**Analytic Provenance Dataset**

**Sina Mohseni, Eric D. Ragan**

We conducted a series of user studies involving an exploratory data analysis scenario with textual data. For the analysis scenarios, we selected data analysis tasks with sufficient complexity and scope to allow the exploration of various topics and hypotheses.

To this end, we chose three analysis scenarios from the IEEE VAST Challenge datasets [] More specifically, our case study used the following data sets: 2010 mini-challenge \#1, 2011 mini-challenge \#3, and 2014 mini-challenge \#1. Each of these data sets is a synthetic set of text documents about intelligence analysis scenarios designed to be close to real world scenarios. Examples of documents included paper articles, email intercepts, telephone calls intercepts, and web blog posts relevant to the given scenario. Text documents varied in length from single sentences up to multiple paragraphs.

While all of the data was in plain-text format, some of the documents primarily consisted of numerical data related to financial transactions. To accommodate practical time constraints for user studies while still allowing appropriate complexity for exploratory analysis, the 2011 and 2014 data sets were limited to 152 documents for participant analysis sessions, and 2010 data set had 102 documents and did not need limiting.

To analyze the text data, participants use a document exploration application where text documents could be viewed, searched, and spatially manipulated to support organization. In this application, documents are placed in a 2D space. Figure (1) shows a participant working with the exploration tool.

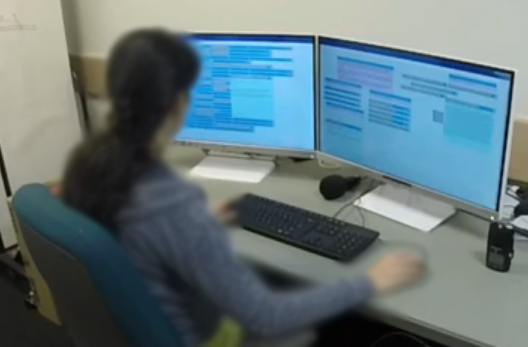


Figure 2 Study setup

We opted for a basic document exploration tool because of its easily-understandable design and its similarity to many real tools and workspaces used to work with documents and files.

When loading the data in our document explorer tool, each document starts as collapsed with only its title visible. Users could “open” any document by double clicking the title bar or by clicking a dedicated button on the document's title bar, and this would expand the document to a window containing the text of the document. The document could be collapsed back to the title in the same way. Within an open document, users could highlight text by selecting it, right clicking, and activating a menu item. When a window has highlighted text, the window could be “reduced to highlight”, which would hide all text in the document except for the highlighted content. Figure 2 shows the document explorer tool.

At the beginning of the study documents were arranged in left display not having any specific order or grouping. Users clicked and dragged documents, freely re-arranging texts in their workspaces. They could also create editable notes windows in the same workspace.

When using the “search” functionality, both matching words within windows and the windows themselves were highlighted. Users could also draw connection lines across document windows, which created a line to visually denote relationships.

Each participant trial involved using the document explorer tool to conduct an open-ended text analysis task with one of the data sets. The workspace included a note that instructed participants about their task. In all three data sets, the task was a fictional riddle scenario (e.g. weapons dealings, kidnappings, terrorist activities). The note instructed participant to explore documents to identify key people and

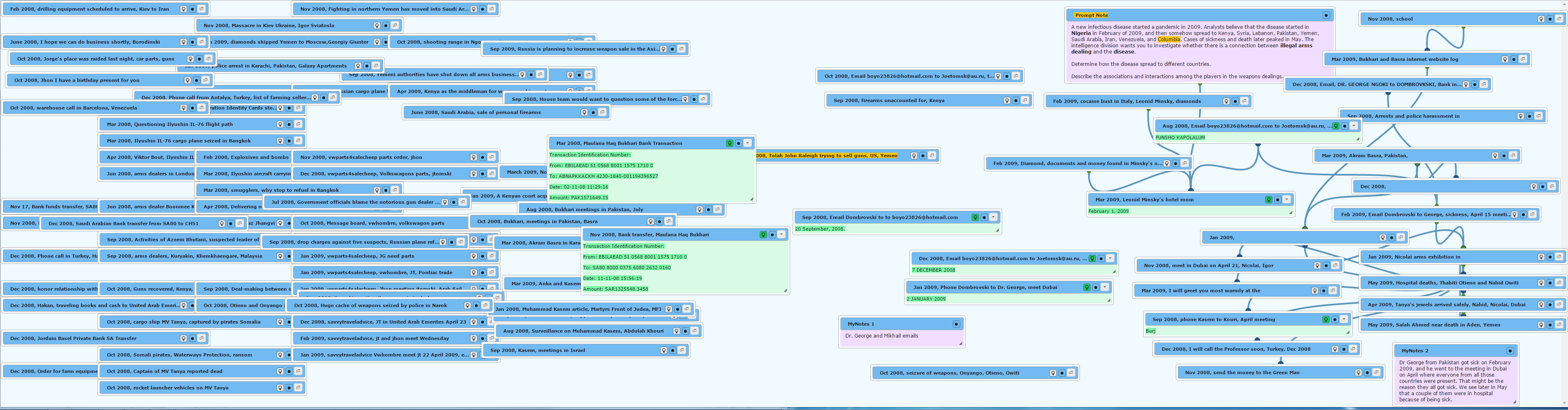


Figure 1 document explorer tool

events, and to find connections among them. These instructions gave participants a goal to work towards, but the prompt was sufficiently vague as to provide freedom in the analysis processes.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

You are free to use the *Provenance Analytics* *Dataset* for research purposes. Athough all publications which use this dataset should cite following work.

Rhema Linder, Alyssa M Pena, Sampath Jayarathna, and Eric D Ragan. 2016. Results and Challenges in Visualizing Analytic Provenance of Text Analysis Tasks Using Interaction Logs. Logging Interactive Visualizations and Visualizing Interaction Logs (LIVVIL) Workshop (2016).

Please contact Sina Mohseni (sina [dot] Mohseni [at] tamu [dot] edu) for any questions or comments regarding this dataset.