

A Tale of Alderwood: Visualizing Relationships in a Diverse Data Collection

Summer Adams

Susan Gov

Sheena Lewis

Kanupriya Singhal

College of Computing
Georgia Institute of Technology
Atlanta, GA

ABSTRACT

In this paper we present a novel implementation of classic visualization techniques. Our problem domain is the VAST contest sponsored by the IEEE Symposium on Visual Analytics Science and Technology. We begin by understanding the data sources and nature of the data provided. Considering this information, we select each visualization within our application to provide a wide variety of views into the data set which aids in uncovering otherwise difficult to determine relationships. We focus on tight integration among the various visualizations which is critical for usability and effectiveness. And finally, we summarize our lessons learned and propose additional features for future versions of the application.

Keywords

Text, graphs, maps, VAST contest

INTRODUCTION

The VAST contest is part of the IEEE Symposium on Visual Analytics Science and Technology. The primary task of the contest is to determine what, if any, illegal activities are happening in the small, fictitious town of Alderwood, Washington, given a collection of various data. More specifically, the data set consists of almost 1200 news stories, a few photos, a few maps of the town and vicinity, a few files of other mixed materials, and a couple of pages of background information. This is an incredibly diverse data set that poses a number of challenges.

Text analysis and parsing was by far the most difficult aspect of this project. This is not surprising considering that natural language processing is still an open research domain. Techniques explored in the past to visualize textual data include visual maps [4], PaperLens [3], and query visualization [1]. Inspiration for part of our design was derived from these sources.

Beyond the higher level text visualization, we also chose to visualize the entity networks within the data. Ideas for this visualization were stimulated by social visualizations such as Vizster [2]. Vizster is limited to the people entity type which was not sufficient in our case so we expanded the model to include additional entities. Our modifications to

the model drew upon prior work done on graphs and networks.

A timeline was implemented to show temporal relationships among entities. There is a significant body of work that has been completed regarding visualizations of time series data and once again we were motivated by some of this previous work.

And finally, a map visualization was added to represent locations in the data collection.

Given the quantity and nature of the data involved, we decided to focus our efforts on visualizing the relationships between data entities. Thus, it followed that multiple visualizations would be required in order to show the various relationships among the specific entities. The interactions between the separate visualizations was just as critical, therefore we concentrated much of our attention on that aspect of the project as well.

DATA ANALYSIS

Data Acclimation

After collecting the data set from the VAST contest, it became clear that before considering a visualization, we first had to understand the type of information provided in the data set (e.g. articles, photos, voter registration list, phone logs, etc.). We began to gain familiarity with the data by dividing the 1200 articles amongst the four group members. Each group member skimmed through this subset of the articles in order to understand the information that was being provided and to gain insight into the types of information that may be relevant to understanding the story of Alderwood. The group collaborated and discussed interesting or unexpected facts that were collected from our brief individual interactions with our portions of the data set. From this initial interaction with the articles, we collected various types of information such as information regarding the key players in the town of Alderwood and

their relationships with other people and occurring events. After some analysis, it became clear that these relationships, which are embedded in the textual representation of the data set, are important to understanding the overall story of Alderwood. Therefore, through our initial analysis, we began to explore all the potential connections between pieces of the data and how a visualization could augment a manual analysis of the story of Alderwood.

Data analysis was a challenging aspect of this project. Familiarization with the data required considerable effort. Once we had obtained a general understanding of the data, we then focused on the data structure appropriate for the data set. Finally, we developed methods to automatically manipulate the data to convert it into formats to use in the visualizations.

Data Structure

Evaluation of the data structure led us to define five separate entity types. We determined the types should be Person, Organization, Location, Event, and Activity. A Person represents an individual; an Organization represents any collection of people; a Location represents anything that has a physical instantiation; an Event represents a specific incident that occurs at given a point in time; and an Activity represents an incident that occurs over a period of time and typically consists of a number of events. Examples of each entity type from the data set are included in the table below. Each data value is associated with a single entity type.

ENTITY	VALUE
Person	Rex Luthor, John Torch, Dr. Philip Boynton, Ronald Reagan
Organization	Alderwood City Council, Alderwood Police Department, U.S. Department of Agriculture
Location	Alderwood, Boynton campus, Wine Country Road
Event	City council meeting, attempted robbery, new casino opening
Activity	2004 election campaign, develop mad cow disease test, study plan for voting machines

Table 1: Entity types from the data are listed along with examples of each type.

Data Manipulation

Perl scripts were written to automatically extract entities from the data set. Perl was selected because it provides a powerful mechanism by which to massage and control data.

Word counts were extracted using a script designed to parse through each news story. The script incorporated stemming using the Porter stemmer, stop wording, and text frequency-inverse document frequency (TF-IDF). Stemming was included to transform words down to root form in order to reach a more accurate count. For example, “swimming” and “swims” both become “swim”. Stop wording was included to remove common words that are not key to the semantic meaning such as “a”, “the”, and “and”. TF-IDF effectively results in normalization of the words across documents. In other words, a word that occurs ten times in a small document is more meaningful than if it occurs ten times in a large document therefore the smaller document is more relevant to the word than the larger. All options were created as variables, allowing the user to turn them on or off at will.

To extract entities from the data initially, a combination of Perl scripts and a parser was used. The Perl scripts were implemented for rudimentary processing followed by processing by the parser. The parser we employed was the Link Grammar Parser available from the School of Computer Science at Carnegie Mellon University. We built a Java interface into the parser, which was written in C, for convenience purposes.

All data was ultimately converted into formats appropriate for each visualization we created. Perl scripts were applied to this task as well. An XML file was generated for the initial treemap view. Subsequent XML files were generated from within the application at run-time for detailed views within the treemap. For the social network view, all XML files are generated at run-time based upon filter criteria selected by the user. Two matrix files were produced to capture the relationships between entities and between specific news stories and entities. Each of these files was constructed with the top section representing pertinent information about either the news story or the entity as appropriate. In both files, the bottom section consisted of a matrix of 1’s and 0’s indicating whether or not a relationship existed between two entities or between the news story and the entity.

SYSTEM DESIGN

As stated in the previous section, we gathered multiple relationship matrices from the data extraction. We then were faced with the task of creating a visualization that affords extensive analysis of relationships in Alderwood. When examining potential visualizations, we considered the users of the visualization (i.e. governmental investigative agents) and the purpose of the tool (i.e. to afford analysis). Due to the fact that certain visualizations of the past have been successful at displaying particular

relationships, we decided not to create a completely new visualization that shows relationships. Instead, we decided to seamlessly integrate “classic” visualizations that have been known for showing relationships well. Therefore, we decided to combine multiple visualizations and interaction techniques in a novel way in order to demonstrate relationships in Alderwood.

The classic visualizations that we use are a treemap, a node-link graph, and a timeline. These types of visualizations were the correct visualizations for us because users are familiar with the data and it allows for easy analysis. Additionally, by using traditional visualizations, we were able to focus on the interaction of these different visualizations. By using these types of visualizations, we are also able to provide users with a high-level view of the data rather than only a detailed view. Furthermore, each visualization compliments the others by providing strengths that are not demonstrated by the others.

On the top left of the visualization, there is a treemap that displays selected entities (which the user requests to search on) as well as their relationship with other entities. The hierarchy from the treemap is created through the relationship matrix (i.e. the matrix the holds the relationships between all of the entities). The treemap uses a square-ified formula from the prefuse package. The initial view of the treemap presents a view of the hierarchy of the data based upon entity type. Each entity associated with a particular entity type is displayed in the corresponding section.



Image 1: Treemap panel.

Moreover, there is also a map visualization at the bottom left of the application. This visualization was created from

the map provided by the VAST contest. The map displays location entities as the user chooses to filter the visualization.

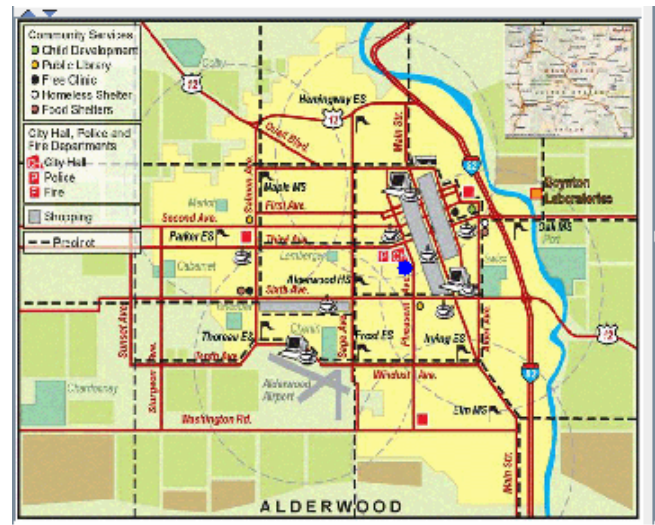


Image 2: Map panel.

The application also has a visualization of the articles and other documents in a textual view and timeline view that is located in the top center panels. The entire articles list are displayed as well as only the AND of the searched items after the user chooses to filter on particular entity items.

Articles	
Name	Date of Article
Alderwood to probe voting machines	11/16/2004
Forum	10/29/2004
Masked man attempts purse snatchi...	10/08/2004
Alderwood School Board Briefs	10/04/2004
Man arrested for using fake ID	10/04/2004
Briefs	09/20/2004
Forum	09/08/2004
Burglars strike in Grandview	09/08/2004
Alderwood City Council Briefs	08/26/2004

Image 3: Document panel.

The timeline visualization lists the articles in a temporal view that is not simply textual. This affords the transfer of preattentive information to the users regarding the articles and when they were published.

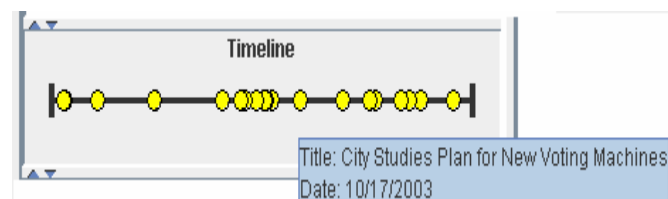


Image 4: Timeline panel.

The middle of the application screen contains a link-node graph that displays information about entities that are selected to be filtered upon. This graph provides a visual representation of the entities that are related to the filtered items. The colors of the nodes differentiate the various entity types and related entities are connected by edges. If two entities being filtered upon do not share any relationship with each other, the link node graph easily indicates this by showing each entity without an edge between them. Instead, each entity is displayed as a standalone cluster representing existing relationships with other entities. When zooming in on the graph, the nodes are displayed and can be moused over in order to reveal the entity name.

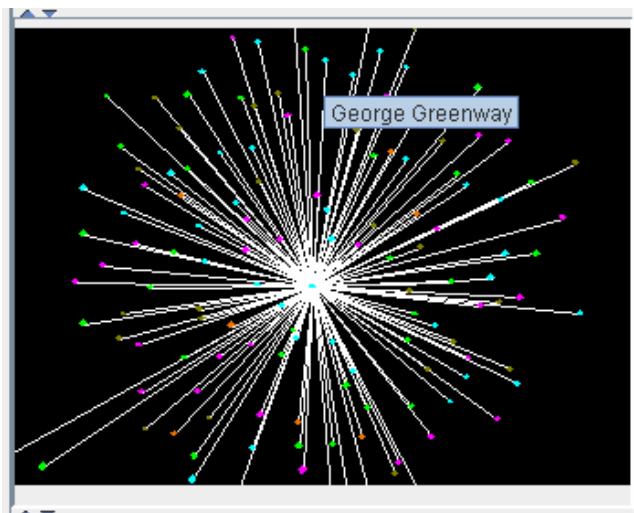


Image 5: Social network graph panel.

In addition to these classic visualizations, we provided users with a detailed view of the data by displaying each entity item on the right of the screen. Therefore, if users are interested in information about a particular entity, (e.g. a person, organization, activity, event, or location), they can search for the entity on the right side of the visualization. The entities are organized into sorted lists in individual tables, where each entity is placed in a separate table. These tables provide the name, date, and the frequency (i.e. the amount of times the entity is mentioned in comparison to the other entities) pertaining to an entity. The entity lists can be sorted by any of these attributes, which can prove helpful if the user wants to view the entities in temporal order (i.e. the first date the entity was mentioned) or to understand how frequent the entity is mentioned.

People		
Name	Date of Birth	Frequency
Aaron Bones	03/04/2002	
Adelia Goedhart	04/02/2003	
Alex de la Cruz	08/26/2004	
Amanda Martinez	09/20/2004	
Amber Hansen	07/20/2003	

Organizations		
Name	Date	Frequency
Alderwood Chambe...	04/02/2003	
Alderwood Chevron	03/25/2003	
Alderwood City Cou...	01/23/2002	
Alderwood Daily Ne...	03/23/2003	
Alderwood Elks Lod...	07/12/2003	

Events		
Name	Date	Frequency
absent from meeting	02/10/2002	
accept project as co...	09/09/2004	
adopted Joint Memo...	02/19/2002	
adopted resolution ...	05/11/2003	
Alderwood Port Dist...	09/20/2004	

Activities		
Name	Date	Frequency
2004 election	07/12/2003	
bad feelings between	03/23/2003	
check forgeries	12/31/2003	
develop mad cow di...	02/19/2002	
election campaign	02/04/2004	

Locations		
Name	Date	Frequency
10 North Eighth Street	09/08/2004	
100 block of Wine C...	09/08/2004	
100 East South Hill ...	05/25/2003	
107 W. Lincoln Ave	10/04/2004	
111 East Lincoln Av...	03/25/2003	
1328 Road 28	08/05/2004	
1620 East Edison	01/23/2002	

Image 6: Data tables panel

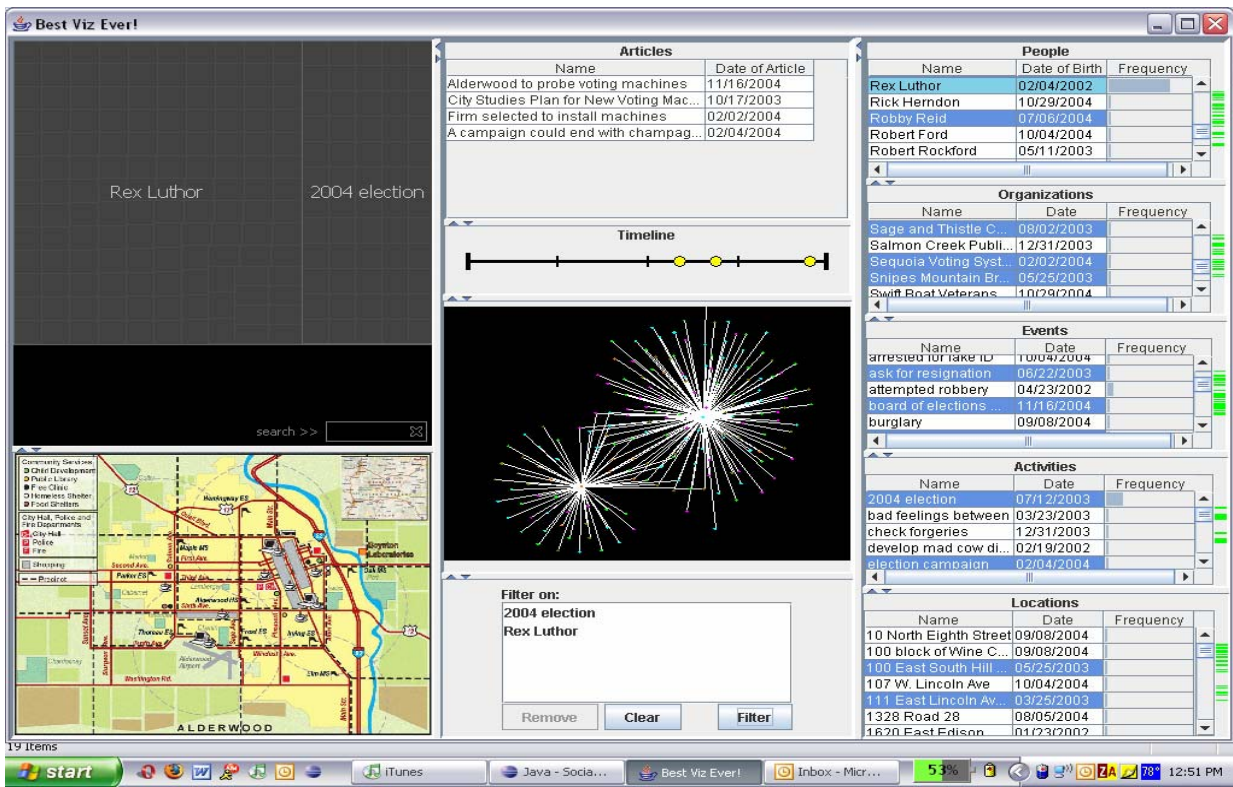


Image 7: Full image of all visualizations in the application.

Overall, these are the visualizations that our application contains. In the next section, we will discuss the interactions that the system supports.

SYSTEM INTERACTION

The interaction of the visualization is the most important part because it demonstrates the linking of the “classic” visualizations as well as provides an opportunity for users to interact with a new system in a way that shows relationships about the people of Alderwood.

Currently, the system is driven by the detailed data on the right of the screen. The user would first have to select an entity on the right of the screen. By single clicking on one of these entities, the related entities are highlighted in the other tables. This is an interaction that is necessary if the user wants to see how one entity is related to the others. For instance, if the user wants to filter all the people that are related to Alderwood High School, he or she would not have to filter first because the highlighting would provide whether a person is related to Alderwood High School.

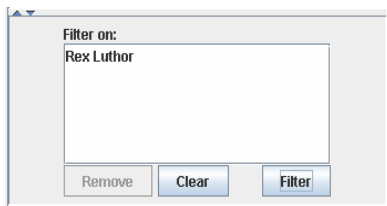


Image 8: Filter panel.

Furthermore, we recognize that the list of entities may be long, which may make it difficult for the user to see which related entities are highlighted. Thus, we created a side panel that corresponds in location to the highlighted rows of a table. Clicking on one of the green bars forces the entity table to jump to the location that was selected. For instance, if there are many entities at the bottom of the table, the user can just click the bottom dashes that represent the highlighted relationships and the table will jump to the highlighted entities. This side bar panel affords the user the opportunity to quickly estimate the number of relationships a particular entity has in common with all the entities. This can be useful when an investigator who is trying to familiarize himself/herself with the names and events that occur in the town quickly by showing the relationships each entity has the other others, which could be an indication of its importance.

By double clicking on an entity list item, the entity list item will appear on the query table, which is at the bottom center of the screen. This means that when the user clicks the “Filter” button, all of the visualizations will filter on all of the items that are located in the query panel. Inside of the query panel, the user is also able to remove items and clear the entire table.

By clicking on the filter, all the visualizations are updated with respect to the items that are set to be filtered. The tree map will only show the entities that have been filtered upon and the relationships that entity has with other items. The sizes of the nodes inside of the treemap are an indicator the frequency of the entity. For instance, if a user searches on the mayor of Alderwood, taxes may be a large block inside of his section of the treemap because the mayor may be strongly connected to taxes and the creation or elimination of new tax laws for the city. The nodes in the treemap are highlighted when moused over as well. The user is also able to search for items in the treemap by using a search tool that is specific to the treemap panel. This search feature allows the user to dynamically search for items within the treemap by typing in letters that are not necessarily in the order of what he or she is looking for. However, through this dynamic search, entities can easily be found within the treemap without much effort from the user. The population of these entities also demonstrates how easily relationships are shown using our visualization.

As for the map visualization, the filter only shows relationships that are locations on the map. If the filter does not produce any locations, then the map visualization will not be updated and will stay static. However, if the filter does provide locations, the map visualization will be updated to display these locations. It is recognized by the team that the map that is displayed may not be the easiest to read; however, it was important for us to preserve the integrity of our data set by using the map that was given to us in the original data set by the VAST contest. Therefore, we continued to use the map that was provided despite it is not as clear and concise as we would like it to be.

In addition to updating the treemap and map visualization, the social network visualization is also affected. The social visualization shows the entities that each item is related to and how they are related to each other. As stated above, if there is no relationship between a pair of entities being filtered upon, these entities and their relationships will be displayed as separate, disconnected clusters. The social visualization provides a way to visually see the relationships using a node-like graph, which most users are familiar with. By mousing over the nodes in the social network, the user is able to clearly see the entities names. Therefore, if they are interested in searching on a particular node, he or she can easily find it in the entity list to the right of the screen.

As for the articles, when the filter button is selected, the articles that are related to all of the filtered entity items are displayed in the table. This information provides the user with direct access to the articles and the photos that the relationships are gathered from. Therefore, the user is able to understand further how the filtered items are related through the articles. This is a way for agents to understand relationships without being overwhelmed with all 1200 articles. For instance, if a user wants to understand the

relationship that mayor has with the city council, then he or she could filter on these two entities and only read the articles that contain information on both the mayor and city council. This may provide the user with direct knowledge from the raw data set but by allowing only a subset of articles to be shown, the user may feel less overwhelmed by the data.

In addition to the table of articles, the temporal timeline provides information about the articles. In fact, the temporal timeline provides a list of the filtered articles with respect to time, such that users should be able to quickly gather when the articles were written and an estimate of the amount of articles written without heavy calculations. The user is able to mouse over the timeline to see the article title and the date the article is written.

STRENGTHS AND WEAKNESSES

We believe our application has a number of strengths. First and foremost, following Shneiderman's mantra, it permits the user full access to all of the data in a central location without overwhelming the user. This facilitates the user acquiring multiple insights into the same data at the same time and creates an opportunity for the user to uncover information that may not have been as obvious using standalone views. Because the visualizations complement each other, they are much more powerful together than any single visualization of this data would be on its own. The user also has a number of options when interacting with the data. The application permits overall views into the data as well as detailed views. This gives the user the capability to get a high level picture of what may be interesting and then to drill down into detail based on explicit entities. Another strength of the application is that it shows relationships very easily and in multiple views. It shows trends in the data as well. Even though the application affords many capabilities, it is simple to use and interpret. And finally, our visualization application is practical. The interactions make sense and incorporate common features that users have become accustomed to in software.

One weakness with the application is the automated data extraction. Much work was done to extract information from all data sources without human intervention, however there were some sources that required hand manipulation on occasion.

IMPLICATIONS AND RESULTS

The results of our visualization were exactly what we hoped for in that it affords a clear, concise view of relationships. By viewing the relationships, we expected to be able to find connections in the data that would otherwise be hard to observe using the original VAST data set. After completing this project, our hypothesis was true in that we were able to see connections between the different entity types, which allowed us to see anomalies in the data that helped point to suspicious activity in the data. For instance, we found out there was an election from a section in the tree map. From the election entity, a connection to

John Torch was observed using the social network visualization. Therefore, we decided to look for more data on John Torch. After filtering on Torch, we were able to find out that Torch was running in the election, however due to an affair scandal, he dropped out of the elections. From this information, it leads us to believe that Torch could either have been set up, or he really could have really been having an affair with a worker from Boynton Laboratories. This is just a small example of how interesting data can emerge from showing relationships.



Image 9: Treemap with election campaign detail data.

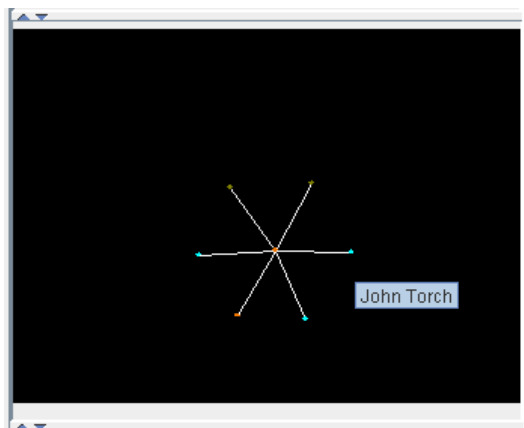


Image 10: Social network for John Torch.

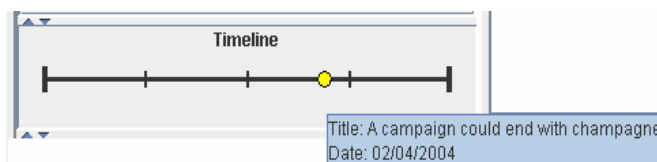


Image 11: Timeline with article for John Torch.

Though our visualization worked the way we anticipated, there are many things that we did not know before going attempting this project. Having knowledge about the

things that we know now, we recognize that there are certain aspects of the project that we would have done differently or prepared for in advance.

One aspect of the project that we did not take much consideration to is the work necessary to mine such a large and varied data set. Having text files, pictures, phone logs, Excel files, etc., we did know understand the extent of effort that is needed to extract useful data from this information and visualize it such that it is helpful. This became problematic because there are not many proven technique that have complete success with data minding. Because this is an on-going field of research, we had to try multiple ways of extracting the relationships from the data set. It became an extremely tough and tedious process that should have been given more time and attention in the beginning of the project.

In addition to extracting that data that we wanted, it took an extremely long time to decide on the entities because we first had to understand the data set thoroughly enough to decide on a method of collecting “important and useful” data. After figuring out that relationships were important to the data set, we had to decide how to discuss the relationships between different types of things. Thus, we came up with the five different entity types. This was difficult in that we had to thoroughly debate on the types of entities to ensure their usefulness and significance to users of the application. After debating on these entities types, there were issues with coming up with strict definitions for the entity. These definitions were essential to making sure there was no overlapping of the entities.

Lastly, our first attempt at achieving a visualization was completely different then described in this paper. We attempted to decide on visualizations for the data set without truly understanding the problem and knowing how we wanted to tackle the problem. Therefore, we came up with a visualization that was in search of a problem rather than a visualization that that solved our problem. After speaking with the professor, we realized that our visualization ideas needed to be completely rethought and that we needed to decide on a concrete task that we wanted our visualization to focus on. Hence, the approach that we decided upon was to show relationships well. However, the time that we spent determining an incorrect visualization was time wasted. So if we were to redo this project, we would begin with finding a good task (e.g. showing relationships well) rather than focusing on the visualization or the data set. From our experience, after deciding upon a good task, the visualization was relatively easy to decide upon.

Detailed Theories

Based on our insight into the data we believe there are a few crimes being committed. First of all, we believe the mayor of Alderwood, Rex Luthor, is involved in taking bribes from private organizations for personal gain. In particular, Rex Luthor received personal gain from the

contract with the company that provided the new electronic voting machines for the 2004 elections. We also believe that he saw John Torch as a threat to his office and had him set up by Laurel Sulfate and Boynton Labs to embarrass him into dropping out of the race. We believe Boynton Labs is involved in paying government officials for special favors, mostly because there is evidence that high level politicians in Washington state intervened on their behalf to cancel an investigation into their practices by the U.S. Department of Agriculture. Other minor crimes include burglaries and attempted robberies. Our visualization shows connections between these events as shown in the image below. We believe these are not significant in the overall scheme of the shady dealings in Alderwood.

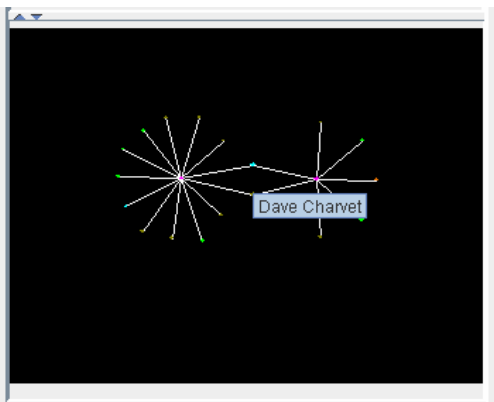


Image 12: Social network on burglaries and attempted robberies.

FUTURE WORK

For future work we would first of all suggest incorporating additional map views into the visualization, possibly including the ability to pan and zoom. We would also add user control to the timeline, allowing the user to specify periods of time that are of interest in order to view more or less detail. Additionally, by incorporating supplementary domain knowledge about the data set into the automatic extraction utilities, their accuracy and effectiveness could be improved. Another modification we would make in a future version would be to add the ability for the node sizes in the treemap to vary based on a frequency of occurrence metric. Adding the capability to interact with the social visualization graph would be one more nice feature to add in the future. Ultimately it would be great to integrate a feature set related to building and supporting hypotheses based on the data from the visualizations. This would require almost an entirely additional application but would further highlight the value of our application.

CONCLUSION

Analysis and understanding of disparate data sources is an incredibly challenging problem. One must first fully appreciate the nature, depth, and coverage of the data set. Then based upon this understanding, a taxonomy and extraction tools must be developed. Depending on a person's background, this may not seem like a difficult problem, however as mentioned earlier, it is still an area of open research that receives a significant amount of attention. Though once the underlying data structures are understood it becomes possible to define the relationships within the data. From there, the visualizations can be selected and implemented that bring forth the relationships for the end user in a much more effective manner than manual text analysis.

In the end it is essential to focus on the tasks the user will be performing and to relate that back to the data set being used. We believe our visualizations in our application and the interaction we created between them accomplishes this goal. We combine some classic visualization techniques in a new and interesting way with full consideration for the data set from the VAST contest. We make the relationships in the data the center of attention thus facilitating the first step an investigator would take when exploring the history of a small community such as Alderwood. This combination of features provides a solid base from which to build if one were to expand the application for assisting investigators with later stages of their inquiries.

ACKNOWLEDGMENTS

We would like to thank Dr. Stasko and Bob Amar for their feedback on our project and approach over the course of the semester.

REFERENCES

1. Havre, S., Hetzler, E., Perrine, K., Jurrus, E., and Miller, N. Interactive Visualization of Multiple Query Results. In *Proc. IEEE Information Visualization Symposium, 2002*.
2. Heer, J. and Boyd, D. Vizster: Visualizing online social networks. *InfoVis 2005 IEEE Symposium on Information Visualization, 2005*.
3. Lee, B., Czerwinski, M., Robertson, G., and Bederson, B. Understanding Research Trends in Conferences using PaperLens. In *CHI 2005 extended abstracts on Human Factors in computing systems*, Portland, OR, 2005. ACM Press.
4. Lin, X. Visualization for the Document Space. In *Proc. IEEE Visualization 1992*, Boston, MA, pages 274-281, Los Alamitos, CA, October 1992. IEEE Computer Press Society.