall non-normalized as to best represent the magnitudes of the quantity of discoveries. A large increase in recall is seen between the Twitter Only' condition and the conditions where Clarisense was present. A slight drop in total discoveries seems to occur between the 'Clarisense Only' condition and the conditions with Clarisense AND the original Twitter data. Participants were grouped by whether the recommender was available (no=60,yes=137) and a single factor analysis of variance was run, showing a significant decrease of 60% when Clarisense was not present ($F = 92.87, p < 0.01$). When we compare the 'Clarisense Only' condition with the conditions that provided context and explanation facilities, we see a 25% drop in recall rate ($F = 19.54, p < 0.01$). Thus H_1 is supported.

We hypothesized that the presence of Clarisense and its presentation would have a significant impact on which events were discovered by participants. To test this, we considered only the events which we determined that Clarisense 'missed' or underrepresented due to its filtering and reporting mechanism (Figure 2). To qualitatively determine what Clarisense had 'missed', we decided that a lenient anomaly score threshold should be chosen that would give Clarisense a precision

Condition	Precision	Recall
Twitter Only	0.14	2.77
Clarisense Only	0.60	8.23
Clarisense with Context	0.44	6.44
Clarisense with Explanation	0.50	5.85

Table 5. Mean precision and recall for each interface configuration. Participants that only interacted with the anomaly recommender were able to incorporate more of its reported events in the same time period and take advantage of its precision.

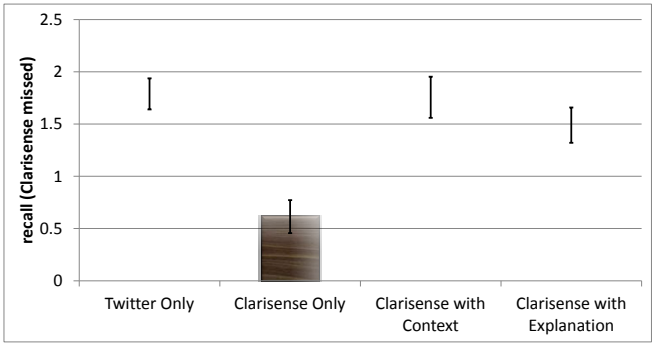

Figure 2. Participants that interacted with the original data were able to consistently find events that the anomaly recommender missed or classified as relatively less interesting

of at least that of the worst participant in the Twitter Metadata only condition (0.034). We settled on a score threshold of one standard deviation above the mean (0.15), which corresponded to a precision of 0.035. Referring to Table 4, this means that Clarisense reported 12 events total (slightly more than half) from 339 pieces of evidence total, which results in a list of 10 events that Clarisense missed. In the 'Clarisense Only' condition, participants only appear to be half as likely to discover one of these events. We again grouped the participants by whether they had the original Twitter data available (no=51,yes=146) and ran another single-factor ANOVA between conditions that contained the original Twitter dataset and the 'Clarisense Only' condition, finding a 63% decrease ($F = 33.65, p < 0.01$) in underrepresented events when only Clarisense was present. This supports H_2.

Looking at Table 5, it can be seen that the relative proportion of values within precision and recall was roughly the same, indicating that an increase in recall corresponded to a similar increase in precision, unfortunately reducing the usefulness of the precision metric when interpreting results. Still, an ANOVA revealed large differences between the conditions ($F = 57.44, p < 0.001$), with big differences between 'Twitter Only' and the other treatments ($p < 0.001$ for all), but also a difference was found between 'Clarisense Only' and 'Clarisense with Context' ($p < 0.001$), and another difference between 'Clarisense Only' and 'Clarisense with Explanation' ($p < 0.027$). Thus H_3 is supported.

Figure 3 shows the overall estimation error from our questionnaire which followed the evidence collection task. The vertical axis shows the average difference between actual and estimated distinct blockages for each treatment. Since the scales of the ground truth were similar (disabled vehicles: 29,

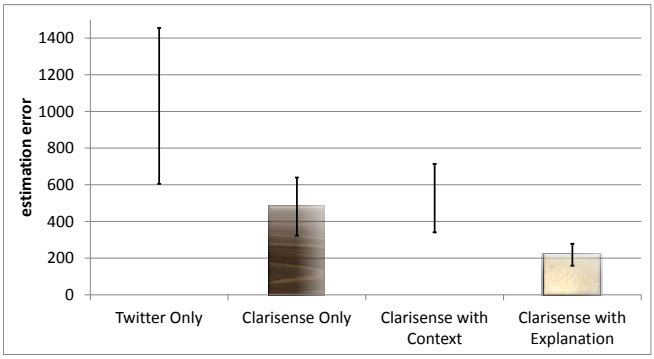

Figure 3. After the open-ended task, participants that interacted with the anomaly recommender consistently had a better understanding of the quantities in the data. The presence of explanation and provenance improved estimation further.

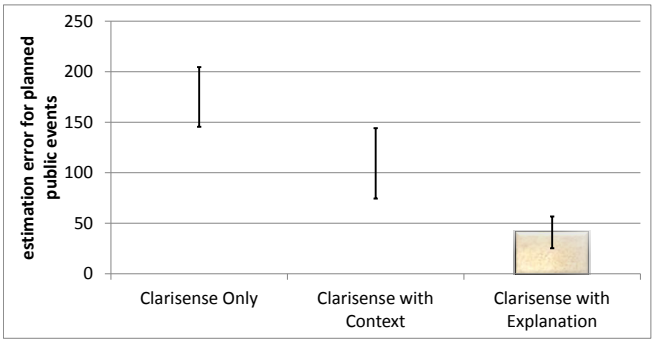

Figure 4. The anomaly recommender was much more likely to report major public events. The level of explanation greatly decreased the participant's error in perception for the frequency of these types of traffic blockages.

damage: 6, police/riot/protest: 9, planned public events: 12) we aggregated these results into one graph. A value of 0 indicates perfect accuracy. Participants were much more likely to overestimate than underestimate. A large increase in estimation accuracy can be seen between the condition where Clarisense was absent (Twitter Only) and the other three conditions. Another drop can be seen between the Clarisense explanation condition and the conditions where less explanation is given. We ran a single factor analysis of variance between the 'Twitter Only' condition and the conditions with the recommender, showing an estimation error decrease of 60% ($F = 4.8, p = 0.030$). We also tested the 56% drop in estimation error for 'Clarisense with Explanation' against the other two Clarisense conditions, finding it fell just short of the 0.05 significance level despite its notable effect size ($F = 2.99, p = 0.087$). To further investigate the decrease in estimation error for the full explanation condition, we plotted the estimation parameters for each type of blockage individually (Figure 4). The most notable of these was a large difference in planned public events - there was a 31% decrease in estimation error in the 'Clarisense with Context' condition and a 75% decrease in estimation error in the 'Clarisense with Explanation' condition. For the latter, we ran another single-factor ANOVA and found ($F = 4.10, p = 0.046$). These findings support H_4.

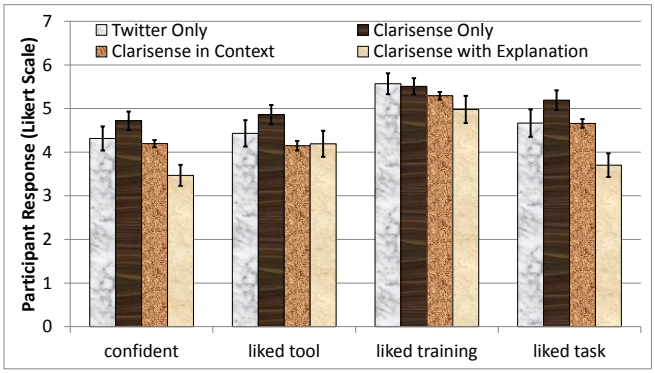

Figure 5. Presence of the original dataset and an increase in explanation corresponded to a decrease in confidence and enjoyment of the task.

Finally, we tested the impact of the increasingly complicated interface and explanation of Clarisense on the participant. Figure 5 shows the results from our post study questionnaire. The answers were provided on a Likert scale (1-7). From left to right, the questions were 'How confident were you using the tool to complete the task?' (confident), 'How much did you like the interface tool?' (like), 'How much did you enjoy the training portion of the task?' (training), 'How much did you enjoy the evidence collection portion of the task?' (task). Presence of the original Twitter data appeared to decrease both confidence and enjoyment of the task, with the largest drops (27% decrease in confidence, 29% decrease in task enjoyment) being between the 'Clarisense with Explanation' and the 'Clarisense Only' conditions. To test the differences between 'Clarisense Only' and 'Clarisense with Explanation', we ran two more single-factor ANOVAs yielding ($F = 15.28, p < 0.01$) for the 27% confidence drop and ($F = 7.074, p < 0.1$) for the 29% task enjoyment drop. More analysis was run for enjoyment of the training session and likeability of the tool, but nothing below the accepted significance level was found. Thus, H_5 and H_8 are supported, but we do not find enough evidence here to support H_6 and H_7.

DISCUSSION

The results uncovered numerous trade-offs related to diversity and quantity of reported events. H_1, H_2, and H_3 indicate that the recommender was useful overall when exploring the dataset, but strongly affected which events were reported. This result is not too surprising given that Clarisense presented several key events to participants even without any interaction. Participant records indicate that they drew their evidence from Clarisense with 50% likelihood in these conditions (remaining evidence came from the original Twitter data). Unfortunately, it seems that when participants spent time employing their own search strategy on the Twitter data, it detracted from the rate at which they considered and incorporated the recommender's discoveries. However, presenting the original Twitter data in the Fluo interface allowed the participants to develop search strategies that yielded different discoveries than Clarisense - note that two of the events discovered were not represented in the 'Clarisense Only' condition (gas leak, 07/26 market st protest), and the remaining 8 were classified as significantly less interesting by the recommender. Of particular note is the 'terminator' event, which saw remarkably higher probability of recall when both the Twitter data and Clarisense reports were present. Evidently, even our novice participants were able to develop search strategies that consistently contributed at least a few novel discoveries to the analysis process.

H_4 indicates that recommender presence had a positive impact on the ability to estimate quantities in the data. The dataset may have simply been too large to gain a good understanding in the limited time frame, but it seems as though the recommender was able to provide a better 'orientation' in the data in a much shorter time. We reason that when the participants scanned the informative tweets in the Clarisense column they formed a reasonable estimate of what was present in the dataset, since Clarisense's recall was comparatively high. The drop in estimation error when concerning planned public events becomes even more meaningful when we consider Clarisense's strategy for reporting interesting anomalies in the dataset. On Twitter, large public events usually have distinct key terms and usually hashtags associated with them that only appear in conjunction with the event. As such, Clarisense is more likely to view these as anomalous than other types of events in the dataset. From Table 4, it can be seen that planned public events dominate the top 5 most anomalous events from Clarisense's perspective. The 'Topics' column likely gave participants additional insight into the quantities and thus reduced estimation error, since it becomes obvious that Clarisense is very interested in large public events from scanning the top topics.

Our usability results (H_5, H_8) indicate that explanation facilities can potentially drop both a user's confidence and make the process of search more stressful. The drop in confidence with increased explanation may mean that participants were comparing the complexity of their own strategy against the one used by Clarisense, thus feeling like their contribution was not as significant. Across treatments, the participant's enjoyment of the training session and the likeability of the tool (H_6, H_7) did not appear to vary much, which was surprising due to the varying length of the session and complexity of the tool based on the treatment.

To get more insight on the training session and the likeability of the tool, we examined the qualitative feedback that was given at the end of the task. Though the free-text responses were too noisy to provide evidence for supporting H_6 and H_7, they still provided insight. By and large, participant comments indicated that the 'Clarisense Only' treatment was the easiest to use and the most "intuitive." Participants in the 'Clarisense with Explanation' condition often reported that the interface was too complicated or difficult to use, despite their competitive results (though, this feedback agrees with the 'confidence' scores). Participants across all conditions reported that the experiment was very interesting, e.g. "this certainly stretched my brain!" and that the training session was very helpful. Also recall that participants struggled much more on training questions related to the Twitter metadata and interface usage (seen by all participants) than training ques-

tions related to Clarisense, which helps explain the similar responses from participants about the training phase.

Key Takeaways

Recommend First, Search Second: Interfaces should highlight results from a recommender when a user begins the process of data exploration, but general search and exploration tools should always be available. The participants in our experiment benefited greatly from the recommender presence, consistently reporting better estimations of the quantities in the data over those that received no recommender. Participants that could search over the original dataset were still apt to do so, and through their own search strategy discovered events that the recommender missed. The recommendations themselves may also serve as catalysts for initial search strategies or improve learning [3][24], which can greatly help new users, novices, or those working with new datasets.

Contextualize and Explain Recommendations: Both the introduction of the original Twitter data and more explanation facilities appeared to help participants understand and contextualize Clarisense's search strategy, which greatly decreased their estimation error with respect to the events that Clarisense over-represented (public events). Explanation facilities should carefully explain the search strategy of a recommender to users when this is appropriate and put the recommendations in context to avoid these errors. Though not every recommendation system is the same, in domains where decisions are costly, perception biases could be disastrous. For example, analysts of epidemics might ask: what is the relative severity of illness x and hazard y at a specific location - which problem should more resources be allocated to? In other examples, web developers for digital storefronts may want to avoid creating misconceptions about the variety and quantity of items that his store has available, or a library might want to emphasize the impression of diversity among its titles. Ongoing research [9] is still trying to understand the effect of recommender systems on diversity of items delivered to users. It may be possible that users will abandon these services and others (e.g. Netflix, Amazon Prime, Hulu) if they misestimate the diversity of items in the catalog and develop a negative perception of the service.

Recommend to New Users, Explain to Returning Users: In this experiment, full explanation of the recommender decreased user confidence and enjoyment of the open-ended search task. Previous research has also shown the cost of digesting explanations from recommender systems [12]. However, in this case, the presence of the daunting Twitter dataset also appeared to contribute. While most of the participants in this task could be classified as novices in the field of information analysis, they were also new to the tool and some were new to the concept of Twitter. By creating and maintaining models of users, different configurations of the recommender and search tool might be shown at different times. For instance, a digital shop could minimalize their storefront and only initially show recommendations until the user requests a targeted search. When regular customers are established, the store can begin explaining/contextualizing recommendations so that the user can synthesize the recommendations with their own search strategy, potentially finding new products.

Limitations and Future Work

The evidence collection portion of the task was limited to 15 minutes and all users were essentially novices with the Fluo interface. Given more time and more comprehensive training, it is possible that users would have reached a 'saturation point,' where all useful information from the recommender would have been exhausted and more events from user-contributed search strategies would have emerged. To get more insight into the learning curves associated with the system, and therefore a better indication of performance, the authors plan a near-term follow up study to evaluate performance with the user interface without a time restriction.

The specific interaction methodology of Fluo was not varied between conditions, the only variation was in the amount of provenance and metadata shown. In future experiments, different interaction techniques to support different user intents (to better support a variety of user preferences) could be tested in tandem with different explanation facilities. It may be the case that different interaction techniques lend themselves to better presentation of contextualized explanation.

CONCLUSION

In this work, we sought to answer the questions: 1) how can an interface be designed to maximally leverage user and recommender search strategies for exploratory search? 2) how do recommendation algorithms change user perception of an underlying data set? and 3) what are the positive and negative effects of explaining recommendation algorithms in this context? An ($N = 197$) user experiment was run to determine the impact of explanation from a recommender during an exploratory search task. Results show that the presence of recommendations in the interface allowed participants to quickly find more interesting items and improved estimation ability. The presence of search tools allowed participants to develop their own search strategies and find items missed by the recommender, and participants in the most explanatory condition were able to avoid a perceptual bias that affected other participants (70% reduction in bias when estimating public events). We conclude that designers should carefully evaluate time costs and negative impact on user experience when providing explanations, but explanations remain an important design consideration due to their positive impact on a user's understanding of recommender search strategies.

ACKNOWLEDGMENT

The authors would like to thank Coty Gonzalez and Jason Harman at Carnegie Mellon University for their advice on this research. This work was partially funded by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

1. Ahn, J.-w., and Brusilovsky, P. Adaptive visualization for exploratory information retrieval. *Information Processing & Management 49*, 5 (2013), 1139–1164.
2. Amar, R. A., and Stasko, J. T. Knowledge precepts for design and evaluation of information visualizations. *Visualization and Computer Graphics, IEEE Transactions on 11*, 4 (2005), 432–442.
3. Arnold, V., Clark, N., Collier, P. A., Leech, S. A., and Sutton, S. G. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *Mis Quarterly* (2006), 79–97.
4. Bilgic, M., and Mooney, R. J. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI*, vol. 5 (2005).
5. Bostandjiev, S., O'Donovan, J., and Höllerer, T. Tasteweights: a visual interactive hybrid recommender system. In *RecSys*, P. Cunningham, N. J. Hurley, I. Guy, and S. S. Anand, Eds., ACM (2012), 35–42.
6. Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6*, 1 (2011), 3–5.
7. Chang, R., Ziemkiewicz, C., Green, T. M., and Ribarsky, W. Defining insight for visual analytics. *Computer Graphics and Applications, IEEE 29*, 2 (2009), 14–17.
8. Daly, E. M., Lecue, F., and Bicer, V. Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM (2013), 203–212.
9. Fleder, D. M., and Hosanagar, K. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*, ACM (2007), 192–199.
10. Giridhar, P., Amin, M. T. A., Abdelzaher, T. F., Kaplan, L. M., George, J., and Ganti, R. K. Clarisense: Clarifying sensor anomalies using social network feeds. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom 2014 Workshops, Budapest, Hungary, March 24-28, 2014* (2014), 395–400.
11. Gratzl, S., Gehlenborg, N., Lex, A., Pfister, H., and Streit, M. Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 2023–2032.
12. Gregor, S., and Benbasat, I. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.
13. Herlocker, J. L., Konstan, J. A., and Riedl, J. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, ACM (New York, NY, USA, 2000), 241–250.
14. Herlocker, J. L., Konstan, J. A., and Riedl, J. Explaining collaborative filtering recommendations. In *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work* (2000), 241–250.
15. Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., and Kobsa, A. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, ACM (2012), 43–50.
16. McCarthy, K., Reilly, J., McGinty, L., and Smyth, B. Experiments in dynamic critiquing. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, ACM (New York, NY, USA, 2005), 175–182.
17. North, C. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE 26*, 3 (2006), 6–9.
18. O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., and Höllerer, T. Peerchooser: visual interactive recommendation. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM (New York, NY, USA, 2008), 1085–1088.
19. Oppenheimer, D. M., Meyvis, T., and Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45*, 4 (2009), 867–872.
20. Paolacci, G., Chandler, J., and Ipeirotis, P. G. Running experiments on amazon mechanical turk. *Judgment and Decision Making 5* (2010), 411–419.
21. Parra, D., Brusilovsky, P., and Trattner, C. See what you want to see: visual user-driven approach for hybrid recommendation. In *IUI'14 19th International Conference on Intelligent User Interfaces, IUI'14, Haifa, Israel, February 24-27, 2014* (2014), 235–240.
22. Ren, D., Hollerer, T., and Yuan, X. ivisdesigner: Expressive interactive design of information visualizations. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 2092–2101.
23. Schaffer, J., Abdelzaher, T., Jones, D., Hollerer, T., Gonzalez, C., Harman, J., and O'Donovan, J. Truth, lies, and data: Credibility representation in data analysis. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on*, IEEE (2014), 28–34.
24. Sinha, R., and Swearingen, K. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*, ACM (2002), 830–831.

25. Tintarev, N., and Masthoff, J. A survey of explanations in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, IEEE (2007), 801–810.

26. Tintarev, N., O'Donovan, J., Brusilovsky, P., Felfernig, A., Semeraro, G., and Lops, P. Recsys' 14 joint workshop on interfaces and human decision making for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*, ACM (2014), 383–384.

27. Yi, J. S., ah Kang, Y., Stasko, J. T., and Jacko, J. A. Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1224–1231.

28. Yi, J. S., Kang, Y.-a., Stasko, J. T., and Jacko, J. A. Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, ACM (2008), 4.

29. Zhang, J., and Pu, P. A comparative study of compound critique generation in conversational recommender systems. In *In Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006*, Springer (2006), 234–243.

30. Zhang, J., Wang, Y., and Vassileva, J. Socconnect: A personalized social network aggregator and recommender. *Information Processing and Management 49*, 3 (2013), 721–737.