EPIC Collab: Supporting Asynchronous Collaboration in Big Data Analysis Systems

Rsha Talal Mirza
Department of Computer Science
University of Colorado Boulder
Boulder, USA
rsha.mirza@colorado.edu

Kenneth M. Anderson
Department of Computer Science
University of Colorado Boulder
Boulder, USA
kena@cs.colorado.edu

Stephen Voida
Department of Information Science
University of Colorado Boulder
Boulder, USA
svoida@colorado.edu

***Abstract*—The rise of big data has led to the creation of large datasets that require teams to collaborate to analyze data effectively. Unfortunately, the software systems that collect and analyze large datasets are not often designed to support this kind of collaboration. Accordingly, our work investigates issues related to supporting collaboration in big data analysis systems. We use the domain of crisis informatics and the software infrastructure of Project EPIC as a case study to gain insight into the features that analysts need to effectively perform analysis at scale.**

This paper focuses on supporting asynchronous collaboration among analysts who work in small distributed teams on big data software systems. It describes the challenges faced by researchers who work collaboratively to analyze large crisis datasets (consisting, typically, of Twitter data). It then describes the work performed to redesign an existing big data analysis environment to substantially improve its support for collaboration. The impact of this research lies in its ability to improve the work of similar teams performing large-scale data analysis. While our work is based on insights gleaned from crisis informatics, we believe that our design, results, and lessons learned are broadly applicable to other application domains.

***Index Terms*—groupware, collaborative analysis, synchronous collaboration, big data, crisis informatics**

I. Introduction

Big data has received considerable attention in the last few years. Numerous studies have been conducted on extracting valuable information from the data, especially large social media datasets, using different analysis techniques and statistical methods (e.g., [1], [2]). Many big data software systems have been built to collect this kind of data to enable researchers to analyze data together (e.g., [3], [4]). Building such systems requires understanding the needs of analysts and knowing what kinds of work they need to perform on their data.

Project EPIC (Empowering the Public with Information in Crisis) is a research project that requires the use of data-intensive software systems [3], [5]–[7]. It allows analysts to collect and analyze large amounts of Twitter data to answer questions related to crisis informatics [7]. However, analysts still face many challenges when working on data analysis tasks together. They need to work on a system that better captures their real practice of crisis analysis as related to data sets containing tweets from Twitter. Additionally, due to the volume of data that analysts have to deal with, they need to work collaboratively around big data within the system.

A groupware perspective [8], [9] provides useful insights into the system design issues associated with supporting this kind of work. However, the heterogeneity and volume of the data that analysts need to share and collectively analyze in big data software systems differ significantly from the characteristics of many of the groupware systems presented in the literature. These big data analysis tools feature workspaces with millions of shared "documents" that need to be manipulated, sorted, and coded, and much of the focus of the work is asynchronous, due to the increased reliance on time-consuming, distributed computation for sifting through the large, heterogeneous sets of data. Moreover, the kinds of activities involved in the work of big data analysts are also different: Analysts need to coordinate their tasks around some research questions they need to investigate. They need to shift their attention from one activity to another to complete work faster since some jobs might be time consuming to complete. Therefore, they need to work with big data systems that have been designed to facilitate their collaborative work and serve their particular needs.

Accordingly, our goal is to provide analysts with a concrete implementation of a system that enables teams to collaboratively analyze big data from Twitter. We seek a better understanding of the following research questions:

- What challenges do researchers face when attempting to collaboratively analyze large data sets in the domain of crisis informatics? What features are needed to support collaborative work in this context?
- How can a big data analysis software system be designed to support crisis informatics research while supporting collaborative work by its users?

To provide insights into and a practical context for solving this problem, we first conducted an investigative user interview study to understand the needs of crisis informatics analysts. Then, we re-designed a key component of Project EPIC called *EPIC Analyze* to respond to these new requirements. Finally, we conducted a small-scale evaluation to assess the design of the new prototype. Our results confirmed that the new system enables better collaboration within teams who work on big data analysis for crisis informatics research and contributes to a broader understanding of how to support collaborative work within teams of analysts working on large data sets.

II. Background and Related Work

A. Big Data Analytics

Big data is defined by five characteristics: volume, velocity, variety [10], [11], veracity, and value [12], [13]. Big data analytics is the process of performing sophisticated analytical techniques on large sets of data [14]. The data analysis process generally involves collecting data from multiple resources, organizing it, and then analyzing it to produce valuable facts and figures [15]. Analysts working on big data analytics often start by applying statistical analyses to *explore* their data. Then, a wide range of expertise is required, such as machine learning, data mining, and information visualization, in order to *apply* algorithms to the data to obtain comprehensive results from the entirety of the data set. There are many useful systems and stand-alone tools that allow analysts to conduct these activities, including systems like Data Wrangler [16], Google Charts, and Open Heat Map. However, analysts face many challenges when working with big data, such as dealing with the growth, expansion, scale, and processing speed of the data and choosing the right data storage to store structured and unstructured data [12].

B. Crisis informatics Research

Crisis informatics examines "how information and communication technology is used in emergency response" [17]. It studies the socio-technical relationships among people, technology, and information during mass emergency or mass convergence events. During disaster events, social media platforms act as a tool to disseminate information to the public to send messages, post pictures and videos, and seek help [18]. Data generated on these platforms need to be collected and analyzed to better understand the interactions between information, technology, and people during crisis events.

General-purpose analysis tools like Jupyter Notebook[1] and Tableau [2] are frequently used by crisis informatics researchers to perform data analysis and visualization activities individually. However, these tools are neither well-suited for sifting through the very large volumes of data commonly generated during crisis situations, nor do they provide any significant capabilities for supporting synchronous or asynchronous collaboration. Other special-purpose tools such as Social Web Analysis Buddy (SWAB) and VizCept are used by researchers to analyze social media data. SWAB is a system that allows researchers to analyze Twitter datasets collaboratively with a focus on studying student-produced content on Twitter, while VizCept is a tool that is built to support synchronous data analysis between small teams in collaboration. Theses two systems offer useful services to analysts but they are more customized to other domains and they lack integration into the big data systems that analysts use to get data.

Project EPIC was founded in 2009 to help crisis informatics researchers perform studies during disasters or mass emergency events by providing a robust data collection and analysis platform. Project EPIC's software infrastructure was designed to collect a high volume of Twitter data generated by the public during disasters [3]. Over the years, Project EPIC has developed two primary data-intensive software systems—*EPIC Collect* and *EPIC Analyze*. *EPIC Collect* is a software system that collect billions of Twitter data across hundreds of mass emergency events. Tweets are collected and stored in a database based on sets of based on sets of keywords specified by crisis informatics analysts during a specific period, and these keyword lists are reviewed and updated when needed via a simple web application. *EPIC Analyze* is a web-based analysis software system that supports analysts in filtering and analyzing the collected data to answer questions related to crisis informatics. Analysts can review the list of existing Project EPIC datasets and see the keywords that were used to collect the tweets contained within. Analysts also can view the tweets contained in the dataset and perform various filtering, searching, and sorting operations. Additionally, *EPIC Analyze* offers an annotation interface that allows analysts to classify tweets, make comments on them, and save this information for future analysis (see Figure 1). As such, support for collaboration is only possible via these textual annotations, and this represents an impoverished mechanism for allowing team members to understand the work that is being or has been performed on the dataset, since annotations are connected to specific tweets and are not effective for communicating about larger, overarching analysis goals or the rationale behind a sequence of work activities.

C. Groupware and Awareness

The main goal of groupware is to enable collaboration between users in systems. Key to this goal is adding support for awareness. Awareness refers to "an understanding of the activities of others, which provides a context for your own activity" [19]. Awareness help to reduce group coordination efforts on tasks [20] and can support distributed work by helping groups to better communicate and interact with each other [19]. Studies have shown the importance of supporting awareness at multiple levels [21]: high-level awareness of other activities helps collaborators to coordinate their activities to avoid duplicated work and build upon previous results [22], whereas low-level awareness allows better work-sharing between participants.

Awareness has been categorized into many types, such as social awareness, workspace awareness, and activity awareness. Social awareness is about supporting the user's knowledge about other collaborators in a social or conversational context,

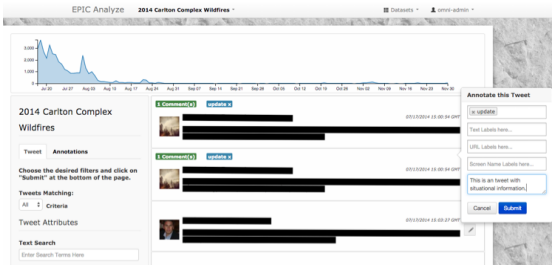

Fig. 1. *EPIC Analyze*: Annotation form

[1]https://jupyter.org
[2]https://www.tableau.com

such as these individuals' presence, level of engagement, interest level, and emotional state [23].

Workspace awareness is the knowledge that a user has about the actions of other collaborators (i.e., *who*, *what*, *when*, *where*, and *how*) in the shared workspace, in both the present and the past [24]. Other studies have expanded upon this theoretical framework to support asynchronous change awareness in workspaces by explaining how to support each of these elements in asynchronous environments [25]. Issues related to workspace awareness in CSCW systems have been well-studied, especially the problem of overloading groups with information about workspace activities (e.g., [26]). One solution to this problem is to provide users with a filtering profile for selection of awareness components. This solution improves awareness by producing a simple interface that is easy to configure to help users tackle the problems of awareness and information overload together, while increasing system usability.

Activity awareness is another key type of awareness that supports a group's knowledge about the past and present of interrelated activities. It is based on sharing activities of individual workspaces, not the shared workspace [27]. Supporting activity awareness helps users to better coordinate their activities with the group to achieve complex goals [28] and implies supporting both social awareness—see above—as well as action awareness [28]. Action awareness is about informing users about other collaborators' interactions with shared objects [28]. Other work in the literature has expanded that definition to cover understanding overall group activities that are performed to achieve larger, shared goals in collaboration. Activity, here, is defined as a sequence of actions that are performed toward reaching a group's shared goal [29]. Good notification systems should be designed to support exchange of activity awareness according to each individual's needs and preferences [28]. Implicit and explicit sharing of information between users is also important and has shown its value in helping groups to achieve their collaborative goals [30].

Several information visualization techniques allow users to visualize the histories of collaborative activities as a means of fostering activity awareness. One of the most common is the timeline chart: a graphical linear representation of the histories of activities. Timeline charts have been used to represent personal activities [31] and to represent a team's collaborative activities in synchronous [32], [33] and asynchronous modes [34]–[37]. The types of operations supported in these timelines include navigation, filtering, annotating, and exporting. The technique has been used to support different domains, ranging from data analysis [32], [36] to video editing [38] and software development [39] to education [33]–[35], [37].

III. Existing System and User Needs Analysis

Designing effective collaborative systems is challenging. Therefore, we conducted a small-scale interview study that contributes insights that address our first research question. The interview investigated the current practices and needs of Project EPIC's big-data analysts. It also asked them to brainstorm features that could better support their collaboration. The goal of this study was to determine the most-needed collaborative features to make analysis tasks in the presence of big data easier and faster.

The study consisted of a set of interviews to gather the user requirements for the design of a new version of *EPIC Analyze* that would provide explicit support for collaboration in crisis dataset analysis. The interviews were conducted with eight crisis informatics data analysts who all had experience working on a big data system before, during, or after crisis events. The participants came from four different contexts: two university research groups studying crisis informatics technologies, one large open-source software project that includes individuals focused on crisis response tool development, and an independent research institute investigating similar issues. Interviews were conducted in-person or via video conference. One analyst was excluded from our analysis after his interview revealed that he was not actively working within a group.

A. Findings

Current Practices and Analysis Workflow. We asked about the workflows that analysts follow and found that these practices vary—sometimes significantly—from group to group. However, they also share some commonalities: typically, workflows include obtaining the required datasets, dividing them, and performing basic descriptive analyses to reach results that guide teams to deeper and more nuanced analysis.

Analysts faced many challenges when following these practices. These challenges largely arise because *every analysis event is different*; workflows depend on the specific questions that need to be asked. It is challenging for teams to remember every analysis step they have performed and every decision they have made. Teams must also use numerous scripts with different datasets. These scripts and their associated datasets need to be organized so that teams know which program is related to which event, and which dataset is used with which script. Changing current workflows is difficult and that makes plugging queries and tools into their current pipeline difficult.

The second challenge is that *there is no specific tool available to streamline analysis work*. Teams need to use lots of different tools, which makes it difficult to keep track of all the past analyses that were performed across different tools.

The third challenge is related to *dataset versioning*. Sometimes, a member changes the underlying datasets, which creates confusion for other members working on the same data. Copying datasets from one place to another and transforming data from one format to another compound these problems. Analyst 5 relayed a story about an instance in which her team needed to extract some data from their SQL database to an Excel spreadsheet to do some data analysis and annotation. After adding their annotations to the data, when they tried to store the information back into the SQL database, they "suddenly found a lot of missing records in the database" due to changes committed by other team members. Additionally, maintaining updated datasets across all team members is difficult because each member often has his/her own copy of the dataset from which to work. For example, if a tweet is deleted from one ana-

lyst's copy, that change is not automatically propagated to different copies of the same datasets "owned" by other analysts.

The next challenge is related to *data collection*. Data are collected in many different formats, but many analysts are not equipped to handle this heterogeneity. It is also difficult to access historical data, which limits the types of questions that teams can ask.

Scalability is also another challenge. Datasets are big and that makes processing them very slow. In addition, if a dataset is too big, that means some types of algorithms cannot run on them locally, which means the data need to be uploaded to the cloud to use suitable software that can apply those algorithms at scale. Analyst 3 complained that it is difficult to get "simple answers" from the data without having to query the whole datasets, but also noted the much more significant time investment in running these full-scale queries.

The next challenge is that *there is no consistency* in data analysis work. Each analyst has his or her own way of working, a known characteristic of knowledge work [40]. This often means that there is no "information architecture" that governs local analysis processes and practices, e.g., how files should be named, "which makes a lot of confusion" [Analyst 5].

The last challenge is that there is always *redundant work* being performed across the team. Analysts end up writing numerous scripts to support their work, and some of these scripts have already been written by other team members. Redundant work also occurs when teams try to set bounds on the data; this means taking a large data set and filtering out information that does not fall within a particular set of days or within specific geographic bounds. As a result, portions of these filtered datasets end up being duplicated, muddying the boundaries of each analyst's work.

Duration of Analysis Tasks. Participants also talked about how hard it is to estimate analysis time due to the inability to predict inconsistencies in datasets, which may require cleaning, validation, identification of outliers, and dealing with time zone differences. One participant reported that it may take a month, another said it takes 6 to 8 weeks, and others reported that it can take up to two semesters (approximately nine months) to complete the work.

Collaborative Tasks. The participants reported needing to collaborate and coordinate with each other when working on early tasks of data analysis—sharing general ideas and goals, answering initial questions related to datasets, dividing tasks among members of the team, and setting up standards to follow. Participants also discussed collaborating when collecting datasets from different sources, annotating tweets, and mining datasets. They also have to work together when visualizing networks and when writing papers to publish their results. In addition, they commonly share their outcomes: analysis results, figures, pictures, and statistics. They also sometimes need to assign analysis tasks that require stronger programming skills to other team members.

Other Data Analysis Programs and Tools. Numerous different programs and tools are used in analysis tasks. Some are used for collecting data, such as the Twitter API, the OpenStreetMap API, and the Overpass API. Others are used for communication, such as Slack and email, or they are used for visualization, scripting, data sharing, and data storage.

Requested Features to Support Collaboration. We asked participants to brainstorm new collaboration services for big data software systems. Some of the (many) ideas included:

- *Data visualization tools*, including overviews of tweets' geographic distribution and overall dataset statistics.
- *A unified analysis environment.* One participant pointed out the importance of a single platform as a way to minimize software configuration overhead. Another said that one workplace, with all collaboration services integrated as embedded services (in contrast to a suite of stand-alone tools) would improve workflow and productivity.
- *Shared storage, history, and provenance tools*. Suggestions included a shared common repository on the server to keep all files and documents related to each event in one place; a common place to document all things related to each dataset; an ability to track the transformation of datasets and to store all dataset versions; and the ability to save all the steps that have been taken by team members when they analyze and mine datasets, so that teams can easily write papers that describe their collective work.
- Most participants requested *notifications of colleagues' activity*. One participant said, "I want to be notified when teams add, modify, or delete keywords from events." Another reported wanting to see pop up information with links to relevant changes. Another requested the ability to track the activities of team members to keep him up-to-date with what has been done without bothering his team every time with many questions. A fourth suggested having access to the progress of team members when they work on the same dataset and to peek in on the results that they obtained from running scripts on that dataset.
- Other suggestions included the *integration of more traditional groupware tools*, including a chat system or a digital whiteboard for sharing ideas within the team.

Despite frequent suggestions to capture and share activity history, some participants expressed concerns about documenting analysis work. One said, "documenting is good, but no one is doing it." Another wondered how the system could encourage data analysts to better document their work.

Reflecting on this set of user requirements, we decided to focus on the design and implementation of asynchronous collaborative features, since these captured the largest unexplored part of the tool space and best aligned with known benefits of groupware systems. Furthermore, the suggestion of adopting one unified environment on a server to embody all of a team's work is already instantiated with *EPIC Analyze*, albeit with only rudimentary collaboration support tools. The feature of storing all dataset versions for events, especially the original dataset, was partially supported in *EPIC Analyze*, but not front-and-center in the interface. The system also allowed users to create a sample or template from a dataset, but that sample was not available for viewing or re-use by other team members.

B. Summary of User Requirements

In addition to our initial interviews, members of our research team had previously attended to observe many meetings with crisis informatics researchers and reviewed many scientific papers that enumerate common analysis processes for crisis events [1], [6], [7], [41], [42]. As a results, We settled on carefully redesigning *EPIC Analyze* to meet the following criteria:

- The system should provide users with the *core and specific analysis features* that serve their needs in this domain. That can be achieved by supporting, at a minimum, a typical analysis workflow, which includes obtaining a dataset that is collected during a crisis event, constructing one or more crisis informatics research questions, collaboratively analyzing the data by performing sequences of operations on the datasets to generate results and, finally, reporting these results to answer the questions.
- However, since every analysis task for disasters is different, and the workflow needs to be changed according to the questions that need to be answered, the new system should be *flexible* to support different analysis workflows.
- The system should allow teams to cooperate on analysis work—for example, checking the work of a team member and being able to extend prior work; that is, allowing individuals to reuse datasets and sequences of analysis activities from a previous crisis event in new contexts. These cooperative features are key for reducing analysts' time and effort and help to reduce redundant work.
- Documenting is a tedious task for analysts who want to focus on analysis and not on generating metadata about an analysis. Therefore, the system should *support activity awareness* while *minimizing users' workload*. The system should capture all team interactions within the system and provide a *simple, interactive visualization of the interaction history* to better support collaborative work.

IV. Designing Awareness Into Collaborative, Big-Data Analysis Systems: *EPIC Collab*

Here, we present the design of an all-new version of EPIC Analyze, a system that we now call *EPIC Collab*. *EPIC Collab* contributes insights to address our second research question. A key aspect of the new design is that it is centered around analysts' research questions to better align the tool with how analysts work. Analysts can use the system to document their research goals, and the system tracks all subsequent work that is carried out to answer each question. As a result, the interface is organized using three tabs that reflect different facets of the current crisis event: *Research Questions*, *Twitter Datasets*, and *Team Contributions*. The *Research Questions* tab shows information about the research question(s) that the team has added to this event (Figure 2). There are two ways to review these questions. The first displays all research questions in a table format, and, from this interface, a user can add a new question to an event or bring a question that was added to another event into this event using the *Import* button.

When the user clicks on any question, detailed information about the question is displayed. A research question can be part of another research question; this allows analysts to group their work into more and more finely-grained questions that are each more straightforward to answer. An editing history on the right side of the window shows who created, imported, and/or updated each research question, information that supports the awareness of others' actions performed on various research questions. This detail view also informs the user about the sequence of actions that occurred to change the research question (*How* category), the type of actions performed on the question (*What* category), the users who performed those actions (*Who* category), and when these actions were performed (*When* category) (Figure 3).

A second aspect of this tab displays the research questions as a tree to allow users to visualize the hierarchical structure of all research questions. This interface supports the awareness of activities by helping the user to understand the overall activities of the team on the research question objects (Figure 4).

The second tab, *Twitter Datasets*, shows the collated twitter datasets associated with the event (Figure 5). After opening a dataset, a user can first select a research question to work on, and then perform different analysis activities on the corresponding tweets dataset. These activities include: search; comment; annotate; save the results of a query as a new dataset; export the dataset; and find the lists of top 10 tweets or users who have the highest favorite, retweet, and follower counts.

The third tab, *Team Contributions*, shows information about the team and their activities in analyzing the event. There are two interfaces in the tab. The first shows a list of team members working the event and their respective roles; this interface is intended to facilitate social awareness. The second interface shows all activities that have been performed by the team over time (Figure 6). This "timeline" view is intended to provide awareness of all activities performed by members of the team on all of the event's datasets. It can also be used to determine the sequence of analysis actions performed over time by a specific user on a set of objects (tweets) within a dataset.

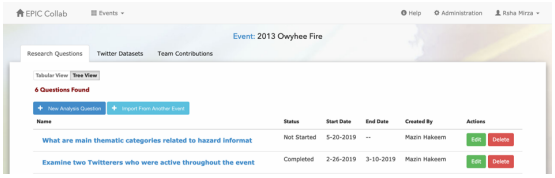

Fig. 2. *EPIC Collab*: Research Questions tabular interface

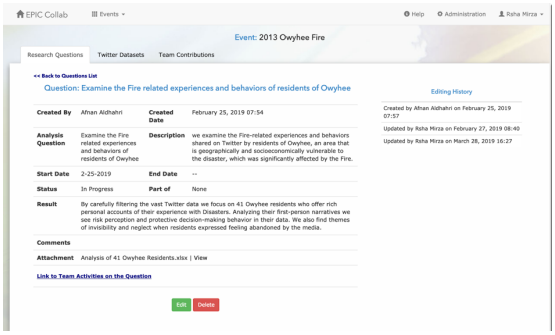

Fig. 3. *EPIC Collab*: Research Question information page

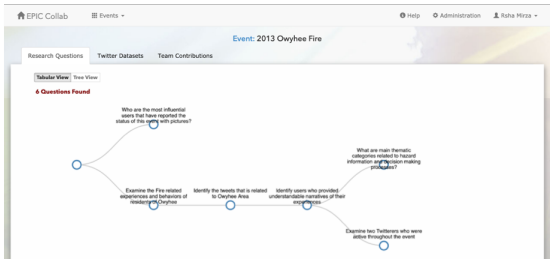

Fig. 4. *EPIC Collab*: Research Questions tree interface

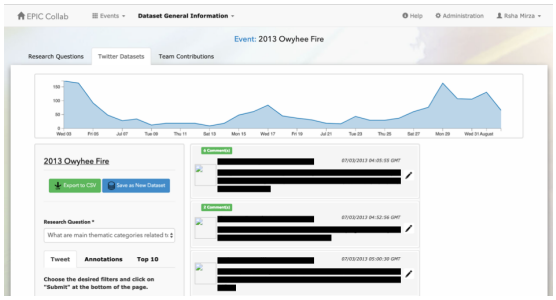

Fig. 5. *EPIC Collab*: Tweets dataset interface

Fig. 6. *EPIC Collab*: Timeline-based Team Activities interface

The timeline shows the activities performed by each team member over the course of the analysis. These activities are color-coded, with each color representing one research question. The user can hover over the question number to see the actual question and can click on the question to reveal more detailed information. The date range of the timeline activities is displayed at the top of the timeline. The left side of the interface shows the number of activities presented in the timeline. Under that, there is a legend that shows the graphical icons that represent the type of activities presented on the timeline. The user can zoom in or out to obtain more insight or context into the collaboration. The user also can drag the bottom gray rectangle to scroll horizontally—through time—on the timeline.

When a user hovers over an activity icon in the timeline, more information is shown in a tooltip (Figure 7). This information aims to support action awareness in the system. It informs the user of five elements of information about the action:

- *What* category: *What is the action performed?*
- *When* category: *When was the action performed?*
- *Who* category: *Who performed the action?*
- *Where* category: *Upon what objects was the action performed?*
- *How* category: *How did the action occur?* or *What query was used to perform the action?*

In addition, the user also can click on any activity icon to see its result. The *X* icons on the activities are only displayed on those activities that were performed by the current user; the user can click on this icon to delete any activity (removing it from all users' timeline views). This feature allows the user to perform exploratory analysis on a dataset and then delete it if it does not return desired or useful results.

The system also allows the user to build upon previous work performed by another team member. This can be done by forking from an activity in the timeline. To do that, the user can select an activity in the timeline to view its result in a read-only mode, allowing the user to see the results of the prior query. If the user chooses to use this query as a "jumping-off" point, she can click on the *Edit* button to continue working from that set of results. In the edit mode, a notification confirms that the activity has been forked, and all editing and analysis buttons are enabled. If the user goes back to the timeline from this point, then she can see that there is a new arrow drawn linking the original activity with the new activity, representing the fork.

The timeline might seem overwhelming because of the detailed awareness information that it provides. Therefore, a filter mechanism is included on the bottom of the timeline to allow the user to display only the information about which she is interested. This mechanism allows the user to filter the activities on the timeline by one of three categories: research question, dataset used, and activity type. Once the user selects (or deselects) a filter, the timeline is updated automatically.

The last aspect of the system specifically designed to support cooperative work is a feature that enables the user to apply previous sequences of work to a new analysis context. The system provides the user with the capability to copy all activities that were performed to answer a research question in one crisis event into another event. To use this feature, the user starts in the current event—the one *into* which she wants to import the activities. When she clicks on the *Import from Another Event* button on the *Research Questions* tab, she is prompted to select an event to import *from*, as well as the research question she wants to replicate (Figure 8). The *Import* action imports the selected question (as well as any sub-questions(s)) from the other event, any dataset(s) created, and all team activities performed to answer the question. Since each crisis event is associated with a different set of collected tweets, the import action does not import the comments or tweet annotations, because these two activities are linked to

Fig. 7. *EPIC Collab*: Information revealed by hovering over timeline activities

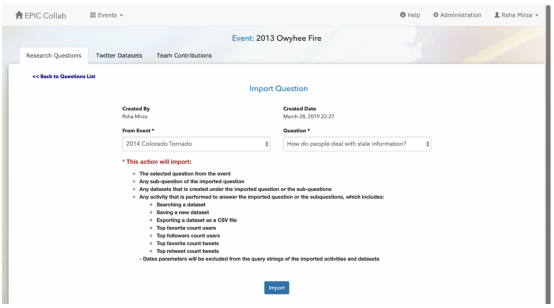

Fig. 8. *EPIC Collab*: Import research question Interface

specific tweets' IDs in the first dataset, and those tweets' IDs might not exist in the second. Further, since each crisis event has different start and end dates, all date-related parameters are excluded from query strings used in imported activities and to generate imported datasets. The *import* action adds the imported research question and its sub-question(s) to the *Research Questions* tab. It also adds the imported datasets to the *Twitter Datasets* tab and all of the imported activities to the third tab, *Team Contributions*. These activities are added to the timeline of the user who performed the *import* action.

V. Evaluation

We conducted an informal task-based evaluation of the system using experienced crisis informatics researchers to assess the design of our newly integrated collaborative features and to evaluate how well the system supports users in collaboratively analyzing big data. In this case, we tested the collaborative features by evaluating the degree to which the system provided social and activity awareness. We assessed how well the system helped the user to be aware of who can contributed to the analysis of a specific crisis event are and what their roles are/were. We also evaluated how well the users could identify what activities their team members had carried out, to date, and what the states of activity-relevant shared objects are. Finally, we studied how well users could take advantage of the cooperative work functionalities described in the previous section.

1) Participants: The usability study was conducted with five crisis informatics researchers broadly affiliated with our research group (but not directly involved in the design of the *EPIC Collab* system) who had prior experience working to collaborative analyze large datasets of crisis events. This number of participants is often a sufficient number for a usability study to evaluate a system in its early stages according to Nielsen's discount usability method [43], and subject-matter experts like those we recruited for this study can provide qualitative findings necessary to help us assess and further refine the system design prior to iteration and large-scale deployment.

2) Study Design and Procedure: The study involved a one-hour laboratory session. It consisted of three methods to collect data from the participants: a pre-test questionnaire, a set of tasks to be performed on the system by the participant, and a post-test questionnaire. First, the pre-test questionnaire collected participants' demographic information, which we use to describe the participants in aggregate. Then, a recorded video of the system was played to the participants to demonstrate its main features. After that, a set of task scenarios was used to evaluate the system's support for collaboration; we asked participants to perform the Concurrent Think-Aloud moderating technique while performing a sequence of specified tasks:

- **Activity awareness** (15 tasks):
  - **Social awareness**: Who are the team members and what are their roles? (1 task)
  - **Action awareness**: What are the states of activity-relevant objects? (10 tasks—5 in the *What* category, 1 in the *When* category, 2 in the *Who* category, and 1 each in the *Where* and *How* categories)
  - **Activity history**: What are the sequences of actions performed over time? (2 tasks)
  - **Overall team activities**: What are the actions performed by the whole team? (2 tasks)
- **Cooperation** (3 tasks):
  - Building upon the work performed by another team member (2 tasks)
  - Copying the work performed by another team member (1 task)

Finally, we delivered a post-test questionnaire to obtain participants' qualitative and quantitative feedback about the system's features and to determine participants' overall satisfaction with the system. Participants were asked a variety of subjective questions to gain insights about the effectiveness of our new features. Participants were also asked to complete the CSUQ computer system usability questionnaire, a validated, quantitative measure of perceived system usability [44].

3) Findings: ***Pre-test Questionnaire.*** The pre-test Questionnaire showed that the age of the participants ranged from 18 to 46 years or older. All of whom had more than three years of experience as big-data analysts. Their levels of proficiency ranged from elementary to proficient. Four were female, and four were employed full-time. The analysis tools they used varied from basic tools, such as Excel, Google Drive, and email, to more advanced tools, such as Pandas, Jupyter, and Spark.

Testing Tasks. The participants were asked to perform 18 testing tasks using the system. 15 tasks were successfully completed by all participants. Only three of the tasks—one task asking participants to determine *who* had completed an activity, one task asking participants to perform an *Import* task, and one task asking participants to identify what activities had previously been accomplished in an imported activity—were not successfully completed by some of the participants. The primary reasons for these task failures included poor visibility of some interface components; participants not recognizing that various aspects of the interface were, in fact, clickable links; one participant's conceptual misunderstanding of how the *Import* feature worked; and a couple of instances of misunderstood task instructions. Overall, the usability problems uncovered by the tasks were all relatively minor and could largely be addressed with minor interface redesigns.

Post-test Questionnaire. Participants provided a wealth of information in the post-test questionnaire about the new collaborative features integrated into the system. The data from

the questionnaire is presented below, organized by broad categories of feedback:

Overall Experience. We were pleased that "very excellent", "amazing", "cool", "fascinating", "powerful", and "WOW" were among the participants' responses when asked to describe their overall experience using *EPIC Collab*. All participants expressed strong, positive feelings about the system. They found the system to be intuitive to operate, easy to use and learn, and useful. In addition, one of the participants was very excited and said that she would love to get a copy of the system to work on her data as soon as possible.

Specific Positive Aspects of the System. Participants were asked to list five aspects of the system they felt most positive about. All participants liked the timeline-based *Team Contributions* interface. Three mentioned the *import* feature, discussed in detail above. Two noted the tree-based, hierarchical depiction of research questions. Other aspects of the system praised by participants included the clear representation of activities, the "helpful" icons displayed on the timeline, the ability to see the results of other members' activities, and the ability to fork previous activities as a starting point for new data explorations.

Negative Aspects of the System. The participants were also asked to list five aspects of the system that they were less enthusiastic about. One participant reported that the *X* icon, which is displayed in the timeline activity, was very sensitive and did not like the fact that there were no images or media displayed in the tweet dataset interface. Another participant mentioned that the system did not support filtering the timeline chart by team member, and was disappointed that the system did not support exporting the annotations and comments with datasets. The last participant disliked that the system forced the user to articulate a clear research question before allowing him to complete any analysis steps. He was also frustrated that the URL links in the tweets dataset interface were not clickable.

Type and Amount of Information. Participants were asked whether the information collected and presented to them about team activities would be beneficial for collaboration and cooperation, and all agreed with this statement. One liked the ability to check in on the results of others' activities. Another participant appreciated being able to hover over the activity icons to quickly retrieve associated activity metadata. A third participant found the information useful because it allowed them to not duplicate the work of or the research questions explored by others. However, one participant felt that it would be valuable to be able to discount or remove activities from some users, especially if they are just learning.

Missing Information. The participants made it clear that additional features would likely enhance their ability to coordinate and collaborate within their teams. One participant wanted to be able to easily find the description of a new dataset, especially if it was created using a complicated query. Another wanted to be able to comment on the timeline's activities, bookmark some of them, and be able to view them later. This participant also requested a to-do list feature that would enable task assignments for individuals and the team, as a whole. A third participant expressed interest in collapsing timeline activities to focus on entire analysis sessions; he felt that this would help him get an overall view of what people were doing.

Confusing Information. Participants were asked whether any part of the information presented in the system was confusing or difficult to understand. One of them responded that the import action was very powerful, but that some of its effects were hard to anticipate or understand. Another found the dates displayed in the timeline charts to be confusing. One participant was expecting the graph at the top of the *Tweets Dataset interface* showing the distribution of tweets over time to automatically refresh as other information on the page changed. Another participant felt that the click-ability of the tree-based hierarchy of research questions was not intuitive.

Other Comments. A variety of additional features were suggested for the system, including being able to: quickly sort tweets in the tweets dataset interface; export a list of all posters, with or without their post content; export annotations and comments added to tweet datasets as a CSV file; add an analysis action that computes and returns a list of the social media users with the largest number of replies; increase and decrease the font size in the tweet results to skim through more tweets quickly; use more distinguishable colors for the research questions; link names of team members in the *Team Member* interface with their activities in the timeline, and integrate maps to display the tweets' geo-information.

CSUQ Evaluation. Participants' quantitative evaluations of the system are shown in Figure 9. The average scores are calculated by adding all participants' scores together for each statement and then dividing these numbers by the total number of answers obtained for each statement. Low scores are better than high scores, due to the anchors used in the 7-point scales.

VI. Discussion

Our evaluation of the *EPIC Collab* system gave us an initial assessment of how successful we were in creating a big data analysis software system to support collaborative crisis informatics research. In general, our subject matter experts agreed that we accomplished many of our goals. Indeed, the system usability ratings provided by participants indicate a positive assessment of our design solutions and resonance with the particular kinds of collaboration support essential to and expected in this kind of work.

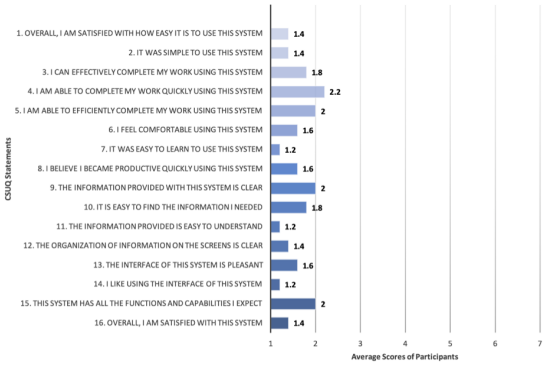

Fig. 9. Average responses to the individual CSUQ items asked during the usability testing study (lower scores are better)

We acknowledge a number of opportunities for addressing usability shortcomings with *EPIC Collab*, particularly in those areas in which our participants did not successfully complete their prescribed tasks (e.g., discoverability problems with the links in the *Research Questions* interfaces, drawing better attention to the editing history information in these pages, and better guidance around and explanations about the process for and effects of activity importing) and for which multiple instances of qualitative subjective feedback pointed out weaknesses. The other potential avenues for improvement can be clustered into two broad categories of novel features: *collaborative features* and *analysis features*. The most-requested collaborative features include the ability to comment on the timeline activities, bookmarks for team activities for later viewing or re-visitation, a calendar with deadlines, a to-do list for individuals and the team, a progress notes feature, a chat system, and information about what other members are currently working on in order to prevent activity overlap and analytic redundancy. Recommended additional analysis features include adding support for inline coding of tweets; adding an analysis action that returns the list of "top" tweeters, those with the highest reply count; a lightweight way to create new crisis events in the system; and the ability to export annotations and comments with datasets, the list of top tweet producers, and the list of all users, with and without their tweets' content.

The data collected as part of our study can be used not only to improve our own system in future iterations, but also enable us to reflect on how our empirically-informed user requirements for supporting collaboration in big data analysis platforms might be applied to other large-scale analysis contexts.

EPIC Collab was designed to provide users with both *core and domain-specific collaboration features*. We focused our development on supporting the most common data analysis and transformation activities. These activities differ from those found in many other groupware systems because they are much more reliant upon facilitating asynchronous collaboration (necessary because of the distributed computation necessary for managing the volume of big data) and ensuring alignment among team members with respect to *process*, as opposed to detailed information about specific information artifacts.

We also created a system that was *flexible* in supporting different analysis workflows and—particularly with the timeline visualization of team members' activity histories—allowed team members to gain insights into how others conducted their analyses. We demonstrated that our timeline and hierarchical research question designs were usable and effective in supporting the activity awareness of analysis workflow in crisis informatics—again, pulling the focus away from annotations on specific information items (i.e., tweets) and onto the broader arc of analysis. This is one of the more important takeaways of our research—differentiating between supporting collaboration that is focused on the artifacts, themselves, and providing tools that focus on sharing a higher-level overview of *process* and *questions explored*. We expect that these design approaches can also be used to represent collaborative analysis workflows in other domains. Both interfaces reflect the overarching structure of big-data analytic practice, but also respect individual differences in carrying out those analyses. Both also render these individual differences visible for accountability and verification purposes across the team.

EPIC Collab provides teams with cooperative features—for example, checking in on the work of a team member or extending prior work via our *import* mechanism. Paying specific attention to those tasks and analytic practices that are error-prone, repetitive, or that introduce the possibility of redundant or unneeded work can serve as important resources for design. The popular computer science adage of "make the common case fast" is important here, but so is the idea that collaborative systems like these should reduce the overhead of computationally-expensive, time-intensive, or collaboratively complicated activities as much as possible. This is an important point, especially in big data contexts in which the volume, velocity, variety [10], [11], veracity, and value [12], [13] of the analytic focus make even common tasks more difficult to carry out without this kind of dedicated tool support.

Finally, we focused on *supporting activity awareness* while *minimizing the workload* for *EPIC Collab*'s intended users. Our participants were quite frankly excited about being able to visually explore the history and evolution of the analysis that they had (hypothetically) undertaken previously, as well having access to their colleagues' trajectories of work. Thinking about ways in which automatic activity detection and logging, implicit social awareness tracking and sharing, and activity sequence re-use capabilities can be incorporated into future systems all seem like relatively low-cost and high-effectiveness mechanisms for amplifying the cognitive and social cognition capabilities of the skilled analysts who are tasked with carrying out this kind of work.

VII. Conclusion

In this article, we explored fundamental issues of collaboration in big data analysis software systems. We focused on supporting asynchronous collaboration among analysts who work within small distributed teams on big data software systems; in this case, a crisis informatics research platform. We present the empirical work that we did to understand the challenges faced by researchers who work on collaborative big data analysis tasks in the crisis informatics domain, as well as these individuals' collaborative needs when working within these systems. This research led us to identify a series of user requirements for this domain and to design, implement, and evaluate a collaborative big-data analysis system, *EPIC Collab*. The results of our evaluation confirmed that *EPIC Collab*'s collaboration features are useful for teams who perform big data analysis for crisis informatics research, and our research contributes to a broader understanding of how to support collaborative work within similar teams of analysts who work on large data sets.

One aspect of crisis informatics research is that it often deals with an intense objects of study—natural disasters with heavy damage and large loss of life—with tight analysis timelines. From supporting that domain, Project EPIC and this research, resulting in the *EPIC Collab* system, have revealed much about

what analysts need when working together in this extreme environment. This knowledge will be useful as others take the lessons learned from our fieldwork and the features of our system to other domains and their associated analysis tasks.

REFERENCES

[1] L. Palen, K. M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald, "A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters," in *Proc. ACM-BCS Visions of Comput. Sci. Conf.* Brit. Comput. Soc., 2010, p. 8.

[2] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, "Analyzing (social media) networks with nodexl," in *Proc. of the fourth Int. Conf. on Communities and technologies*. ACM, 2009, pp. 255–264.

[3] K. M. Anderson and A. Schram, "Design and implementation of a data analytics infrastructure in support of crisis informatics research (nier track)," in *Proc. of the 33rd Int. Conf. on Software Engineering*, ser. ICSE '11. ACM, 2011, pp. 844–847.

[4] X. Chen, K. Madhavan, and M. Vorvoreanu, "A web-based tool for collaborative social media data analysis," in *2013 Int. Conf. on Cloud and Green Computing*. IEEE, 2013, pp. 383–388.

[5] A. Schram and K. M. Anderson, "Mysql to nosql: Data modeling challenges in supporting scalability," in *Proc. of the 3rd annual Conf. on Syst., programming, and applications: software for humanity*. ACM, 2012, pp. 191–202.

[6] K. M. Anderson, A. Schram, A. Alzabarah, and L. Palen, "Architectural implications of social media analytics in support of crisis informatics research." *IEEE Data Eng. Bull.*, vol. 36, no. 3, pp. 13–20, 2013.

[7] K. M. Anderson, A. A. Aydin, M. Barrenechea, A. Cardenas, M. Hakeem, and S. Jambi, "Design challenges/solutions for environments supporting the analysis of social media data in crisis informatics research," in *48th Hawaii Int. Conf. on System Sciences*. IEEE, 2015, pp. 163–172.

[8] S. Greenberg, "Computer-supported cooperative work and groupware: an introduction to the special issues," *Int. J. of Man-Machine Studies*, vol. 34, no. 2, pp. 133–141, 1991.

[9] J. Xu, J. Zhang, T. Harvey, and J. Young, "A survey of asynchronous collaboration tools," *Inf. Technol. J.*, vol. 7, no. 8, pp. 1182–1187, 2008.

[10] N. Mohamed and J. Al-Jaroodi, "Real-time big data analytics: Applications and challenges," in *2014 Int. Conf. on high performance computing & simulation (HPCS)*. IEEE, 2014, pp. 305–310.

[11] K. Park, M. C. Nguyen, and H. Won, "Web-based collaborative big data analytics on big data as a service platform," in *2015 17th Int. Conf. on Advanced Communication Technol. (ICACT)*. IEEE, 2015, pp. 564–567.

[12] P. Chandarana and M. Vijayalakshmi, "Big data analytics frameworks," in *2014 Int. Conf. on Circuits, Syst., Communication and Inf. Technol. Applications (CSCITA)*. IEEE, 2014, pp. 430–434.

[13] S. Sharma and V. Mangat, "Technology and trends to handle big data: Survey," in *Proc. of the 2015 Fifth Int. Conf. on Advanced Computing & Communication Technologies*, ser. ACCT '15. IEEE Comput. Soc., 2015, pp. 266–271. [Online]. Available: http://dx.doi.org/10.1109/ACCT.2015.121

[14] P. Russom *et al.*, "Big data analytics," *TDWI best practices report, fourth quarter*, vol. 19, no. 4, pp. 1–34, 2011.

[15] A. G. Shoro and T. R. Soomro, "Big data analysis: Apache spark perspective," *Global J. of Comput. Sci. and Technol.*, 2015.

[16] J. Heer and S. Kandel, "Interactive analysis of big data," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 19, no. 1, pp. 50–54, 2012.

[17] L. Palen and S. B. Liu, "Citizen communications in crisis: anticipating a future of ict-supported public participation," in *Proc. of the SIGCHI Conf. on Human factors in computing Syst.* ACM, 2007, pp. 727–736.

[18] C. Hagar, "Crisis informatics: Perspectives of trust–is social media a mixed blessing?" *School of Inf. Student Research J.*, vol. 2, no. 2, 2013.

[19] P. Dourish and S. A. Bly, "Portholes: Supporting awareness in a distributed work group," in *Proc. of the CHI '92 Conf. on Human Factors in Computing Systems*, 1992.

[20] C. Shah, "Effects of awareness on coordination in collaborative information seeking," *J. of the American Soc. for Inf. Sci. and Technol.*, vol. 64, no. 6, pp. 1122–1143, 2013.

[21] P. Dourish and V. Bellotti, "Awareness and coordination in shared workspaces." in *CSCW*, vol. 92, no. 3, 1992, pp. 107–114.

[22] J. Heer and M. Agrawala, "Design considerations for collaborative visual analytics," *Inf. visualization*, vol. 7, no. 1, pp. 49–62, 2008.

[23] S. Greenberg, C. Gutwin, and A. Cockburn, "Awareness through fisheye views in relaxed-wysiwis groupware," in *Graphics Interface*, vol. 96, 1996, pp. 28–38.

[24] C. Gutwin, S. Greenberg, and M. Roseman, "Workspace awareness in real-time distributed groupware: Framework, widgets, and evaluation," in *People and Computers XI*. Springer, 1996, pp. 281–298.

[25] J. Tam and S. Greenberg, "A framework for asynchronous change awareness in collaborative documents and workspaces," *Int. J. of Human-Computer Studies*, vol. 64, no. 7, pp. 583–598, 2006.

[26] J. M. N. David and M. R. Borges, "Selectivity of awareness components in asynchronous cscw environments," in *Proc. Seventh Int. Workshop on Groupware. CRIWG 2001*. IEEE, 2001, pp. 115–124.

[27] K. Hayashi, T. Hazama, T. Nomura, T. Yamada, and S. Gudmundson, "Activity awareness: a framework for sharing knowledge of people, projects, and places," in *Proc. ECSCW '99*. Springer, 1999, pp. 99–118.

[28] J. M. Carroll, D. C. Neale, P. L. Isenhour, M. B. Rosson, and D. S. McCrickard, "Notification and awareness: synchronizing task-oriented collaborative activity," *Int. J. of Human-Computer Studies*, vol. 58, no. 5, pp. 605–632, 2003.

[29] G. Convertino, D. C. Neale, L. Hobby, J. M. Carroll, and M. B. Rosson, "A laboratory method for studying activity awareness," in *Proc. of the third Nordic Conf. on HCI*. ACM, 2004, pp. 313–322.

[30] N. Goyal, G. Leshed, D. Cosley, and S. R. Fussell, "Effects of implicit sharing in collaborative analysis," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Syst.* ACM, 2014, pp. 129–138.

[31] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, "Lifelines: Visualizing personal histories," in *Proc. CHI '96*. New York: ACM, 1996, p. 221–227.

[32] H. Chung, S. Yang, N. Massjouni, C. Andrews, R. Kanna, and C. North, "Vizcept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis," in *2010 IEEE Symp. on Visual Analytics Sci. and Technol.* IEEE, 2010, pp. 107–114.

[33] T. Vasileva, V. Tchoumatchenko, M. Lakkala, and K. Kosoncn, "Infrastructure supporting collaborative project based learning in engineering education," *Int. J. of Engineering Education*, vol. 27, no. 3, p. 656, 2011.

[34] C. Plaisant, A. Rose, G. Rubloff, R. Salter, and B. Shneiderman, "The design of history mechanisms and their use in collaborative educational simulations," in *Proc. of the 1999 Conf. on Comput. support for collaborative learning*. Int. Soc. of the Learning Sciences, 1999, p. 44.

[35] C. H. Ganoe, J. P. Somervell, D. C. Neale, P. L. Isenhour, J. M. Carroll, M. B. Rosson, and D. S. McCrickard, "Classroom bridge: using collaborative public and desktop timelines to support activity awareness," in *Proc. of the 16th annual ACM Symp. on User interface software and Technol.* ACM, 2003, pp. 21–30.

[36] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," *IEEE Trans. on visualization and Comput. graphics*, vol. 14, no. 6, pp. 1189–1196, 2008.

[37] H. Hornung and M. C. C. Baranauskas, "Timelines as mediators of lifelong learning processes," in *Proc. 11th Brazilian Symp. on Human Factors in Computing Syst.* Brazilian Comput. Soc., 2012, pp. 99–108.

[38] T. Grossman, J. Matejka, and G. Fitzmaurice, "Chronicle: capture, exploration, and playback of document workflow histories," in *Proc. of the 23nd annual ACM Symp. on User interface software and Technol.* ACM, 2010, pp. 143–152.

[39] F. C. Ribeiro, J. M. de Souza, and M. M. de Paula, "Use of information visualization techniques in a collaborative context," in *2015 IEEE 19th Int. Conf. on CSCW in Design (CSCWD)*. IEEE, 2015, pp. 79–84.

[40] A. Kidd, "The marks are on the knowledge worker," in *Proc. of the SIGCHI Conf. on Human factors in computing Syst.: celebrating interdependence*. ACM, 1994, pp. 186–191.

[41] L. Palen and S. Vieweg, "The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat," in *Proc. of the 2008 ACM Conf. on CSCW*. ACM, 2008, pp. 117–126.

[42] L. Palen, S. Vieweg, and K. M. Anderson, "Supporting "everyday analysts" in safety-and time-critical situations," *The Inf. Soc.*, vol. 27, no. 1, pp. 52–62, 2011.

[43] J. Nielsen, "Usability engineering at a discount," in *Proc. 3rd Int. Conf. on HCI on Designing and using human-computer interfaces and knowledge based Syst. (2nd ed.)*. Elsevier Sci. Inc., 1989, pp. 394–401.

[44] J. R. Lewis, "Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *Int. J. of HCI*, vol. 7, no. 1, pp. 57–78, 1995.