Jigsaw: Supporting Investigative Analysis through Interactive Visualization

John Stasko*   Carsten Görg†   Zhicheng Liu‡   Kanupriya Singhal§

School of Interactive Computing & GVU Center
Georgia Institute of Technology

ABSTRACT

Investigative analysts who work with collections of text documents connect embedded threads of evidence in order to formulate hypotheses about plans and activities of potential interest. As the number of documents and the corresponding number of concepts and entities within the documents grow larger, sense-making processes become more and more difficult for the analysts. We have developed a visual analytic system called `Jigsaw` that represents documents and their entities visually in order to help analysts examine reports more efficiently and develop theories about potential actions more quickly. `Jigsaw` provides multiple coordinated views of document entities with a special emphasis on visually illustrating connections between entities across the different documents.

Keywords: Visual analytics, investigative analysis, intelligence analysis, information visualization, multiple views

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 INTRODUCTION

Investigative analysts seek to make discoveries and uncover hidden truths from large collections of data and information. Often, the investigative process involves analysts pouring over sets of textual reports, reading and reviewing the documents to make connections between seemingly disparate facts. Scientists follow this process when they read research papers to learn about related efforts; newspaper reporters perform such analyses when they investigate new stories; law enforcement and intelligence analysts carry out these kinds of investigations when they review case reports.

While reading reports and digesting the information therein, analysts gradually form internal mental models of the people, places, and events discussed in the reports. As the number of reports grows larger, however, it becomes increasingly difficult for an investigator to find relevant information, track the connections between data, and make sense of it all. The sheer number of entities involved may make it very difficult for a person to form a clear understanding of the underlying concepts and relationships in the report collection.

Much like many others, we believe that visual representations can aid people to examine and understand abstract data such as this. For example, Norman has described how visual representations can help augment people's thinking and analysis processes [15]. Card, Mackinlay, and Shneiderman refer to visuals used in this manner as "external cognition aids" [6].

The objective of our research is to develop visual representations of the information within textual document and report collec-

---

*e-mail: stasko@cc.gatech.edu
†e-mail: goerg@cc.gatech.edu
‡e-mail: zcliu@cc.gatech.edu
§e-mail: ksinghal@cc.gatech.edu

tions in order to help analysts search, review, and understand the reports better. We seek to create interactive visualizations that will highlight and identify connections between entities in the reports where entities may be people, places, dates and organizations, for instance. Fundamentally, we want to build visual representations of the reports that help analysts browse and explore them, making sense of all the facts and information contained in the reports.

Our goal is not to replace the reports, however. We firmly believe that analysts must carefully read reports to best understand them. What we seek to provide is a type of interactive visual index onto the reports, a visual analytic system [17, 23] that connects and links entities discussed therein and thus guides analysts toward the reports to read next. Furthermore, the interactive visualizations should provide representations that assist analysts in building accurate and informative conceptual models of the underlying themes, plots, and stories embedded in the report collection. Our approach is human-centered; we want to design an easy-to-use system that puts the analyst in charge of analysis as opposed to relying on algorithmic, automated techniques.

Pirolli and Card performed a cognitive task analysis of intelligence analysts and their work that resulted in a notional model of the intelligence analysis process [4]. Their model is organized around two major activity loops, foraging and sense-making. Our work touches on both loops, helping analysts to choose useful reports to examine next and also to develop schema and hypotheses that fit the available evidence. Pirolli and Card identify several leverage/pain points particularly in need of assistance within analytic processes. Two, in particular, that our work addresses involve 1) the cost structure of scanning and selecting items for further attention and 2) analysts' span of attention for evidence and hypotheses. They comment on the two leverage points, respectively:

> "Our analysts spent considerable time scanning data seeking relevant entities (names, numbers, locations, etc.). The assessment of whether or not an item is relevant also takes time. Techniques for highlighting important information with pre-attentive codings, or re-representing documents (e.g., by summaries) appropriate to the task can improve these costs."

> "Techniques aimed at expanding the working memory capacity of analysts by offloading information patterns onto external memory (e.g., visual displays) may ameliorate these problems."

To address such objectives, we have designed a suite of interactive visualizations and built a prototype system called `Jigsaw` that implements the visualizations as separate views onto a report (text document) collection. The views are connected so that actions within one view can be reflected in the others. We named the system `Jigsaw` because we think of all the different entities and facts in a report collection as the pieces of a puzzle. The `Jigsaw` system should help an analyst "put the pieces together."

In the next section we provide more details about the types of reports that are the focus of analysis for `Jigsaw`. We also describe the entity types that are extracted from a report and serve as the

primary basis for the visualizations. Section 3 reviews the `Jigsaw` system in detail, its underlying data structures, system architecture, event messaging, and each of the different views. In Section 4 we provide a short example scenario of use to better help the reader understand how the system functions. The paper concludes with a discussion of related work and a list of ongoing and future efforts planned for the system.

2 ANALYZING REPORTS

The target artifact of our study is a textual report describing some set of facts or observations from the domain of interest to analysts. We assume that the reports will be in a natural language format and likely of a length of about 1-5 paragraphs. There is nothing inherent in our work to prevent longer reports from being used, but our intended target is a smaller report with a few nuggets of information contained therein.

Analysis can draw on data from varied and distributed sources. The distributed nature of information leads to heterogeneity across the reports in terms of topic, authors, content, style, date and so on. Furthermore, different reports will contain information that is unclear, confusing, or even contradictory. Organizational tools have to consider both the complexity of the information as well as the analysis task.

Below is an example report, taken from the VAST 2007 Conference Contest [18], that provides a flavor of the types of reports on which we are focusing. A large number of different events, items, themes, and stories can be embedded throughout a collection of thousands or even just hundreds of such reports.

Wed Jul 16 17:35:00 2003
(Los Angeles) A package of beef and a letter from "Animal Justice League" claiming that meat had been poisoned in 20 Los Angeles supermarkets was left at the Los Angeles Times, 1st Street offices. The paper also received a phone call taking credit for the action and stating, "We will take direct action against animal abuse in whatever form is necessary to stop the cruelty." A similar threat was made against a supermarket in Upsala in November. No poisoned meat was found at either supermarket.

While other systems such as `IN-SPIRE` [20] focus mainly on themes or concepts across document collections, the primary unit of analysis from reports for `Jigsaw` is an entity. Within any report one can identify a set of entities. For the current version of `Jigsaw`, we focus on the following entity types: person, place, date, and organization.

The goal of the `Jigsaw` system is to highlight and communicate connections and relationships between entities across a report collection. We believe that these connections, when assimilated, help to provide the analyst with a better global understanding of the broader themes and plans hinted at by the particular events and facts documented in the reports.

Obviously, an initial requirement for `Jigsaw` is to identify and extract the entities [8, 11, 14] from each report and, ideally, store them in a format that allows easier analysis and manipulation. Entity extraction, however, is not the focus of our work so we presently are exploring the use of tools and techniques from other researchers in this process.

To help initiate our work, we adopted analysis exercises created by Frank Hughes of the Joint Military Intelligence College [9]. The exercises involve collections of fabricated reports with an embedded master plot. Different reports in the collection hint at this plot and the goal of the exercise is to discover and articulate the plot. To make identifying the master plot more challenging, threads of other unrelated plots are suggested in the reports as well. Analysts in training perform exercises like this as part of their educational process.

To bootstrap development of `Jigsaw`, we extracted the entities in the example report collections by hand. We created an XML file that summarizes all the entities in all the reports in the collection. The file contains a `<REPORT>` node for each report and embedded `<PERSON>`, `<DATE>`, `<ORGANIZATION>`, etc. nodes for the entities within a report. The original report text is included as well.

In the descriptions of `Jigsaw` in the next section as well as in the scenario described later in the article and in the accompanying video, we use a report set from one of the Hughes' exercises [9]. We have altered some of the names and other entities from the already fabricated documents for further anonymization.

3 SYSTEM DESCRIPTION

3.1 Overview

`Jigsaw` provides an analyst with multiple perspectives on a document collection. The system's primary focus is on displaying connections between entities in the documents. We define entity "connection" by simple coincident appearance – two entities are connected if they appear in one or more documents together. Other, more semantically rich models of connection could be incorporated into `Jigsaw` as well, but this simple definition we have adopted seems to be both easy-to-understand and useful.

Jigsaw presents information about documents and entities through four distinct visualizations, called views. Each view provides a different perspective onto the data. The views, which will be discussed in more detail in the following sections, include:

- a tabular connections view containing multiple reorderable lists of entities in which connections between entities are shown by coloring related entities and drawing links between them

- a semantic graph view displaying connections between entities and reports in a node-link diagram, allowing analysts to dynamically explore the reports by showing and hiding links and nodes

- a scatter plot view giving an overview of the relationships between any two entity categories; a closer investigation over a smaller region is supported by range sliders

- a text view displaying the original reports with entities highlighted

User interaction with one view is translated to an event and communicated to all other views which then update themselves appropriately. Through such communication, different aspects of the reports can be examined simultaneously under different perspectives. Users can turn on and off event listening for each view depending upon whether they want the view to stay synchronized with the most recent interactions in other views. Users also can create multiple copies of each view type. This capability allows a view to be frozen at an interesting state (event listening turned off) while a new version of that view continues to receive events and update its state.

`Jigsaw` also provides a query interface for users to search for any entity as well as any string mentioned in the reports. When a query is issued, either the matching entities or the reports containing that string return as a result, and a message is dispatched to the listening views telling them to show the result(s).

Because analysts may want to take notes and draw diagrams to help clarify their thoughts during analysis, we have provided an authoring view within `Jigsaw` using the the Microsoft OneNote[TM] environment with a Wacom[TM] pen tablet as an input interface. OneNote provides freehand and structured editing of documents via pen input. Arbitrary regions of the other views

can also be captured and pasted as backgrounds for reference and note-taking.

We have found that the system is more useful when a set of views can be laid out and easily examined without window flipping and reordering. Due to the large amount of screen real estate required to display its views, Jigsaw ideally should be run on a computer with multiple and/or high-resolution monitors. We use the system on a computer with four displays as shown in Figure 1. Decreasing prices and smaller footprints of LCD monitors have made such configurations more common.
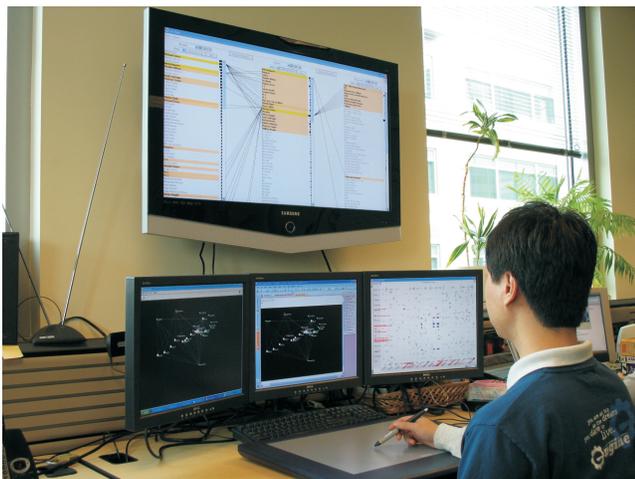
Figure 1: Jigsaw being used on a multiple-monitor computer with a pen tablet for note-taking.

3.2 Data Structure, System Architecture, and Event Messaging

Jigsaw is written in Java and adopts a model-view-controller architecture that separates the data (model) and user interface (view) components. As discussed in the prior section, Jigsaw reads an XML document that stores the extracted entities, tagged by their respective category name and bundled per report. Jigsaw creates Java objects for all entities and stores them in a general data structure model. This data structure is encapsulated in a class that provides an interface for the different view classes to call and retrieve entity-report information in order to build visualizations.

A controller class coordinates event communication from and to the views. Messages dispatched by views first go to the controller which then forwards the message to all listening views. (Recall that each view provides a button to enable/disable event listening.) Currently two types of events exist: select and show. A select event occurs when a user selects either an entity or a report in a view – such a selection is usually performed by a mouse click on the object. As feedback, the entity or report changes color or is highlighted visually. A show event occurs when a user explicitly indicates that an entity or a report should be displayed where it is not currently visible. Users can initiate show events by performing a particular mouse gesture or by issuing a search query.

Each of the four views interprets the two events differently and provides its own style of visual feedback. The Scatterplot View and the Text View are report based: reports are units of interaction and entities such as place or person are only shown in the context of a report. The List View and the Graph View, on the other hand, explicitly present entities as well as reports.

3.3 List View

The List View, illustrated in Figure 2, shows connections between sets of entities. Recall that two entities are "connected" if they appear together in one or more reports. The view consists of a number of lists of entity names. Each list contains all the entities of one specific type. The user can add and remove lists as desired – the number of lists is pragmatically constrained only by the horizontal space in the view. Once a list is displayed, a menu choice at the top allows the user to change the entity type shown in that list. Thus, even the same type of entity can be placed side-by-side in the view. The List View shown in Figure 2 contains two entity lists, persons and places.
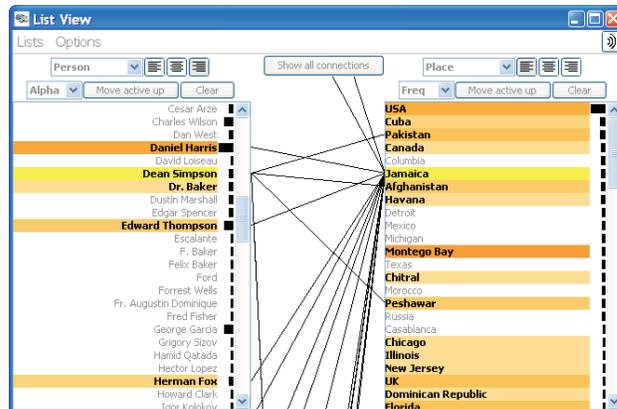

Figure 2: The List View. Selected entities are shown in yellow and connected entities are indicated by the joining diagonal lines and the orange shading. Darker shading represents stronger connections to the selected entities.

If a list of entities is too long for all the items to fit in the view, scrollbars appear to aid navigation. The items in a list can be sorted either alphabetically or by frequency of appearance in different reports in the document collection. This appearance frequency for an entity is represented by a small bar at the right end of each item in the list. A long bar indicates a high frequency and a short bar indicates a low frequency. In Figure 2, the person list is sorted alphabetically and the place list is sorted by frequency.

Entities in a list can be selected by a mouse click on the item and multiple selections are also supported. Selected entities are highlighted in bright yellow and all connected entities in all lists are highlighted in a shade of orange. The brightness of the highlighting on a connected entity indicates the strength of the connection: if the two appear together in only one report, a light orange is used, but if the two appear together in multiple reports, an increasingly dark orange is used as the number of co-appearances rises. Furthermore, the view draws lines between connected entities in adjacent lists to indicate the connection even further. The toggle button "Show all connections" above pairs of entity lists allows the viewer to see all connections at once instead of showing only the connections from selected items. Radio buttons at the top of each list also allow entity names in that list to be either left aligned, right aligned, or centered. This adjustment can help the viewer trace line connections. In Figure 2, two entities are selected and highlighted in yellow: "Dean Simpson" in the person list and "Jamaica" in the place list.

When a list of entities is long and requires scrolling, many connected items may not be visible in the view at any time. Thus, the List View also provides a mode in which all selected and connected entities are automatically moved to the top of the list (via the button "Move active up").

3.4 Graph View

The Graph View, illustrated in Figure 3, represents reports and their entities in a traditional node-link graph/network visualization common in many other systems. Both reports and entities are depicted as labeled circles. Reports are white and slightly larger than the other entities that follow the color mapping: people - red, places - green, dates - blue, and organizations - yellow. Edges from reports to the entities they contain are shown as well. Since entities appear only once in the view, this visualization portrays connections too: an entity in multiple reports will have edges connecting its circle to the white circles representing all those reports.
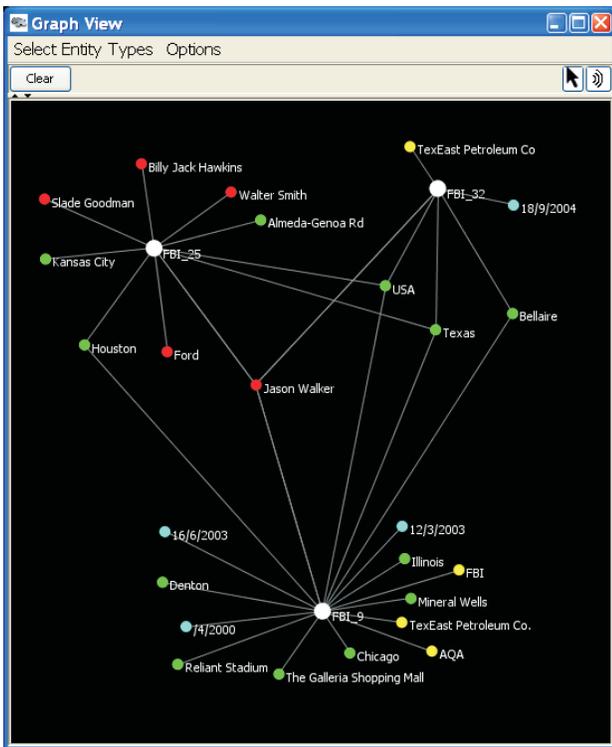
Figure 3: The Graph View. Reports are larger white circles and entities are smaller circles colored by type. Edges connect reports to the entities they contain.

Unlike graph visualizations such as Greenland [22] that present complete graphs consisting of large numbers of nodes, Jigsaw's view does not automatically draw all the reports and entities as one large network. We felt that a layout of such a large network would be overwhelming and difficult to understand, and thus would not be as helpful to the analyst in our context. Instead, Jigsaw's view is incremental. Show events place reports and entities on the display, and then mouse clicks on items can expand or collapse their connections. More specifically, expanding a report shows all its contained entities and expanding an entity shows all the reports in which it can be found.

The view uses a simple layout algorithm. Both report and individual entity nodes are randomly positioned in the plane when they are first shown. When all the entities of a report are first displayed as a group, they are drawn at random positions in a small circle around the report like satellites orbiting a planet. We have found that this simple layout provides reasonable drawings for Jigsaw's needs. In addition, the user can click on any entity or report and drag it to a new location. Dragging a report also moves all its connected entities already displayed that are not also connected to some

other report.

The entities-as-satellites graph visualization also provides another important connections view in Jigsaw since the user can see all the different entities mentioned in a report together. Furthermore, the visualization shows an entity mentioned in multiple reports via the lines drawn from the different reports to that entity. We have found the view to be useful in an interactive exploration mode – the user displays an initial report or entity, then expands the item to reveal its relations, and expands one of those items to reveal more, and so on. This type of interaction alternately reveals reports and connected entities.

A single click on an item simply selects it and dispatches a selection event to the other listening views. Selected nodes have a circle drawn around them. Multiple nodes can be selected via CTRL-key clicks or by rubber-banding a rectangular selection region. The system also provides an inverse selection operation that toggles the selected/unselected state of each node.

Other commands allow node(s) to be hidden (they retain the same position if they are subsequently shown again) and different types of entities to be filtered from the display. The viewer can remove all nodes from the view by using the "Clear" button.

3.5 Scatterplot View

The Scatterplot View, as shown in Figure 4, highlights pairwise connections between entities and it shows the reports containing the coincidences through a pseudo Starfield display [1]. The user specifies, through a pop-up menu on each axis, the entity type to be placed on that axis. All the entity names of that type then are (logically) listed along the axis in alphabetical order for people, places, and organizations, or chronological order for dates. If entities from each of the two axes appear together in a report, a diamond is drawn in the view at the conjunction of the two entity's positions along the respective axes. Since a report can contain more than one entity of the same type, multiple visual representations (diamonds) of the same report can appear together in the view at the same time.

Figure 4: The Scatterplot View. Each axis enumerates a list of entities. Diamonds in the center indicate reports containing particular pairs of entities, one from each axis at the relative x,y position.

Representative entity labels are drawn in a readable font size at equally spaced intervals along each axis to help the viewer. However, it is likely that many more entities exist in each category than can be shown this way. The view displays these other labels in a tiny illegible font size to provide a hint about the quantity of labels

missing. When the user moves the mouse pointer over an entity name, the scatterplot magnifies that item to be readable.

With a large set of reports, the display area can become cluttered with many diamonds representing those reports. To address that problem and help the viewer focus on sets of entities, the scatterplot view provides range sliders on each axis so that the viewer can zoom in on a segment of the axis. The view then updates to only show reports containing entities in that smaller range. We have found this capability particularly useful when dates are shown on an axis as a type of time-series view. The viewer can narrow the display to focus on a small interval of time.

The user can apply a particular color to a specific report. All instances of that report in the view are then shown in this color, even if the user changes entity types on the axes. This capability helps the user track information across varied display conditions.

3.6 Text View

Because the actual text of the reports is so important, `Jigsaw` includes a textual report view as shown in Figure 5. Multiple reports can be loaded into one Text View – the tabs at the top allow the user to select a particular report to display. All the entities in the report are highlighted in colors consistent with the color coding of entity types in the Graph View. A mouse click on an entity generates a selection event that is passed to the other listening views.
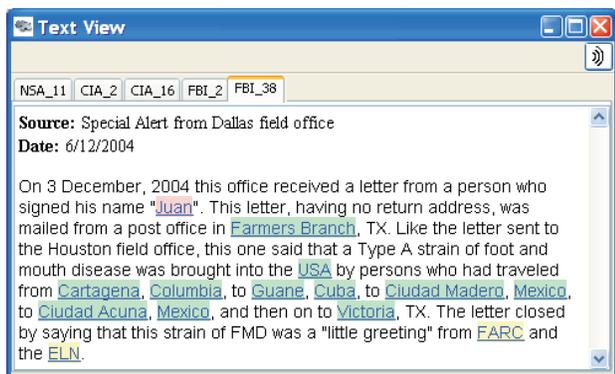

Figure 5: The Text View. Tabs indicate particular reports and the selected tab's report text appears below with entities highlighted and colored by type.

4 SCENARIO

In this section we walk through an analysis scenario with a fictional dataset to demonstrate how `Jigsaw` supports an analyst. Figure 6 illustrates relevant views from the scenario and the accompanying video demonstrates scenario actions as well.

Suppose that an analyst received information regarding a suspicious person named Michael Jones. To learn more about him, the analyst starts `Jigsaw`, opens the dataset, displays the List View, selects *Person* as the entity to be shown in the left list, and sorts the list by frequency. Michael Jones appears at the second position and the long bar next to the name indicates that Michael Jones is mentioned in a number of other reports. In order to explore people associated with Michael Jones, the analyst places *Person* entities in the second list as well and moves the people associated with Michael Jones to the top. The color mappings imply that Martin Clark has the strongest connection to Michael Jones since his name is colored in a dark shade of orange (see Figure 6, List View).

To verify this connection the analyst switches to the Text View to read the reports about Michael Jones. He is mentioned together

with Martin Clark in two reports (FBI_11 and FBI_35) and thus the connection seems plausible.

The analyst switches back to the List View, selects both Martin Clark and Michael Jones, and then puts *Organization* entities into a third list which reveals that both men have connections to the same organizations. The Revolution Now Scholarship Fund has the strongest connections of any organization, so the analyst continues the exploration on it.

The Text View shows two reports mentioning the Scholarship Fund. Report FBI_35 mentions that Michael Jones donated $48,000 to the fund on the stipulation that the donation be equally split among six students, Martin Clark being one of them. The analyst also notes that the six students form three pairs – where students in each pair live close to each other. This raises suspicions that the students might be collaborating.

Proceeding, the analyst brings up the three reports about Martin Clark and William Brown (who both live in Virginia) by selecting them in the List View. Two of the reports were already encountered in this investigation and the third, FBI_41, states that a month ago Martin Clark and William Brown took a cruise together from Hampton to Kingston, Jamaica. Furthermore, both are again on this cruise right now. The report also says that two other scholarship recipients, Thomas Taylor and Robert Johnson, took a cruise together from New York City to Montego Bay last month and they are also currently on this cruise again.

To more closely examine the chronology of events, the analyst selects the four students in the List View, switches to the Scatterplot View and displays *Date* entities versus *Person* entities. After zooming in to the relevant time range, the scatterplot shows the timeline of events for each of the students (see Figure 6, Scatter Plot View). Because she wants to save the current configuration of the List View and the Scatterplot View, the analyst halts event listening in them.

Now, to get a deeper understanding of the connections between the people and places, the analyst moves to the Graph View and displays report FBI_41. After expanding the node for report FBI_41 and filtering out the date entities, the analyst expands the nodes representing Kingston and Montego Bay. The view reveals that both are connected to three report nodes: FBI_14, CIA_10 and NSA_6. The analyst selects these reports and reads them in the Text View.
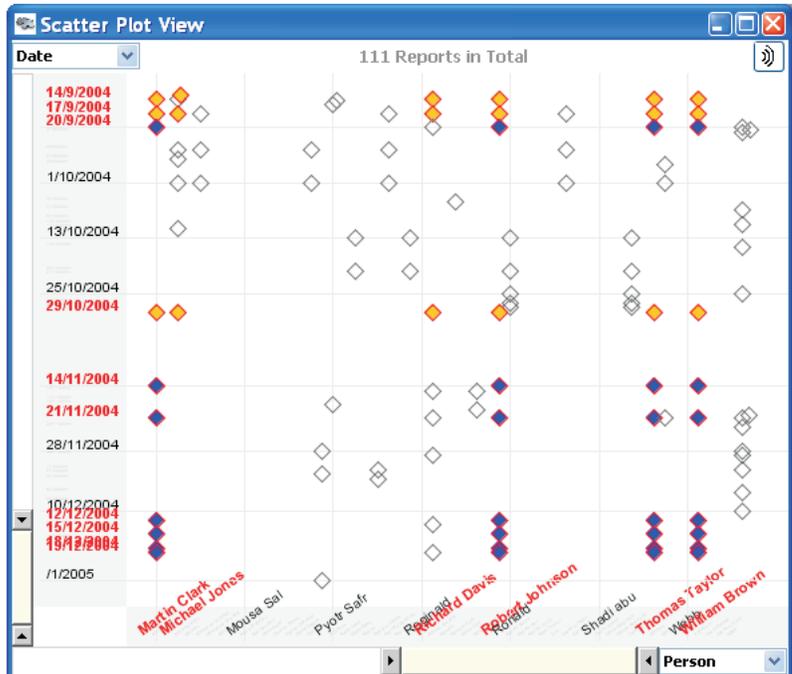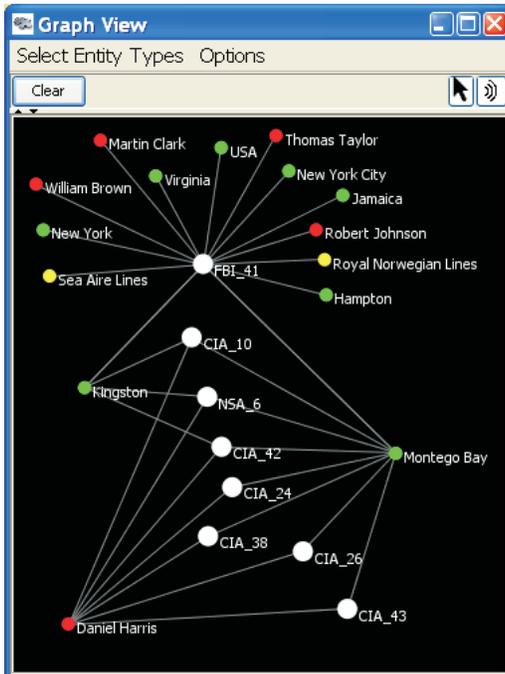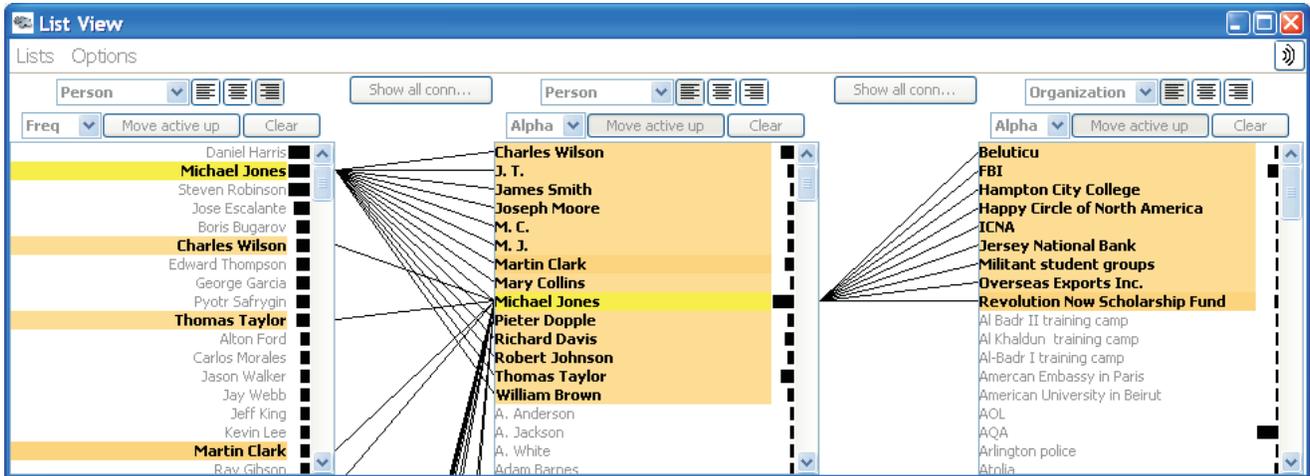
All three reports mention the person Daniel Harris who works in Montego Bay. The analyst issues a query on Harris, showing the man's entity in the Graph View. She expands his node and connections to seven more reports show up (see Figure 6, Graph View). Upon reading these reports, the analyst learns that Daniel Harris traveled from Montego Bay to Kingston on December 1st and passed a package to a person named Edward Thompson. The Scatterplot View shows that this date falls in the range of travel dates of the four students mentioned in report FBI_41.

The analyst concludes the investigation hypothesizing that suspicious activities are planned involving some of these individuals traveling on cruise ships in the Caribbean and with potential packages of interest. The analyst suggests that further investigation be conducted focusing on Daniel Harris and related activities.

5 RELATED WORK

A growing number of research and commercial systems are using visualization and visual analytic techniques to help support investigative analysis. WebTAS from ISS, Inc. [19] is focuses on temporal analysis and fusion of large, heterogeneous data sets. The system combines data mining techniques with a collection of visualizations including ones for link analysis, geographic and timeline representations.

Analyst's Notebook from i2 Inc. [10] provides a semantic graph visualization to assist analysts with investigations. Nodes in the graph are entities of semantic data types such as person, event, organization, bank account, etc. While the system can import text

Figure 6: Views from the example scenario discussed in the text and the accompanying video.

files and do automatic layout, its primary application appears to be analysts manually creating and refining case charts.

Oculus Info Inc. provides a suite of systems for different aspects of investigative analysis. First, `GeoTime` [13] is a system that can be used to visualize the type of report data discussed in this article. `GeoTime` visualizes the spatial inter-connectedness of information over time overlaid onto a geographical substrate. It uses an interactive 3D view to visualize and track events, objects, and activities both temporally and geospatially. Next, the `TRIST` system [12] allows analysts to formulate, refine, organize and execute queries over large document collections. Its user interface is a multi-pane view that provides different perspectives on search results including clustering, trend analysis, comparisons and difference. Information retrieved through `TRIST` then can be loaded into the the `SANDBOX` system [24], an analytical sense making environment that helps to sort, organize, and analyze large amounts of data. The system's goal is to amplify human's insights with computational linguistic, analytical functions, and by encouraging the analyst to make thinking more explicit. The system offers interactive visualization techniques including gestures for placing, moving, and grouping information, as well as templates for building visual models of information and visual assessment of evidence. An evaluation experiment of the `SANDBOX` system showed that analysts using the system did higher quality analysis in less time than using standard tools. `Jigsaw` provides a different style of visual representation of document entity data to analysts; `TRIST` and `SANDBOX` provide more authoring and organizational infrastructure.

`IN-SPIRE` [20] is a system for exploring textual data in document collections. It generates a "topical landscape", either through a 3D surface plot or a galaxy-style view, that supports queries, provides the possibility to analyze trends over time, and allows analysts to discover hidden information relationships among documents. Its goal is to identify and communicate the different topics and themes, and then allow the analyst to inspect the documents more deeply through interactive analysis. `Jigsaw` differs in its focus on exploring relationships among the entities in documents.

Sanfillipo and colleagues at PNNL [16] introduce a system that extracts scenario information from unstructured intelligence data sources. Their system provides multiple views on multiple monitors as does Jigsaw, but it focuses more on language analysis and ontologies to help identify the scenarios and on evidence marshalling views for constructing hypotheses.

Also from PNNL, Wong *et al*. [21] developed the `Have Green` framework, an interactive graph exploration environment. It supports analysts in comprehending and analyzing large semantic graphs that represent concepts and relationships through its powerful analytic capabilities.

The `ENTITY WORKSPACE` [3] is a tool to amplify the usefulness of an traditional evidence file that is widely used by analysts to keep track of facts. It provides an explicit model of important entities to help the analyst to find and re-find facts rapidly, discover connections and identify important documents and entities to continue the exploration. The system is just one of a suit of tools from PARC directed at assisting sense-making [5].

`Jigsaw` differs from the above systems in its focus on representing connections and relationships between entities in document collections. Also, it provides a system model where user interaction is a first-class object, helping to expose the entity connections, and providing for easier extensions to new styles of views.

Probably the closest system to our work is the `KANI` [7] project that includes a component for visualizing entities from textual documents. `KANI` has two main views, a document viewer that highlights entities and their selected relationships and a graph view that shows different entities connected in a node-link structure. The system provides extensive filtering capabilities to the analyst and includes automated associate components that help with activities like hypothesis refinement and assumption testing. `Jigsaw` goes beyond `KANI` in the variety and style of the interactive visualizations provided, but `KANI` has a more complete infrastructure for supporting reasoning and hypothesis formulation.

6 DISCUSSION AND FUTURE DIRECTIONS

While `Jigsaw` provides a number of capabilities that we believe will be useful for investigative analysis, our work has only begun to scratch the surface of what is possible in this area. Numerous avenues of research and extensions to the system are possible in future work. In fact, we have many already underway.

Because the system has yet to be evaluated, that is an obvious missing element. Evaluation should range from basic usability assessments of the views to trial use of the system by real analysts. Their feedback can drive changes and additions to the system.

As mentioned earlier, we have largely avoided the challenging issue of entity identification and extraction. Instead, we are presently exploring external tools that can be used in this process. More broadly, `Jigsaw` must escape its current more batch-oriented model in which entities are extracted from reports a priori and the resulting entity collection is visualized in the system. Instead, entity extraction should be integrated more dynamically within `Jigsaw`. Analysts should be able to read in new reports and remove reports even after analysis with the system has begun. The set of entities and reports visualized in `Jigsaw` should update to reflect dynamic additions, removals, and consolidations. Furthermore, analysts should be able to manually identify entities missed by the automatic analysis.

Closely related to this issue is the challenge of scalability. For larger report collections in which the number of entities in a category can grow into the thousands or beyond, some form of dynamic update and filtering is absolutely necessary. The examples examined with `Jigsaw` so far are modest in size with at most hundreds of entities in a set. Obviously, when the number of entities in a category moves past that, the List View and Scatterplot View which show enumerations of all entities become less useful. Allowing analysts to selectively import reports and/or entities is a logical way of proceeding. Thus, the List and Scatterplot views could function more like the present incremental Graph View: only queried or selected entities are shown.

Investigative analysis often involves information of questionable validity or with estimated likelihoods of probability. Presently, `Jigsaw` has no way of representing such information.

Our own trial use of the system while exploring the example report collections has identified a number of potential enhancements and improvements, many of which already have been implemented. Other potential additions range from detailed low-level operations for individual views to broader, analytic capabilities. For example, when an entity such as a place is chosen in the List View, connected entities such as people are highlighted. `Jigsaw` needs a simple way to then show all those people in the Graph View. The node positioning algorithms in the Graph View could be improved to better use space in that view as well.

Our use of the system also has suggested the need for dedicated geographic and time-series views. In sample analysis sessions, we have noted the absence of explicit views supporting those two perspectives. Of course, adding even more types of views raises issues concerning the multiplicity of views – could an abundance of representations overwhelm analysts rather than assist them? How many different kinds of views can profitably be used together?

Trial use of the system also suggests the need for better tools to help analysts organize their thoughts and document the models and plans they are constructing. Presently, analysts must use independent tools such as pencil-and-paper or the included OneNote[TM] interface in order to capture notes and thoughts. The `ENTITY WORKSPACE` system [3, 5] from PARC suggests a number of in-

teresting evidence-marshalling ideas here as does the Shared Reasoning Layer of KANI [7].

Jigsaw embodies a more structured style of analysis because it operates on categorized entities extracted from plain text reports. Integrating the system with others that provide analysis of unstructured text, such as IN-SPIRE [20], might help analysts form hypotheses by exposing the main concepts and themes across the document collection.

Finally, the need for better tools to augment the process of documenting and presenting the results of analysis has been identified [17]. Perhaps the views within Jigsaw could be captured and annotated to provide visual summaries of the evidence used to reach actionable conclusions.

7 CONCLUSION

Every day investigative analysts are faced with the challenging task of assessing and making sense of large bodies of information. Technological aids that promote data exploration and augment investigators' analytical reasoning capabilities hold promise as one way of assisting analysis activities [5, 17, 24]. In a workshop of intelligence analysis professionals, working groups generated a list of the top ten needs for intelligence analysis tool development. One item was "Dynamic Data Processing and Visualization" that was further elaborated as follows:

"Solutions are needed that transcend what is typically described as "visualization" – in contrast to a predominantly "passive" relationship between the system that displays complex visualizations and the analyst who still must digest and interpret them. What is needed is a much more interactive and dynamic relationship in which the analyst is better able to explore the information within the visualization." [2]

Herein we present Jigsaw, a system designed to assist analysts with foraging and sense-making activities across collections of textual reports in just this manner. Jigsaw presents a suite of views that highlight connections between entities within the reports. Through interactive exploration, analysts are able to browse the entities and connections to help form mental models about the plans and activities suggested by the report data.

Jigsaw is not a substitute for careful analysis of the reports, however. Instead, it acts as a visual index that presents entity relations and links in forms that are more easily perceived, thus suggesting relevant reports to examine next. Other systems sometimes put too much information into a single complex view, with the result that though information may be present, it is harder to discern and is much less flexible from the analyst's viewpoint. Our approach hinges on multiple, easy-to-understand views with simple, clear interactions. In creating the visualizations we leveraged well-known representations from the field of information visualization and augmented them with interactive operations useful for showing connections between entities. The synthesis of all the views and their interactive capabilities that provide an environment for aiding investigative analysis is the main contribution of the research.

ACKNOWLEDGEMENTS

REFERENCES

[1] C. Ahlberg and B. Shneiderman. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *ACM CHI '94*, pages 313–317, April 1994.

[2] R. V. Badalamente and F. L. Greitzer. Top Ten Needs for Intelligence Analysis Tool Development. In *2005 International Conference on Intelligence Analysis*, May 2005.

[3] E. A. Bier, E. W. Ishak, and E. Chi. Entity Workspace: An Evidence File That Aids Memory, Inference, and Reading. In *IEEE International Conference on Intelligence and Security Informatics*, pages 466–472, May 2006.

[4] S. Card and P. Pirolli. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis. In *2005 International Conference on Intelligence Analysis*, May 2005.

[5] S. K. Card. Leverage Points and Tools for Aiding Intelligence Analysis. Unpublished report presented at the 2007 Human-Computer Interaction Consortium, 2007.

[6] S. K. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[7] A. R. Chappell, A. J. Cowell, D. A. Thurman, and J. R. Thomson. Supporting Mutual Understanding in a Visual Dialogue Between Analyst and Computer. In *Human Factors and Ergonomics Society 2004*, pages 376–380, September 2004.

[8] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[9] F. J. Hughes. Discovery, Proof, Choice: The Art and Science of the Process of Intelligence Analysis, Case Study 6, "All Fall Down". Unpublished report, 2005.

[10] i2 - Analyst's Notebook. http://www.i2inc.com/, 2007.

[11] P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins, 2002.

[12] D. Jonker, W. Wright, D. Schroh, P. Proulx, and B. Cort. Information Triage with TRIST. In *2005 International Conference on Intelligence Analysis*, May 2005.

[13] T. Kapler and W. Wright. GeoTime Information Visualization. *Information Visualization*, 4(2):136–146, 2005.

[14] A. Mikheev, M. Moens, and C. Grover. Named Entity recognition without gazetteers. In *Proceedings of EACL*, pages 1–8, 1999.

[15] D. Norman. Visual Representations. In *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley, 1994.

[16] A. Sanfilippo, B. Baddeley, A. J. Cowell, M. L. Gregory, R. Hohimer, and S. Tratz. Building a Human Information Discourse Interface to Uncover Scenario Content. In *2005 International Conference on Intelligence Analysis*, May 2005.

[17] J. J. Thomas and K. A. Cook. *Illuminating the Path*. IEEE Computer Society, 2005.

[18] IEEE VAST 2007 CONTEST. http://www.cs.umd.edu/hcil/VASTcontest07, 2007.

[19] WebTAS. http://www.webtas.com/, 2007.

[20] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *IEEE Symposium Information Visualization (InfoVis)*, pages 51–58, October 1995.

[21] P. Wong, G. Chin, H. Foote, P. Mackey, and J. Thomas. Have Green - A Visual Analytics Framework for Large Semantic Graphs. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 67–74, October 2006.

[22] P. C. Wong, H. Foote, G. C. Jr., P. Mackey, and K. Perrine. Graph signatures for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1399–1413, 2006.

[23] P. C. Wong and J. Thomas. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[24] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The Sandbox for analysis: concepts and methods. In *ACM CHI '06*, pages 801–810, April 2006.