**Overview of the Analysis Tool**

Building on the initially proposed primary visual metaphor of visualizing connections between entities and documents in order to understand the given data set, the tool provides an interactive exploratory graph visualization – in this view, nodes can be either entities or documents and a link indicates that an entity occurs in a document. The graph allows users to add, remove, expand and collapse as they explore the dataset.

In addition to the graph view, which was our primary visual metaphor, we planned to:

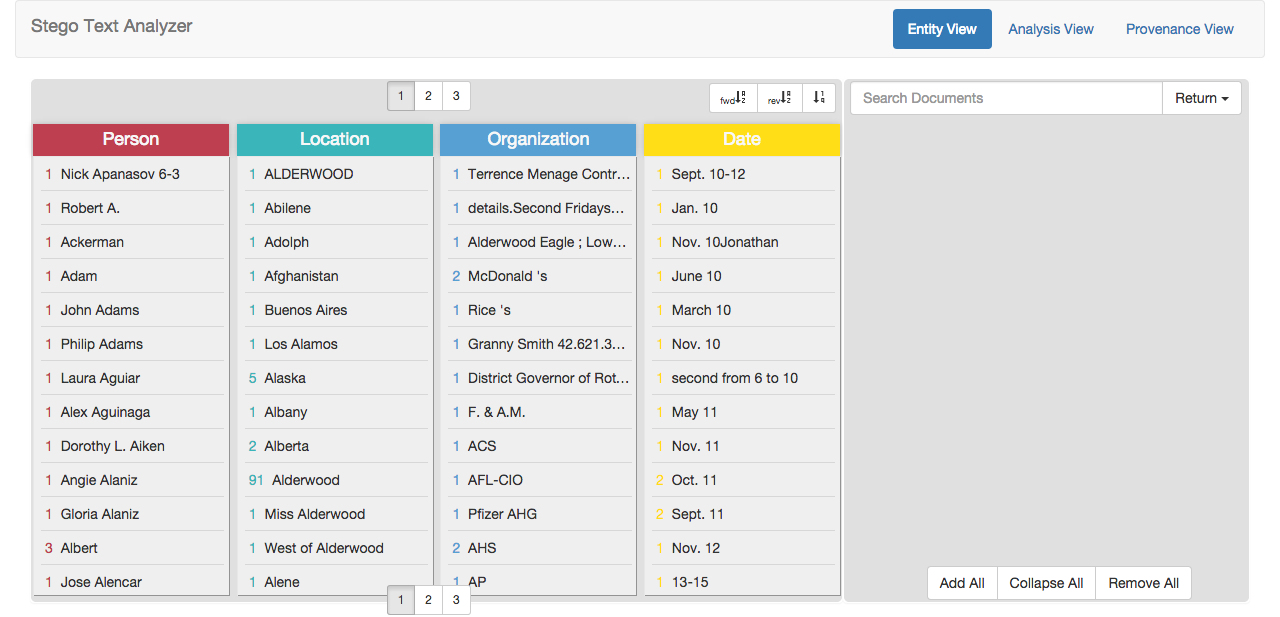
1. Provide users the ability to set up/format/clean the data as per their convenience before and during analysis.
2. Facilitate the user with interactive visualizations that give him a sense of what chronologic order between documents.
3. Have the system generate quantitative values that summarize the session and more importantly help in analytic provenance.

On these fronts, the tool presents:

1. An Entity View, which allows users to edit (rename, change type) or delete entities before and during analysis. This view also provides users the unique feature of creating aliases using multiple entities. For example, a user may decide that “John” and “John Panni” are the same character and hence create an alias. The user can then use this alias in the graph in order to get connections to both the entities.
2. A Calendar View, which is linked with the Graph View. This view assists the user in orienting data temporally while analyzing a relatively complex and growing graph visualization.
3. A Provenance View, which presents the user with a Degree of Interest (DOI) generated for each entity/document the user has accessed during the session. This is presented using the parallel coordinates technique. Each parallel represents a segment of time in the user's session and the value indicates an entity or document's DOI at that point in time. For example, if the user accesses the provenance view after 100 seconds from the session start time, there would be 5 parallels presented to him – 20,40,60,80 and 100. (Scoring logic for DOI explained during the in-class demo) In addition, the user has an option to create an ACH table on the go as he is performing the analysis using the floating window ACH table. Finally, the user can choose to download his session logs, which result in a formatted JSON file with each entry having the timestamp, triggering event and target value – hence enabling better recall.

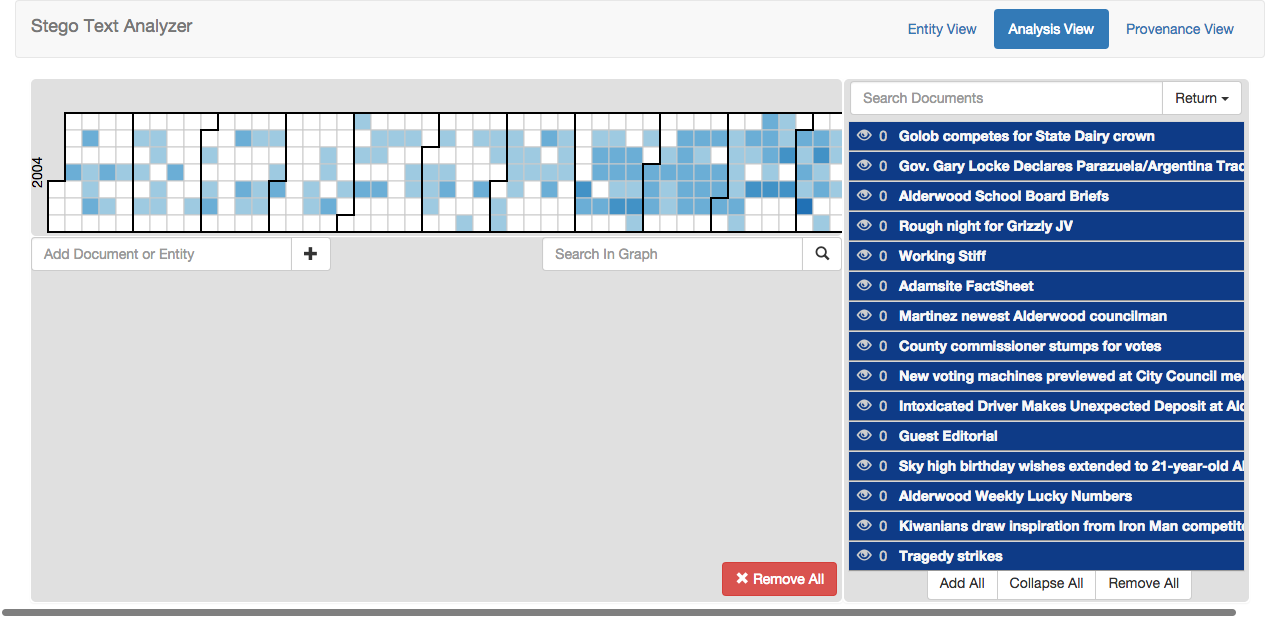
**Recalling the Steps of the Analysis Process**

1. Skim the entity view. Remove any spurious entities and create aliases where obviously necessary. Figure 1 shows the initial entity view.



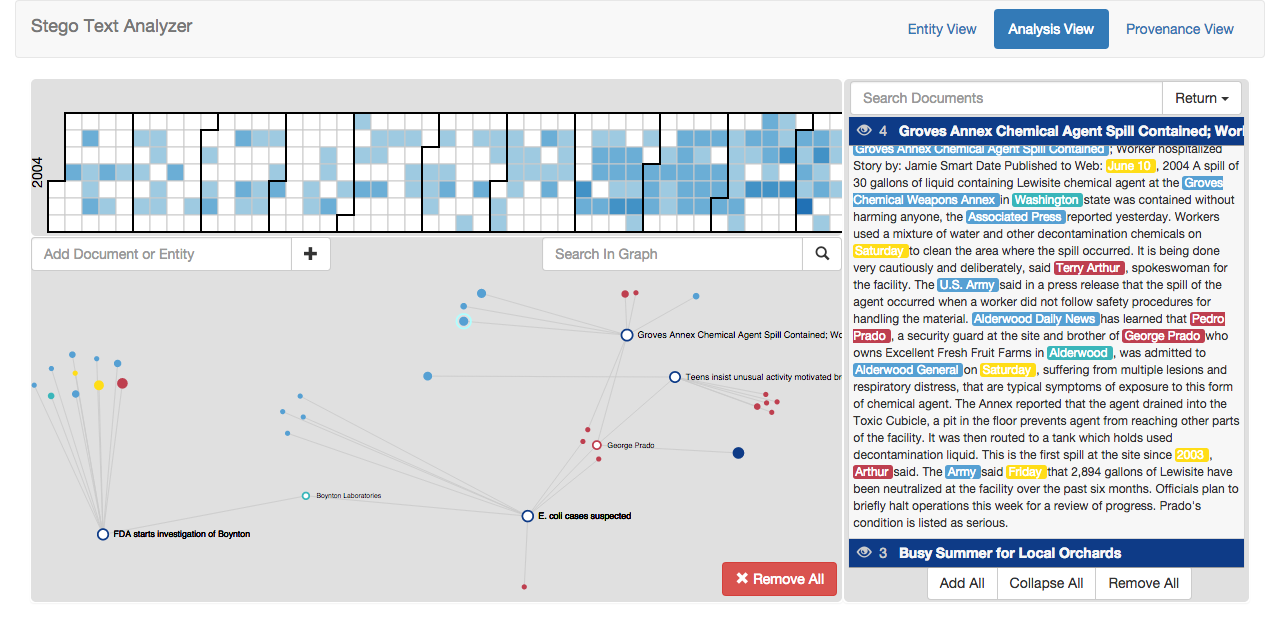
*Figure 1: Initial Entity View*

1. Move to the Analysis View. Press “Add All” in the Documents Panel to begin browsing the documents for interesting titles. Figure 2 illustrates the initial state of the Analysis View with all documents listed.



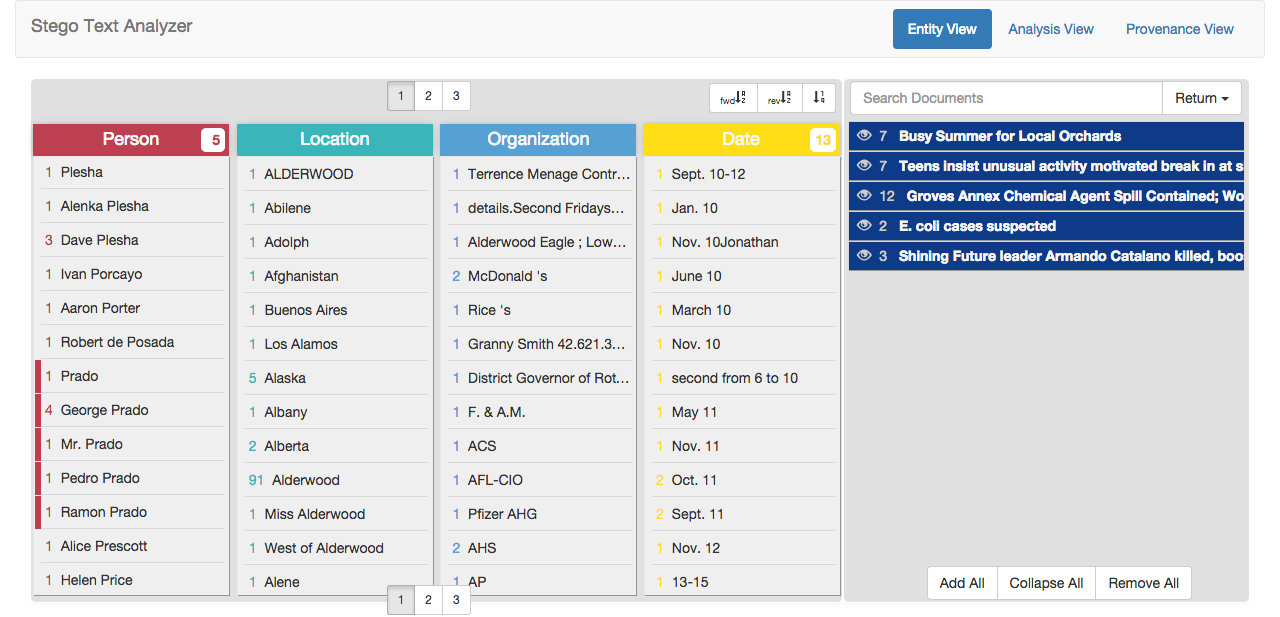
*Figure 2: Initial Analysis View with Documents Listed*

1. Notice a document called “E. coli cases suspected”. It seems interesting, so right click and select “Add to Graph” from the context menu.
2. Expand the newly added document in the Graph View.
3. Remove excess nodes. For example, “Alderwood” is far too general to be of use so remove it from the graph.
4. There is an interesting argument in the document between George Prado and Boynton Laboratories. Expand both of those nodes in the graph.
5. Boynton Laboratories is related to three documents. In one the FDA is starting an investigation. In another the FDA is cancelling the same investigation. The third describes some of the breakthroughs in mad cow disease going on at the laboratory.
   1. Right now this doesn’t seem related to the sick kids. Leave it in the graph view and move on for now.
6. Expand the George Prado node the result of this operation is displayed in Figure 3.
   1. First document seems benign. Establishes him as a leader in the local farming community who is looking forward to a good year.
   2. Second is interesting. Local teens claim suspicious activity going on at an old missile silo. *One of the teens breaking a bottle resulting in a burn and nausea reaction and a smell like a flower garden.* They report believing Prado is involved.
   3. The third mentions a spill of Lewisite at Groves Chemical Weapons Annex George Prado’s brother, a security guard at the annex, was hospitalized with chemical injuries.



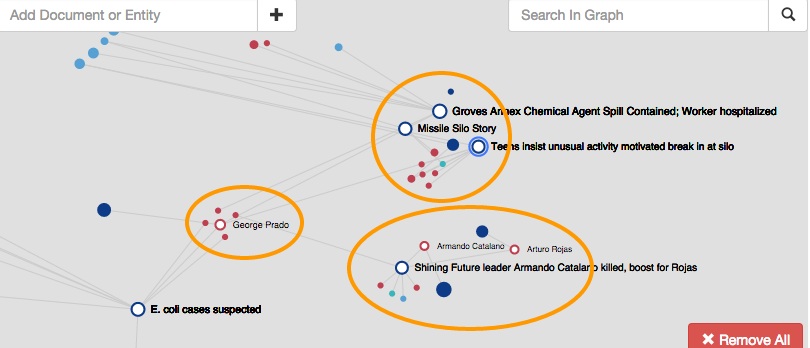
*Figure 3: State of the Analysis View During the Analysis Process*

1. At this point, we are interested in the burns that the teens mentioned after breaking the mysterious bottle. Click “Remove All” in the document panel and perform a search for all documents containing “Lewisite”.
2. Search returns a document called “Lewisite Fact Sheet” which describes the symptoms in detail. It’s a perfect match to the symptoms that the teens described, including a “geranium smell”. Furthermore, the document says that Lewisite has no legitimate purpose as anything but a chemical weapon.
3. The factsheet isn’t linked to any other entities, but is still highly relevant. Add the document to the Graph View and position it in a cluster with documents related to the chemical spill and silo incident.
4. We want to know more about this missile silo incident, so we perform a search for “missile silo”. It returns two more documents related to the same incident.
5. Check the timeline view to see a sequence of events. We wonder if the chemical in the silo could be a byproduct of the chemical spill.
   1. The teens posted a description of their adventure in the high school newspaper in January 2004, five months before the chemical spill. *This timeline tells us the presence of the chemical was deliberate and not the result of an accidental spill.*
6. Given George Prado’s family tie to the chemical weapons facility, and the teens’ belief that he is involved, he is now a person of high interest.
7. To find other relatives of Prado, go back to the Entity View. Reverse sort to see last names. See that in addition to Pedro and George, there is also a Ramon Prado. This reverse sorting is illustrated in Figure 4.



*Figure 4: Illustration of Reverse-Sorting to Find Last Names*

1. Opening the document related to Ramon, there is a discussion of the death of the leader of a South American revolutionary group, Shining Future. The dead leader’s envoy is Ramon Prado. It’s a tenuous connection right now, but add the document to the graph, expand it, and position it near Prado. Add Ramon Prado to a cluster with the other potential family members.
2. Upon expanding and exploring the new document and a few of the neighbors in the graph, notice a new document called “Port Incident”, connected through the Parazuelan President Arturo Rojas.
   1. The document describes an incident in which a ship carrying apples from Washington headed to Parazuela crashed into a dock. The crates of apples burst open and several workers were hospitalized with the same injuries seen previously.
   2. The ship’s captain made political remarks about Arturo Rojas, and is under investigation for a connection to Shining Future.
3. At this point, we observe the clusters that have appeared in the Graph View. One cluster pertains to the spill at the chemical facility and the adventure in the missile silo. Another cluster pertains to Shining Future and the port incident. A third cluster consists mostly of Prado family members, with George as the most prominent, connecting the other two clusters. The clusters are illustrated in Figure 5.



*Figure 5: Demonstration of User-Defined Clustering in the Graph View*

1. We begin to worry about cognitive bias affecting our perception of George. After all there are other apple growers, and the connection is based on the reoccurrence of the family name and a teenager’s testimony.
2. “Remove all” in the document view, start a new search for all documents related to “Shining Future”.
   1. A document describes an intoxicated driver who crashed into a bank and rated about “smelling flowers… but not too much then I die”, and “flowers will bring victory to our shining futures”. Clearly a reference to Lewisite, and strengthens the hypothesis that the Lewisite is shipping to the terrorist organization, but doesn’t strengthen the connection to George Prado.
   2. The search returned a large number of documents, and while scrolling through we notice one with a high view count. It’s the seemingly benign “Busy Summer for Local Orchards” document. In it, Prado is quoted as saying “I'd say our future is as shiny as our apples.” Suspicious, but not concrete.
3. Switch to the Provenance View. Search for high-interaction entities or documents that have escaped our notice.
4. Notice that in the beginning there was high interest in E. Coli-related documents and entities. This helps recall our original though process, which was to figure out what caused the suspected E. Coli outbreak. This prompts us to go back and re-read the E. Coli and Lewisite fact sheets and realize that some of the symptoms may be confused between the two. Possibly, the diagnosis of E. Coli was incorrect. The apple growers may have been incentivized to blame the incident on Boynton labs. Further searching for E. Coli-related documents reveals no additional compelling evidence.
5. End the analysis process.

**Analysis of Competing Hypotheses**

\*\*\* Please see the included xlsx spreadsheet, which contains the ACH Table\*\*\*

We divide the space of possibilities into four hypotheses. In the first hypothesis, George Prado is responsible, acting either alone or as a group, for a plot in which the chemical weapon Lewisite is trafficked to the Shining Future terrorist organization in Parazuela. In the second, the same plot exists, but Mr. Prado is not involved. In the third, Dr. Boynton has caused an E. Coli problem in Alderwood not related to weapon trafficking. In the final hypothesis, there is no suspicious plot. After adding 14 evidences from our analysis process and scoring them, the first two hypotheses are the clear winners, and the second two are highly unlikely. There is a slight edge towards Mr. Prado as the key figure, but only by a score of 9 vs. 8 for hypothesis #2.

**Analysis Tool Strengths**

Our tool successfully integrates several different views in a relatively seamless manner. The tight coupling between the Graph View and the Calendar View led to at least one major insight during the analysis process. It allowed us to easily prove that the Lewisite in the missile silo was not caused by the accidental chemical spill. Similarly, we found a major insight using the tight integration between Entity View and Graph View when we searched for everyone with the last name Prado.

The force directed layout of the graph provides quick and intuitive layouts, yet still affords the user the flexibility to position individual nodes exactly where desired. We found that user-defined clustering was extremely useful, with the force-directed layout actually suggesting clusters as nodes were expanded.

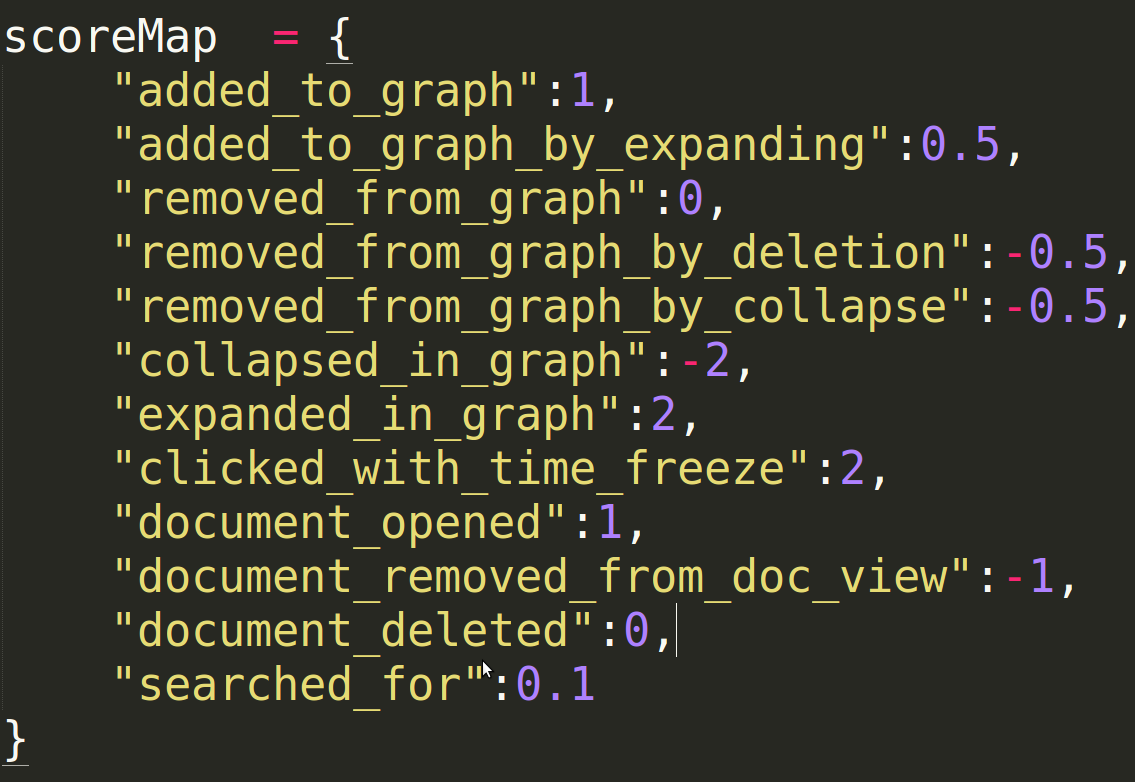
The document search functionality runs on a separate server, and is quite powerful. Multi-term searches are supported, and the analyst can even write simple Boolean functions such as Alderwood NOT Obituaries to find all documents containing the phrase Alderwood without including documents containing Obituaries. The search functionality proved extremely useful when it showed us Mr. Prado’s remark about the apples being as shiny as his future. We did not directly search for this phrase, but the server was smart to find it.

The Provenance View proved quite useful in confirming our hypothesis and increasing our confidence that bias did not skew our results.

**Analysis Tool Weaknesses**

The scoring for the Provenance View is somewhat naïve and could be more meaningful. We found that some entities or documents had a low weight when in fact they were extremely important, merely because we tended to put them in one spot and leave them alone, moving other nodes around them. A better scoring method would require a serious research effort, with weight given to the relative position of nodes in the graph, the frequency of moving nodes around an anchor, specific user behaviors, and more. The weighting scheme that we used is shown in Figure 6.

The Entity View was sometimes difficult to use, since switching to it causes a major context shift away from the Analysis View. If we had a better link that perhaps allowed us to select an entity in the Graph View and show it in the Entity View, then the two views would synchronize better and become easier to utilize.



*Figure 6: Code Snipped showing the Weight Scheme for the Provenance View*

**Collaboration**

Our analytic process first occurred as separate individual efforts to make sense of the dataset and to generate varying hypotheses based on evidence. Each group member used their own analytic techniques within the tool to aggregate, clean data, make connections, and gain insights. After these separate threads of analysis were complete, all of the group members met to discuss their findings in the datasets. The group members shared their findings while displaying the tool on a large display (LCD TV). One of the group members navigated the tool while the others prompted him to take desired actions.

The collaborative analysis began with overlapping hypotheses taking precedence. We found the hypotheses that were the same to be the most interesting. We mostly used the graph view to layout the network of entities and how we thought they were interrelated. We also used the search function to find documents that a group member knew existed and add it to the graph. The timeline brushing was also used to compare time periods of interest. After discussing and analyzing the overlapping hypotheses, conversation then diverged to related hypotheses of our main narrative. We then discussed hypotheses that appeared to be red herrings such as the E. Coli case. After more discussion aided by the tool we decided on our final analysis. We also used the tool to save a log of our interactions so that we would know which actions we took if we needed to recreate or remember our collaborative provenance.

**Handling Cognitive Biases**

We found the Provenance View to be helpful in dealing with cognitive biases. This view allows the analyst to look back and review whether all of the high-interaction documents and entities were truly considered in the sensemaking process. For example, if a cognitive bias caused the analyst to experience a bias against a certain individual, the Provenance View would show that another character has been just as heavily involved, perhaps in other ways, and should be considered more carefully. On the other hand, the Provenance View can help confirm when a bias is justified. In our analysis process, we felt that we may have a bias against Mr. Prado when we saw that our original question of what caused the E. Coli outbreak had been largely overlooked. However, this observation prompted us to go back through the documents and realize that the E. Coli diagnosis may have been incorrect, which strengthened our confidence in our story of the data.

**Conclusions and Future Extensions**

We are happy to say that we accomplished most of the goals that we set out in starting this project, and have used the tool to create a solution to the data set that we are very confident in. In fact, we have added quite a bit more functionality than we original set out to create. The Provenance View is an entirely new addition. We also decided to add user logs and an ACH table pop-out tool to help the analyst recall his process and easily generate hypotheses and evidences on the fly.

Along these lines, one idea for future work is to integrate the scoring for our Provenance View into the ACH table. This functionality could essentially generate a full ACH table by only observing the user interactions with the tool. Some technical things could use some improvement, such as completed brushing on the Provenance View and support for arbitrary data sets. As mentioned previously, we would like to provide a more graceful interaction to switch from Analysis to Entity View. Finally, we would have liked to expose our tool to more users to see how they react and also to determine which actions carry the most weight for the Provenance View.

Overall, we feel that this project improves on existing analysis tools, at least for the dataset in question due to the tight integration between views, powerful entity extraction, search functionality, provenance generation, and overall ease of use.