Visual Analytics for Investigative Analysis

Thomas Absenger, Mohammad Chegini, Thorsten Ruprechter, Helmut Zöhrer

Graz University of Technology
A-8010 Graz, Austria

17 May 2017

Abstract

This survey aims to provide a general overview of tools used for visual investigative analysis. The amount of data available for investigative purposes has reached huge proportions in the last decade. Therefore, methods and tools must be able to cope with the large amount and heterogeneity of data. Driven by the VAST challenges, several tools providing visualizations and analysis have been developed. These tools attempt to assist humans in the process of understanding information existent in data. Using such assistance, investigators are able to detect patterns and anomalies. Some currently available investigative analysis tools will be examined and tested to give the reader an overview as well as a comparison over the course of this survey.

Contents

Credits

This survey was created for the master's course "Information Visualization" at Graz University of Technology and is based on a skeleton provided by courtesy of K. Andrews [2012].

List of Figures

List of Tables

Chapter 1

Introduction

As this report aims to discuss and compare a selection of tools in the field of Visual Analytics for Investigative Analysis, these terms require a proper definition first.

1.1 Visual Analytics (VA)

In order to comprehend K. Andrews [2017] definition of Visual Analytics, one needs to grasp some essential concepts of visualization first:

- Information Visualization: It deals with abstract structures, like hierarchies, networks, and multidimensional spaces.
- Geographic Visualization: It implies a map-based visualization approach where data usually consists of two- or three-dimensional coordinates.
- Data Visualization: This term describes the combination of Information Visualization and Geographic Visualization.

K. Andrews [2017] describes Visual Analytics (VA) as a Data Visualization frontend which is supported by an analytics backend. VA tools and techniques are usually used "to synthesize information; derive insight from massive, dynamic, and often conflicting data; detect the expected and discover the unexpected; provide timely, defensible, and understandable assessments; and communicate assessments effectively for action", as described by IEEE VAST symposium [2006]. The VA field has seen great interest with the rise of information overload which can be turned into a great opportunity. The sciences which are a necessity for VA go beyond those used for usual information visualization. Numerous fields of mathematics, knowledge representation, cognitive and perceptual sciences, as well as management and discovery sciences are significant parts of it.

1.2 Visual Analytics for Investigative Analysis (VAIA)

Adding investigative analysis to VA as a specification means going beyond the typical tasks of pure information visualization, such as finding correlation and spotting outliers combined with the possibility to make use of numerous kinds of data. Visual analyists for investigative analysis seek to develop hypotheses and understand the data thoroughly. What used to be an intense thinking process for highly talented detectives – putting the pieces together and connecting the dots – is tried to be simulated and facilitated [Kang et al., 2011].

The heterogeneity of input data is an important aspect of VAIA, as a much wider variety of data can be taken into account for investigating a case. This will be pointed out further in the section on the test data for the authors' research.

1.3 VAST Contest / Challenge

Developing an insight into various datasets is not only a challenging task, but also hard to quantify. In order to enhance the field of VA and to find metrics which make a comparison possible and useful, experts in the field found that holding an annual competition might be the solution. Therefore, the first visual analytics science and technology (VAST) contest was held in 2006 in conjunction with the 2006 IEEE VAST Symposium [Grinstein et al., 2006]. The VAST challenge offers the chance for developers of tools to see how well their product behaves in real world situations and how it would be rated compared to other developers. But not only developers of free software for the contest's sake, but also vendors may participate and get the possibility to test and re-evaluate their software. As a positive by-product of the initiation of a competition like the VAST Challenge, a benchmark repository of VA tools has emerged [SEMVAST, 2017]. This rich repository collection of winners and participants from the last eleven years deals as basis for the further research and comparison in the following sections.

The very first conducting of the VAST Challenge in the year 2006 plays an important role for this research work. That is because the challenge files from this particular year were used as test data in order to compare tools. The input data which consists of numerous heterogeneous files were entirely made up for the purpose of the competition. It is set in a fictitious town called Alderwood, where shenanigans have taken place. The dubious intrigues are well documented and are produced to give the tool or the user a realistic opportunity to solve the riddle. How well each of the tools performed individually and how they were evaluated can be read in the following sections of this report.

The test data's heterogeneity can be confirmed by this direct listing of all its contained files [Haack et al., 2006]:

- about 1200 news stories plus a few other items collected by the previous investigators
- a few photos
- maps of Alderwood and vicinity (in bitmap image form)
- a few files with other mixed materials, e.g. a spreadsheet with voter registry information or a phone call log (all provided with descriptive information)
- a couple of pages of background information (in text form).

Chapter 2

Tools

The main focus of this survey is to offer an overview of existing tools for visual analytics in the context of investigative analysis. In this chapter, each tool will be described in adequate detail to highlight its capabilities. For each tool we try to at least have a look at the following points:

- A general description of the tool.
- Under which license is the tool available? Is it open source?
- How did the tool perform in the VAST challenges?
- The main features of the tool.
- The limitations of the tool.
- A subjective assessment of the capabilities of the tool.

Not all tools can be examined at the same level of detail as most of the tools are not available for manual testing. Additionally, the amount of resources available for each tool varies greatly.
In the following sections each tool will be described, starting with the most prestigious tool (nSpace), followed by tools which are freely available. The other tools are arranged in no particular order.

2.1 nSpace

In 2006 the company Oculus Info (now called Uncharted Software) published the first version of their tool nSpace. It was created on the basis of the Novel Intelligence in Massive Data (NIMD) research program. NIMD focused on massive data and ways to visualize, investigate and interact with it.
When results were published in 2002, no real system satisfying an analyst's needs existed and working with large amounts of (often heterogeneous) data was cumbersome and included the use of several different tools which were not built for the purpose of data analysis. They proposed using a system of systems combining all relevant components as a possible solution [Jonker et al., 2005].

As nSpace built on top of these results it was designed as a system of systems, each subsystem being one component. Results about the components were published before releasing the whole software. Two main components can be identified which will be described in more detail later:

- The Rapid Information Scanning Tool (TRIST)
- Sandbox

In 2008, a web version of nSpace titled nSpace2 was released. At first, it supported only parts of the functionality of the stand-alone nSpace but was extended to have full feature-support (and beyond). Additionally, the TRIST component received a visual overhaul [Chien et al., 2008].

2.1.1 Licensing

Unfortunately, nSpace is only available for a one-time license fee (and an annual maintenance fee). For this reason, it could not be tested. Even though Oculus Info stated that nSpace2 is open source [Proulx and Canfield, 2015], no source code or any hints about it could be found online.

2.1.2 nSpace and the VAST Challenges

nSpace has won several VAST challenges and awards, starting with the first VAST contest in 2006. Its success can be seen easily in the list of awards they have won [SEMVAST, 2017]:

- VAST 2006 Contest: First Place, Corporate Category
- VAST 2007 Contest: Winner Corporate Category
- VAST 2008 Grand Challenge: Award for Support for Diverse Analytic Techniques
- VAST 2010 Mini Challenge 1: Award for outstanding analysis and accuracy
- VAST 2011 Mini Challenge 3: Award for good analysis & support debrief

Even though nSpace was very successful in the VAST challenges, Oculus Info (now called Uncharted Software) has published an article about how participating in the VAST challenges is not about winning, but about learning from the experience. They state that the challenges have had a huge impact on their software as it could be tested on a wide range of problems, each challenge highlighting shortcomings of nSpace [Proulx and Canfield, 2015].

Besides the iterative improvements of nSpace through the challenges themselves, one of the reasons nSpace is so successful seems to be its completeness. Compared to the other tools in this survey, it supports most (if not all) steps necessary for performing an investigative analysis and can be used as a stand-alone tool. Other tools seem to have more fixed roles in the pipeline of investigative analyses.

2.1.3 Components

As already stated, the architecture of nSpace consists of various components which interact with each other. The component connecting all other components is the Pasteboard. It is used for making selected data available to other components and in this sense acts as a common interface between them.

The typical work-flow can be described as follows:

- Load and retrieve data for use in the TRIST component
- Triage the data within the TRIST components
- Select important or interesting parts of the data and move them to the Sandbox (via Pasteboard)
- Connect the evidence in the Sandbox and build hypotheses (no automated support by the software)
- Repeat until satisfied

TRIST and Sandbox will be described in the following sections.

The Rapid Information Scanning Tool (TRIST)

The Rapid Information Scanning Tool (TRIST) was introduced in 2005 and is the most powerful component of nSpace. Its main purposes are information retrieval and triaging data [Jonker et al., 2005]. A screenshot showing TRIST can be seen in figure 2.1. It highlights the most important aspect of TRIST: multiple linked views displaying different dimensions of the same data source.

To achieve its purposes, it supports various useful features:

- Automatic entity extraction
- Custom queries (including natural language queries)
- Linked views showing different data dimensions
- Compact visualization of results and entity relations
- Document reader for viewing a multitude of data types

Sandbox

The Sandbox component is a workspace used for organizing evidence. It can be seen as a digital pinboard and is used in analogous way. Interesting data points can be dragged around and connected and always remained linked to TRIST and their original data source. Wright et al. [2006] state that analysts using Sandbox have greater productivity as opposed to other (not named) tools.
A screenshot of the Sandbox as used in the VAST 2006 contest can be seen in figure 2.2.

2.1.4 Features

In addition to the features listed for each component of nSpace some additional functionalities can be identified:

- Components can be addressed via WebService APIs, allowing for integration of other tools.
- Data cannot only be imported from local files but from online sources as well.
- nSpace supports collaboration between multiple users and clients.

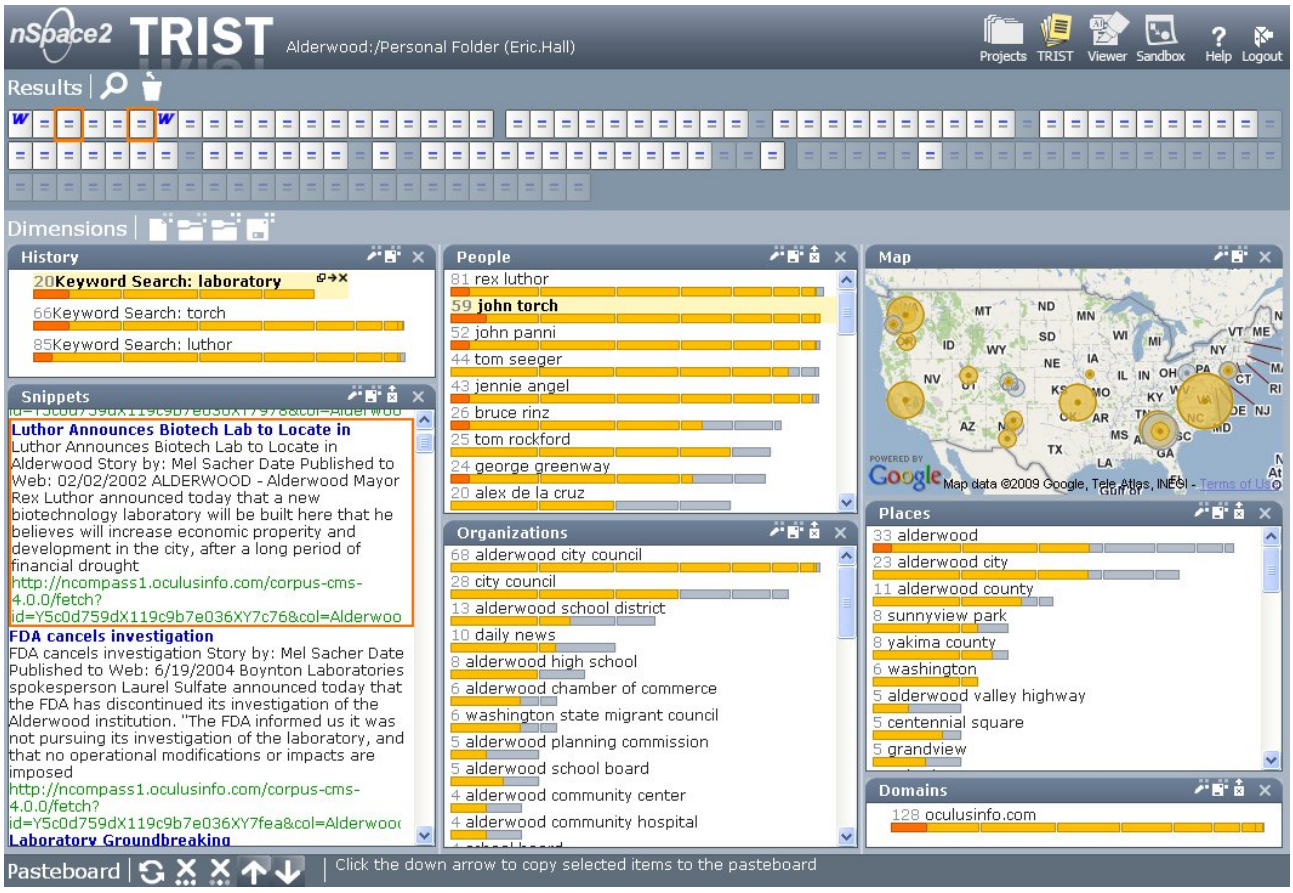

Figure 2.1: The Rapid Information Scanning Tool (TRIST) is mainly used for triaging data. In this screenshot one can see multiple linked views showing extracted entities and selection-dependent highlighting. [Image extracted from Uncharted Software [2017]. © 2017 Uncharted Software. Used under the terms of Austrian copyright law: §42f]

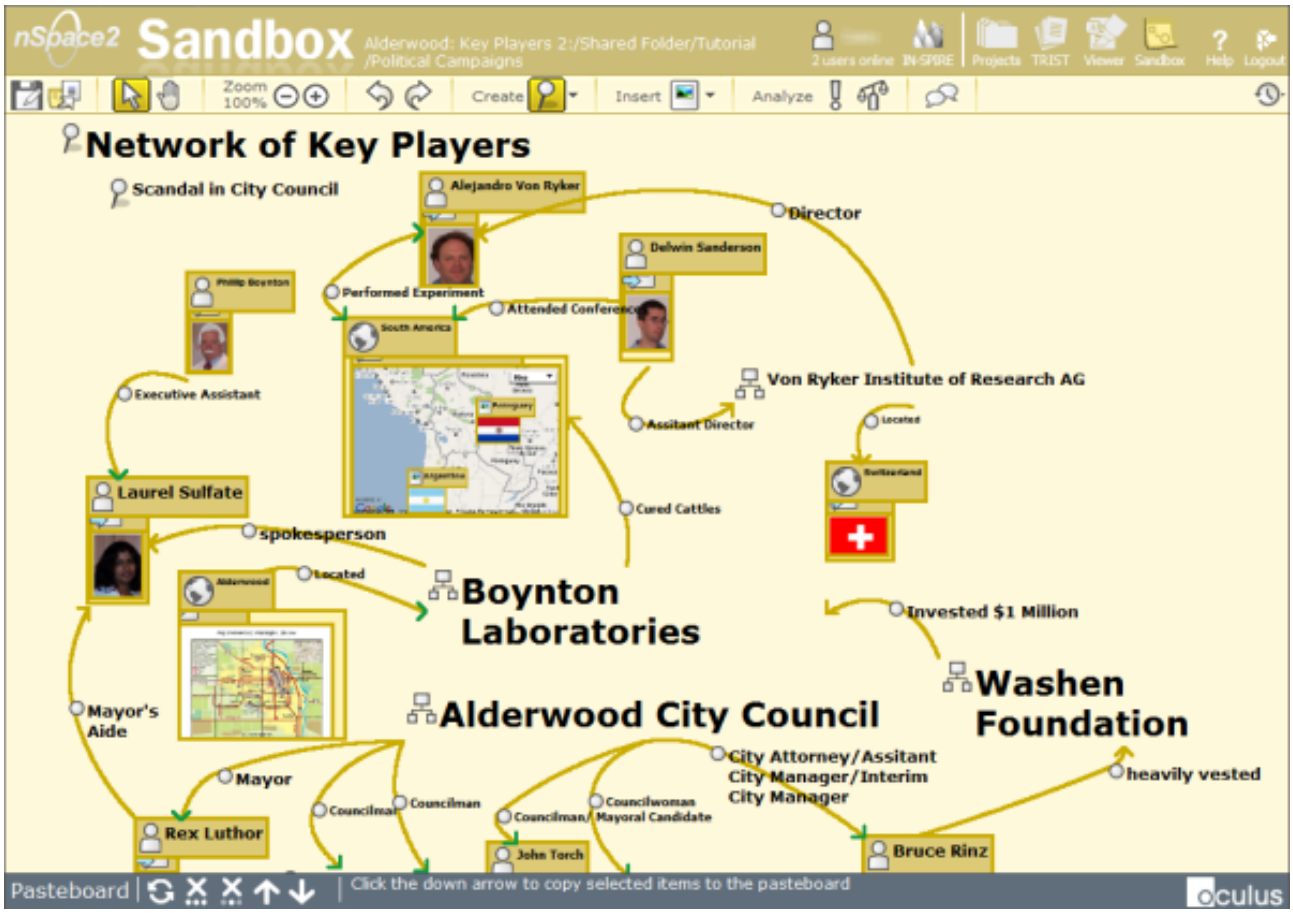

Figure 2.2: The Sandbox component of nSpace is a workspace used for organizing evidence and generating hypotheses. In this screenshot some items are arranged and connected to support a hypothesis for the VAST 2006 challenge. [Image extracted from Uncharted Software [2017]. © 2017 Uncharted Software. Used under the terms of Austrian copyright law: §42f]

2.1.5 Limitations

One of the key limitations of nSpace is the lack of visualization options. As far as can be extracted from the resources available, nSpace supports only the visualization which can be seen in figure 2.1 which only visualizes entity relations with bar charts and geographical visualization.
In the various VAST challenges nSpace was used, it mostly relied on other tools for visualizing the data. The most frequently used tool is GeoTime which is developed by Oculus Info as well [SEMVAST, 2017].

2.1.6 Subjective Assessment

nSpace seems to be the most complete and sophisticated tool of all the tools we examined. The various awards it has won in the VAST challenges show its versatility. For a tool which claims completeness as a system of systems it is however surprising to see the lack of visualization options. Still, the only real downside seems to be the license fee.

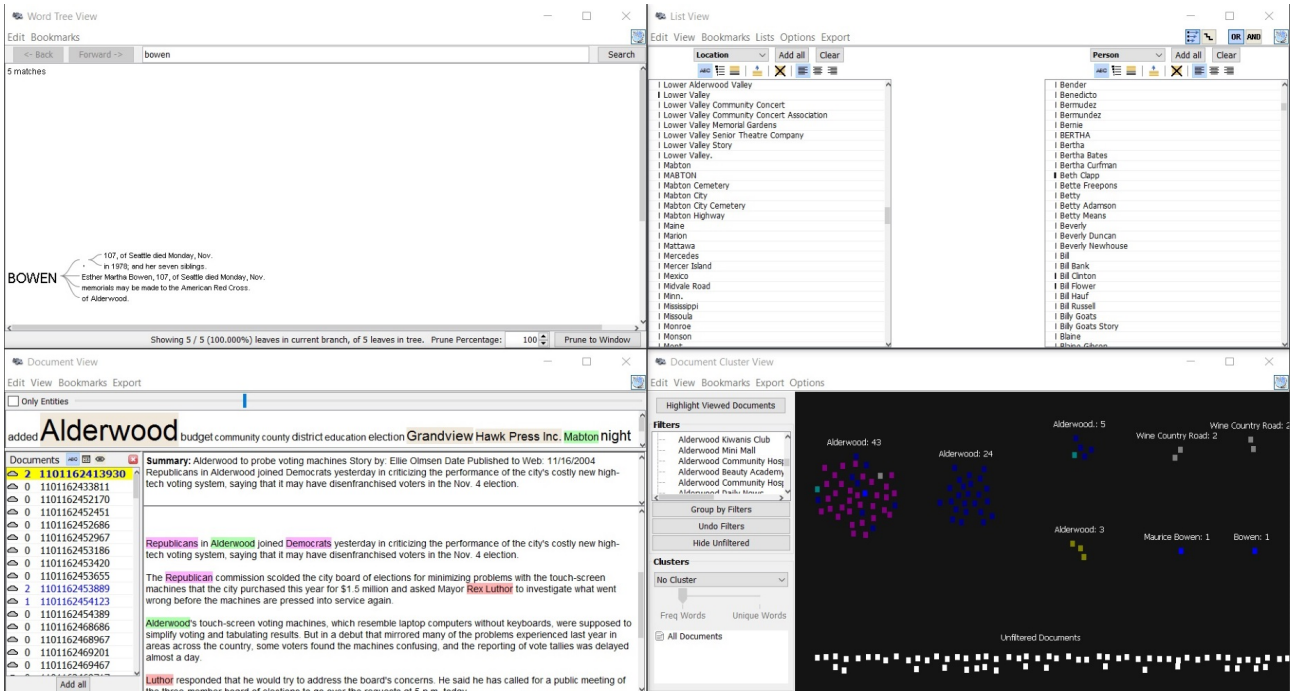

Figure 2.3: At the bottom left of the figure, the document view is shown. The user can explore documents and also observe text clouds of the most popular entities. The cluster view is shown down to the right of the picture, while in the upper right corner a list view is visible. Finally, a word tree view is shown on the upper left side. [Screenshot created by the authors of this survey using Jigsaw]

2.2 Jigsaw

Dashboard

Jigsaw is a tool for investigation analysis that was introduced in 2007 [Gorg et al., 2007]. This tool uses collections of text documents and then analyses and visualizes them in different forms. Jigsaw participated in the VAST challenge 2007 and won the university division of that contest. Jigsaw is more suitable for providing a quick overview of all documents. After importing documents into the system, Jigsaw will extract entities using a chosen algorithm. These entities will later be used to explore data more easily using various views. An overview of Jigsaw is presented in figure 2.3. The shown figure is created using the VAST challenge 2006 data set.

2.2.1 Accepted Formats and Data

As mentioned before, Jigsaw is more of a text analysis tool. Despite the fact it accepts different formats like textual data such as plain texts (TXT) or Portable Document Format (PDF) files in addition to structured data (CSV), it just loads text data from these documents. For example, if there is an image in a PDF-file, it is simply ignored by the software.

2.2.2 Licensing

Jigsaw is free to use, but the source code is not available. In the license agreement of Jigsaw, it is mentioned that the product owners can withdraw any rights of using the software, which means the user should delete the free version of the software from his computer. Also, the user is not allowed to re-engineer the software for any purposes.

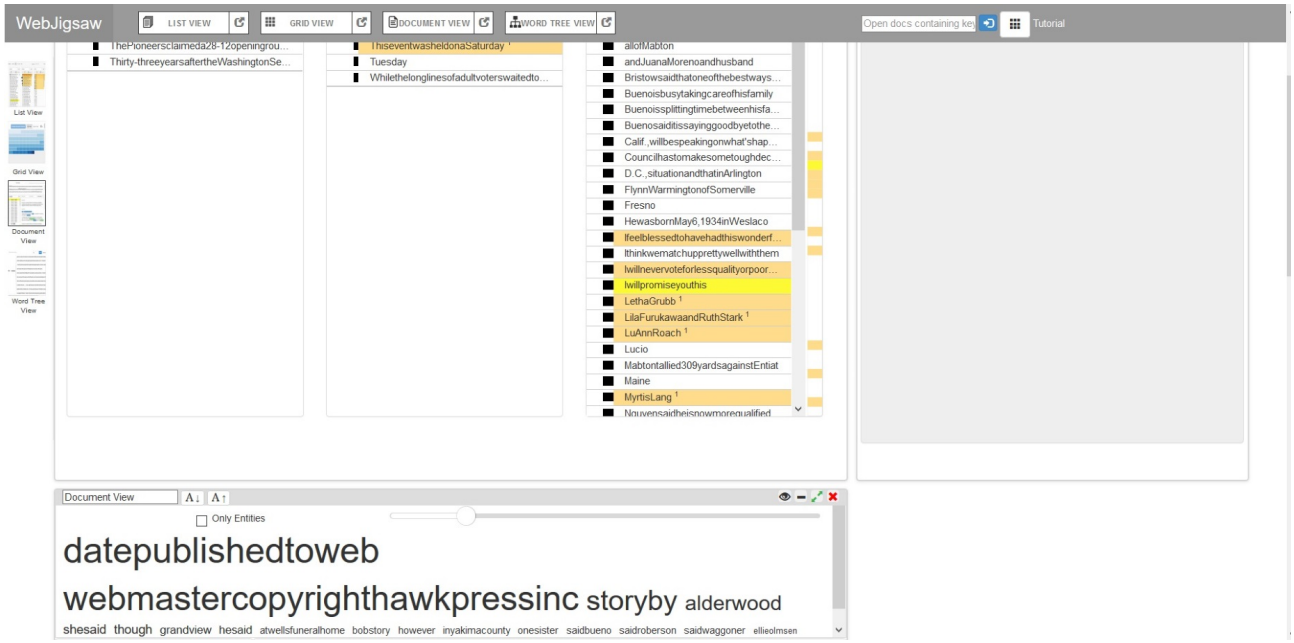

Figure 2.4: WebJigsaw offers the same features as Jigsaw in a web browser. Top right a list view is shown, while on the bottom left the document view is visible. [Screenshot created by the authors of this survey using Jigsaw]

WebJigsaw

There are two versions of Jigsaw which basically do the same thing. First, the standard version of Jigsaw, and second, WebJigsaw. Regular Jigsaw is a Java program distributed as a Java archive (JAR) that runs on any operating system. WebJigsaw represents a web application and offers the same features as the Java version. Although uploading all text documents in WebJigsaw correctly is slightly complicated, preprocessing and extraction of the entities from the data is faster as in the standard version. Figure 2.4 shows an example view of this web application.

Workflow

Importing and making use of data with Jigsaw is performed in four major steps. Figure 2.5, which is extracted from WebJigsaw, demonstrates these steps. First, user imports files into the software. These files will be used to extract entities and entity analysis. Entity analysis phase is supported by different algorithms. We used Spacy algorithm for VAST challenge 2006 data set and it works nicely. After that, the computer or the server will start to compute the entities. This will take some time, depending on the data set. At the end, the user can select different views to visualize data. These views help the user to observe various aspects of documents.

2.2.3 Limitations

For testing, the VAST challenge 2006 dataset was imported into Jigsaw. The entity extraction for the whole dataset took two hours on a normal PC (18GB Ram and Intel Core i7 CPU 3.20GHz). All the views could be opened without any problems. The most important limitation of Jigsaw is its lack of supporting other document formats like pictures and videos. Jigsaw is more suitable for documents which do not have many pictures, like news and academic papers.

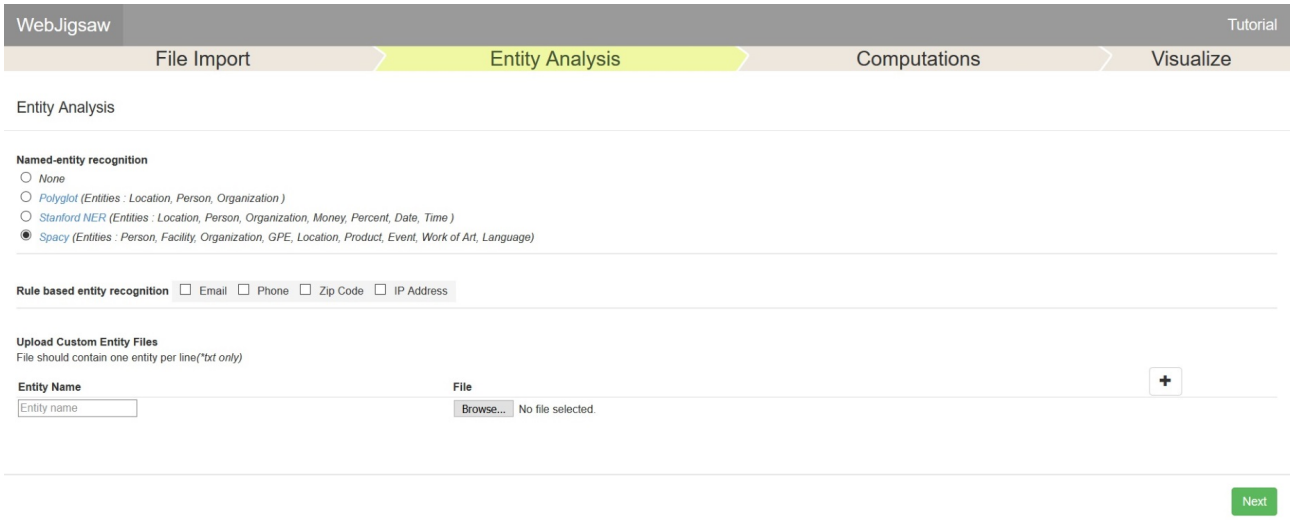

Figure 2.5: Jigsaw work flow as visible in WebJigsaw. The workflow contains import, entity extraction, computation and then visualization. [Screenshot created by the authors of this survey using Jigsaw]

2.2.4 Subjective Assessment

Jigsaw is probably one of the most prestigious, freely available tools for investigation analysis. It is free of cost and fairly simple to use. The learning curve of Jigsaw is flat and although the user interface does not follow modern guidelines, it is powerful and simple to use. The program is fast and only slow when loading lots of documents at the same time. Although Jigsaw uses text documents and does not offer further features for extracting and visualizing entities from those documents, it is a user-friendly and powerful tool for pre-processing and providing an overview over a textual data set.

2.3 Visallo

This tool for investigative analysis is realized as a web application. Its core is freely available and Visallo's source code is accessible on Github under an open source license [Visallo, 2017b]. The company producing this tool also names itself Visallo, however it was also known as V5 in the past. Even earlier, it was called Altamira Technologies Corporation. Although the web application running on a standalone server is available for free, the company offers support and guidance for concrete use cases as well as extensions to the Visallo product core. These additional services presumably bring costs for the customer with them, which account for the company's revenue. Information about how much a customer has to approximately pay for such additional services is not officially available.

2.3.1 Ontologies for Data Structuring

An unique property of this tool is that it mainly uses ontologies to manage the structure of its data as well as the data itself. In context of the semantic web, ontologies are mainly used to specifically define properties of data entities and their relationship to each other. Ontology information is encoded using an XML-based format building on the Rich Document Format (RDF) [W3C, 2017]. Some example file formats which are usually used as a container for ontologies and their corresponding data are Ontology Web Language (OWL), Rich Document Format (RDF), or N-triples (NT).

While Visallo is also able to import basic file formats such as tables or plain text, it works best using above-mentioned formats. Once defined, such ontologies are very efficient in structuring data. However, defining and maintaining these rather complex structures may lead to difficulties which will be further discussed in section 2.3.4.

2.3.2 Licensing

One of the most appealing factors of Visallo is that its product core is available as open source. The precise license under which it is obtainable is called "Apache License, Version 2.0" [Apache Software Foundation, 2004]. This license indicates that anyone can use the software's source code for any purpose if the necessary notices are contained in the end product.

Extensibility

The aforementioned product core is extensible. Developers can use their self-written code to expand the tool's usage possibilities beyond its core functionalities. In addition to that, the company itself offers its customers to enhance the tool for them. Visallo provides extension points for both front-end and back-end. While the back-end offers extension points for functionalities such as import and export as well as additional analytic tools or customized search queries, the front-end can for example be enriched via additional user interface widgets. One of the more popular ways of extending Visallo are web plugins, which combine front-end and back-end functionality. More detailed information regarding code development for this tool can be found under the official developer documentation, which also includes tutorials [Visallo, 2017a].

2.3.3 Features

Visallo mainly focuses on graph structures and maps for investigative analysis. Although, as mentioned above, extensions to this core functionality are possible, this survey will focus on the freely available aspects.

Example Dataset

The main functionalities and features displayed here will be demonstrated using data about an insider threat scenario. This example scenario is hand-crafted by Visallo to elaborate the purpose of their tool [Visallo, 2015].

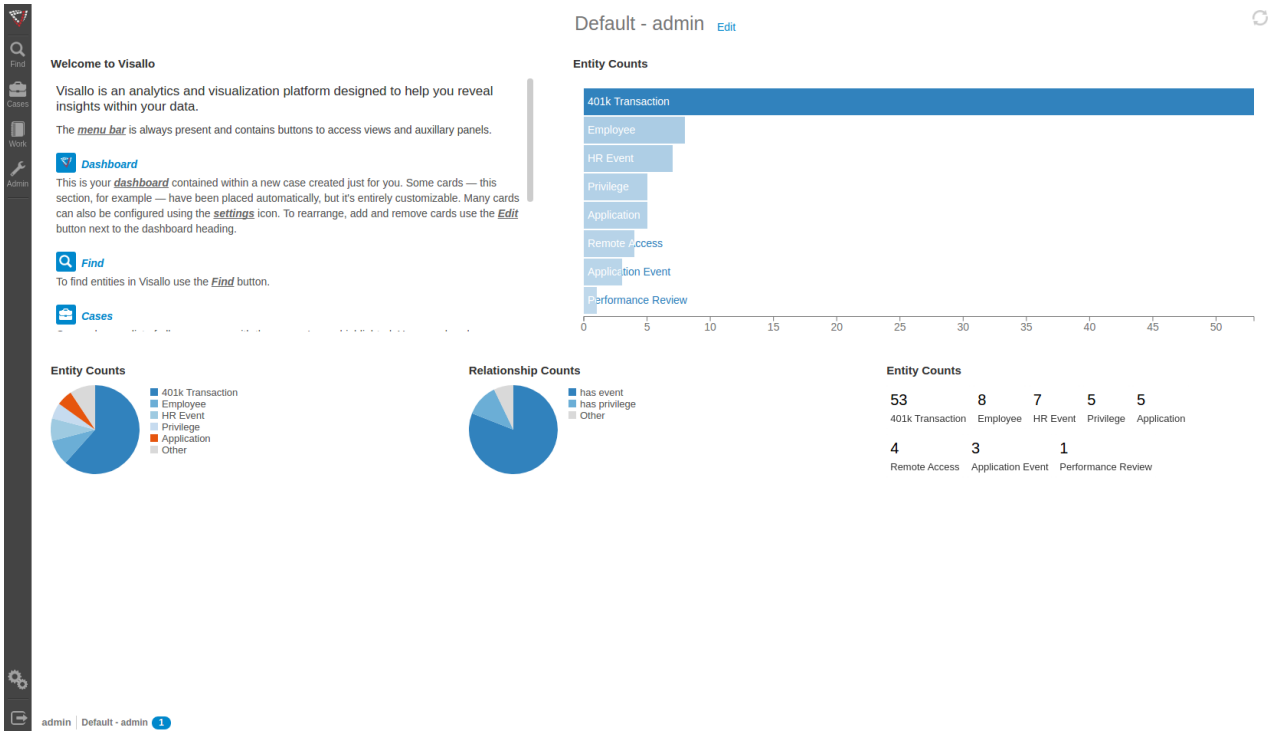

Figure 2.6: Example arrangement of the Visallo starting page (dashboard) after importing data about an insider threat scenario. [Screenshot created by the authors of this survey using Visallo]

The data files and ontologies used for this example were retrieved by contacting their customer support.

Dashboard

After logging in, the starting page of the web application which is realized as a basic dashboard is displayed. This customizable and rearrangeable dashboard can be extended using predefined or self-constructed widgets. It serves as an overview screen and presents a simple view on the currently accessible data. The dashboard and its widgets should be arranged in such a way that gives the analyst a basic idea about the dataset and possibly occurring inconsistencies. For example, when looking at figure 2.6, it is obvious that a suspiciously large amount of transactions were completed recently.

Workbench

While Visallo can also handle map visualization, analyzing graph structures is one of the more sophisticated approaches this tool can take. Therefore, such structures will be the focus of describing the main workbench features.

The graph structure visible in figure 2.7 represents an example analysis of the relations and properties of entity "Employee 0001". The informative section on the right side of the screenshot not only shows essential details about the entity, but also allows interaction such as querying for similar contents. The usage of these interactive components as well as the search window in general can be seen in figure 2.8. In this example, an entity that represents a failed remote access is used to search for other accesses attempted from the same IP address as existent in a property of this node. Furthermore, a timeline feature is included which can visualize the timestamp of any action or dated information. Figure 2.9 demonstrates the identification of events and entities according to a specific day or date range.

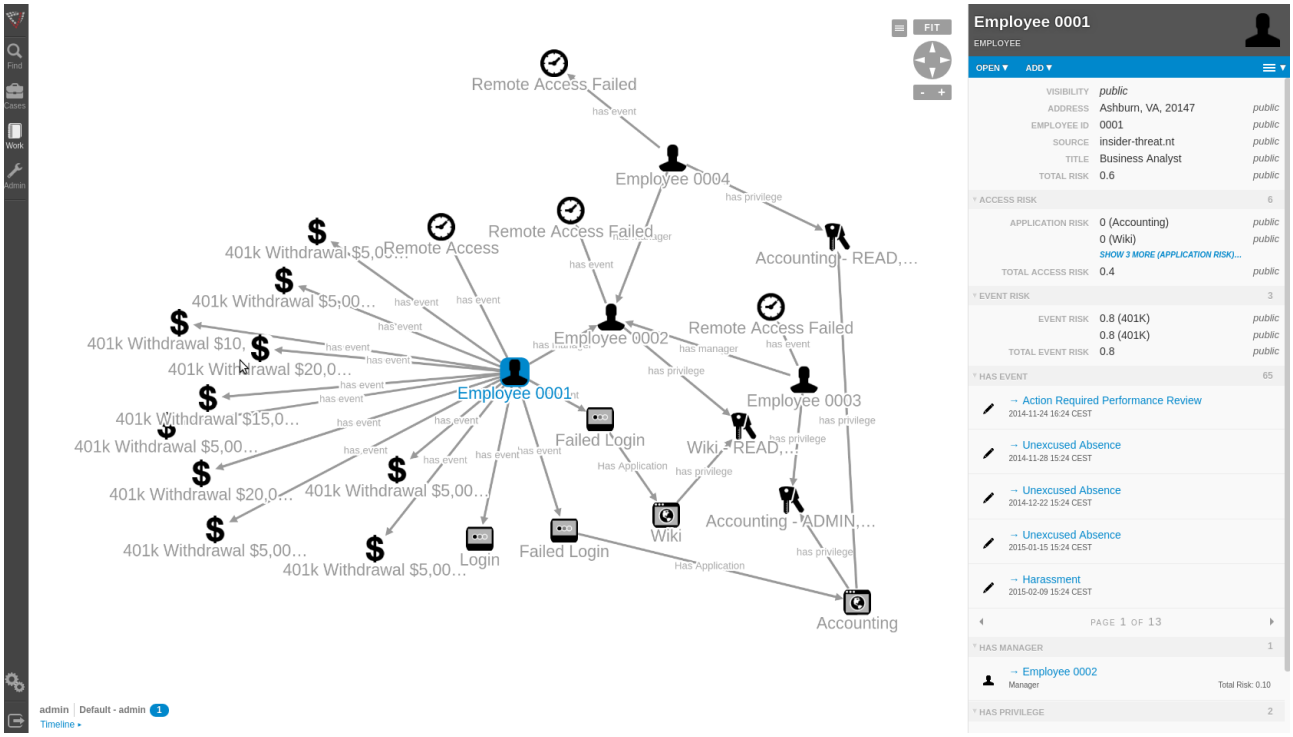

Figure 2.7: Visualization and analysis of the insider threat scenario using Visallo's workbench, with emphasis on relation and properties of the entity "Employee 0001". [Screenshot created by the authors of this survey using Visallo.]

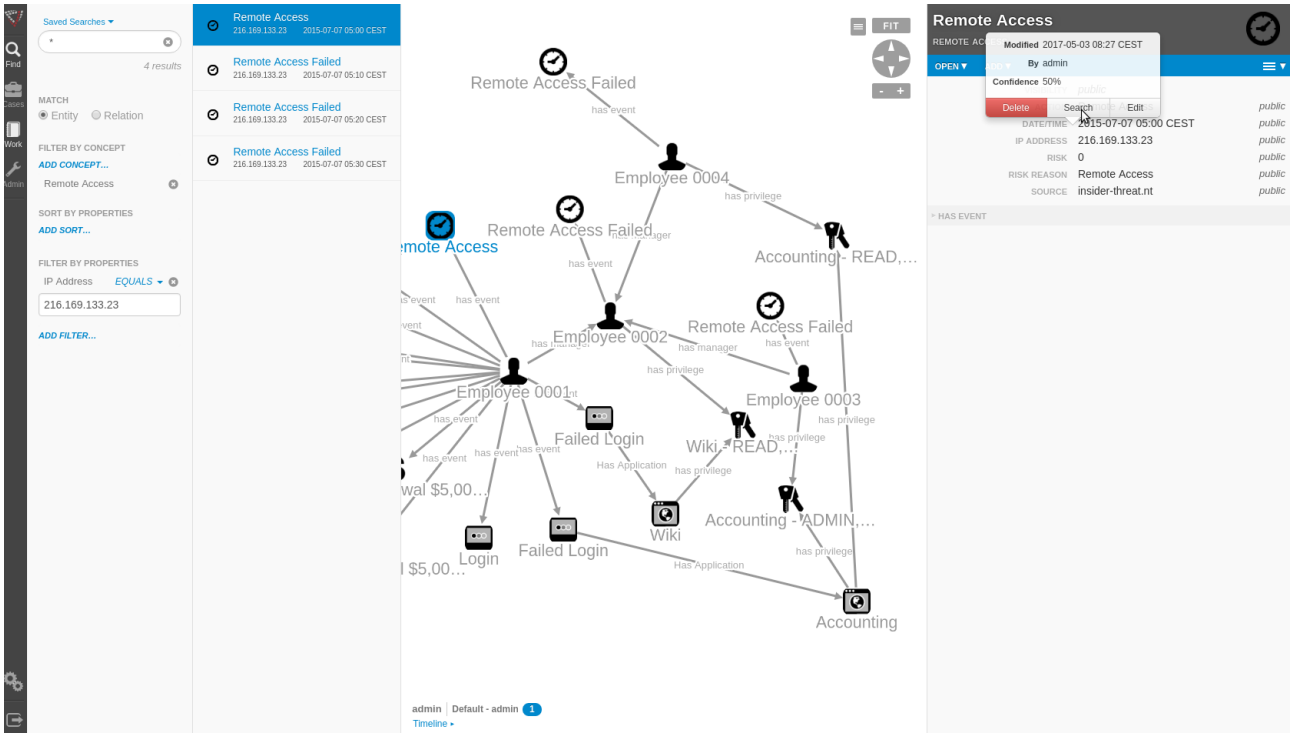

Figure 2.8: An example of executing a search query on the workbench. This specific query was posted by interaction of the user with an entity's properties. [Screenshot created by the authors of this survey using Visallo.]

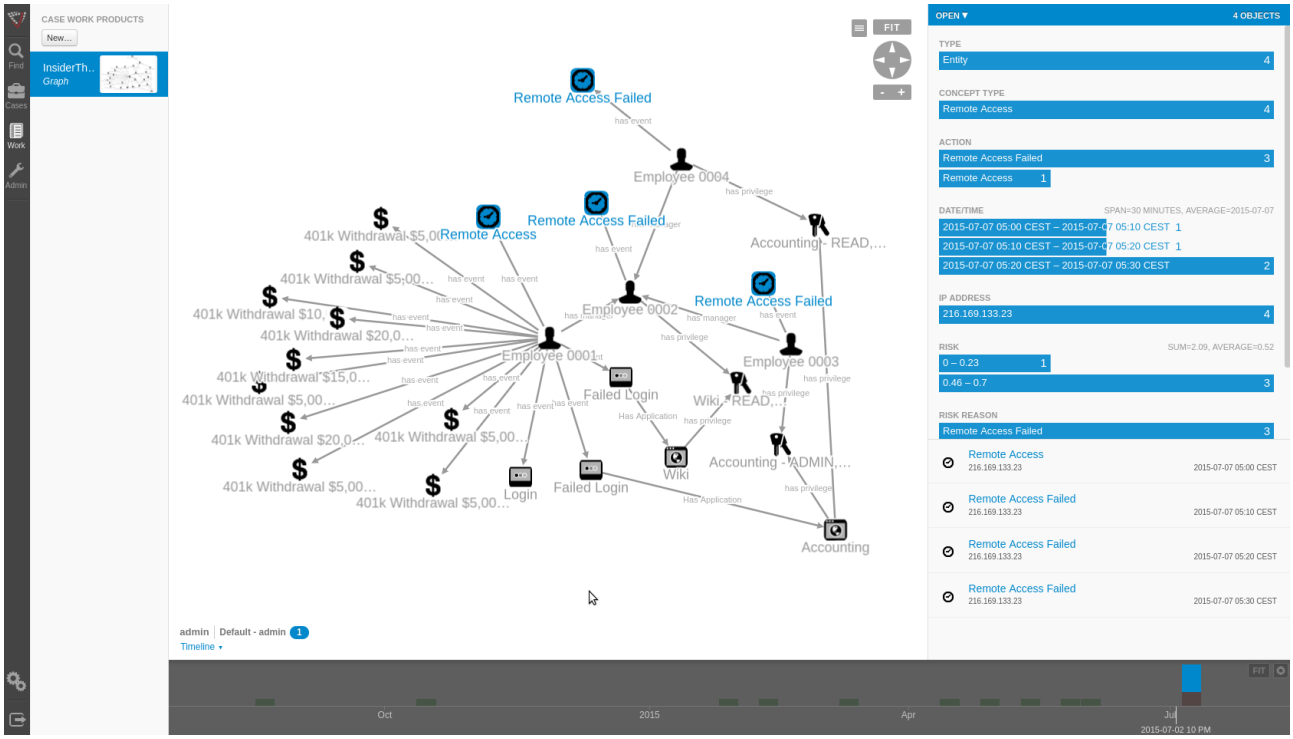

Figure 2.9: More detailed information about multiple entities according to the selection of a date range in the timeline. [Screenshot created by the authors of this survey using Visallo.]

2.3.4 Limitations

Although Visallo provides outstanding functionalities for already created ontologies as well as labeled entity and property data, it has problems handling raw data. Properly preparing an ontology for a given use case is also a quite demanding task.

Extraction of Entities and Properties from Text

In comparison to other tools such as Jigsaw or Nspace, Visallo's core functionality does not support entity text extraction. Although manual extraction of entities and properties from raw text resources and basic structured data is possible, this procedure is very time-consuming and not applicable in practice. The company behind Visallo seems to provide services and functionalities to provide assistance with this task. However, this is not officially disclosed on their website. This limitation was especially made noticeable by the VAST challenge 2006 dataset, because it mainly consists of structured tables as well as raw newspaper articles [SEMVAST, 2010b].

Creating and Maintaining Ontologies

In section 2.3.1 the concept of ontologies was already shortly explained. It has to be stressed that the creation of such complex semantic vocabularies requires huge effort. Although tools which can be used to create ontologies are freely available, the process of setting up an entirely new ontology for a specific use case is complex. Again, Visallo offers help for this problem in terms of services. However, this assistance has to also be requested specifically.

2.3.5 Subjective Assessment

Visallo has some downsides, especially regarding its portability from one use case to another due to its dependency on ontologies. These structures are difficult to generate and maintain. During the feature evaluation this lead to some problems, especially when working with the dataset of the VAST challenge 2006 [SEMVAST, 2010b]. Additionally, the official documentation seems to mainly focus on explaining the possibilities of extending the existing functionally in contrast to actually providing a manual on how to use the software [Visallo, 2017a]. A more detailed documentation would have been useful, especially when setting up the local installation and server needed to run the web application. This lack of information needlessly increases the effort for users which do not want to directly request support for the tool. On the contrary to these downsides, it has to be mentioned that Visallo's features for graph analysis and visualization are a great way of finding patterns in complex use cases, once properly set up. When thinking of using Visallo in a real environment, the aspect of it being open source might also be an advantage.

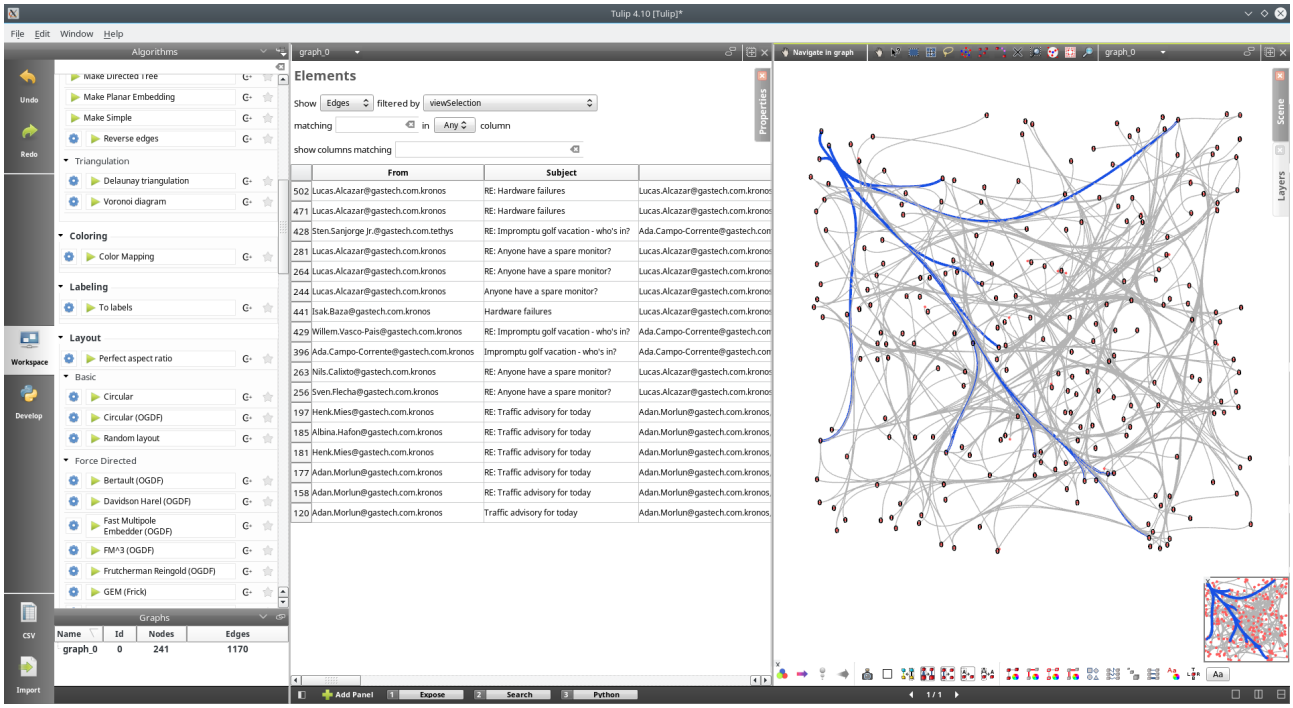

Figure 2.10: Visualization of relational, structured data included in the VAST challenge 2014 [SEMVAST, 2014]. This example focuses on senders, recipients and the subject of e-mails. Edge bundling and curving were performed via Tulip prior to taking this screenshot. [Screenshot created by the authors of this survey using Tulip.]

2.4 Tulip

This project was started by David Auber in 2004 and is still being maintained and actively developed [Auber, 2004]. Besides providing a tool which is mainly used to visualize and analyze graph structures, Tulip also offers a framework which enables the development of further functionality. Tulip is designed to visualize relational data as graphs in an appropriate way. When looking at the submissions of past VAST challenges, some participants used Tulip to visualize and analyze parts of some challenges.

2.4.1 Licensing

Tulip is licensed under GNU Lesser General Public License (LGPL) [Foundation, 2007]. In short, this means that it is allowed to copy, modify and distribute the product. Changes to the original software have to also be published under LGPL. When using existing code licensed under LGPL as a library, the application using this library does not necessarily have to implement LGPL.

2.4.2 Features

The most prominent functionality of Tulip is graph visualization and handling relational data associated to it. Importing structured data is possible in an effortless way. The tool can automatically extract entities and relations from a given table or Character Separated Values file (CSV). In addition to that, it also offers a palette of algorithms in the form of a toolbox, which can be applied to the currently processed graph. Figure 2.10 demonstrates the visualization of a graph as well as displaying data according to the current selection of nodes. Both edge bundling and edge curving algorithms were applied to the graph visible in this screenshot via Tulip.

2.4.3 Limitations

Although Tulip offers functionality which enables automatic extraction of entities, nodes, and edges from relational or structured data, it lacks text processing functionality. Therefore, it is not suited for investigative analysis as a standalone tool with its standard functionality alone. However, Tulip offers a framework API which provides the possibility of adding scripts and additional features.

2.4.4 Subjective Assessment

This tool looks very promising in context of graph visualization and analysis. It enables a quickly set up overview of a given use case represented in relational data or tables. When attempting to use the tool for evaluation, no difficulties regarding the installation or import of data occured. Although its functionalities might be too shallow for investigative analysis as a whole, it still provides an interesting alternative to more sophisticated and expensive tools for examining relational data and graphs.

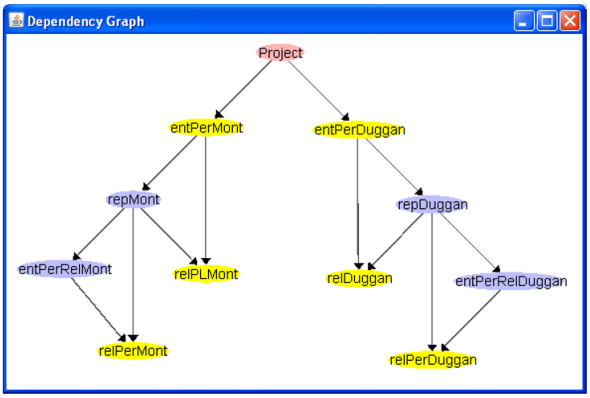

Figure 2.11: This CZSaw dependency graph shows multiple related views and data updates as created when starting with the original data (project node). It is immediately discernible that two different data-perspectives were chosen for the analysis. [Image extracted from Kadivar et al. [2009] under © 2009 IEEE. Used under the terms of Austrian copyright law: §42f]

2.5 CZSaw

CZSaw was developed by Kadivar et al. [2009], a team of students from Simon Fraser University which used Jigsaw as an inspiration. The motivation for developing a new tool was the lack of transparency of the analysis process in existing tools. Even though tools like Jigsaw can achieve outstanding results, the process of how those results were reached is lost. To tackle this problem CZSaw uses two novel features:

- A transactions script recording performed actions in a scripting language.
- A dependency graph modeling the analysis process and the dependencies of the performed steps and created views. Figure 2.11 shows a dependency graph linking multiple related views and data updates.

Three key ideas can be identified:

1. If the performed actions are recorded in a script, this script can later be modified to get more accurate results when an error is discovered later in the analysis process. Hence, a lot of time can be saved, resulting in higher efficiency and will to experiment.
2. Using scripts as the basis of the analysis process allows for automation and reuse of certain steps as well as a quick and powerful way to perform changes in the process.
3. The dependency graph actually visualizes the analysis process. Obvious flaws or alternative approaches may be discovered easily when looking at such a visualization.

The novelty of this approach helped the team around CZSaw to win an award for outstanding analysis and accuracy in the VAST 2010 Mini Challenge 1 which is about identifying key players in illegal arms dealing [SEMVAST, 2010a]. To show the versatility of CZSaw, the challenge was done by two separate teams in the beginning. They later on combined their results with little effort due to the recording of the analysis process [Chen et al., 2010a].

2.5.1 Licensing

CZSaw was freely available, but not open source. Some links to the website hosting the tool still exist, but the website is down and no further information can be acquired (not even from the authors of the tool).

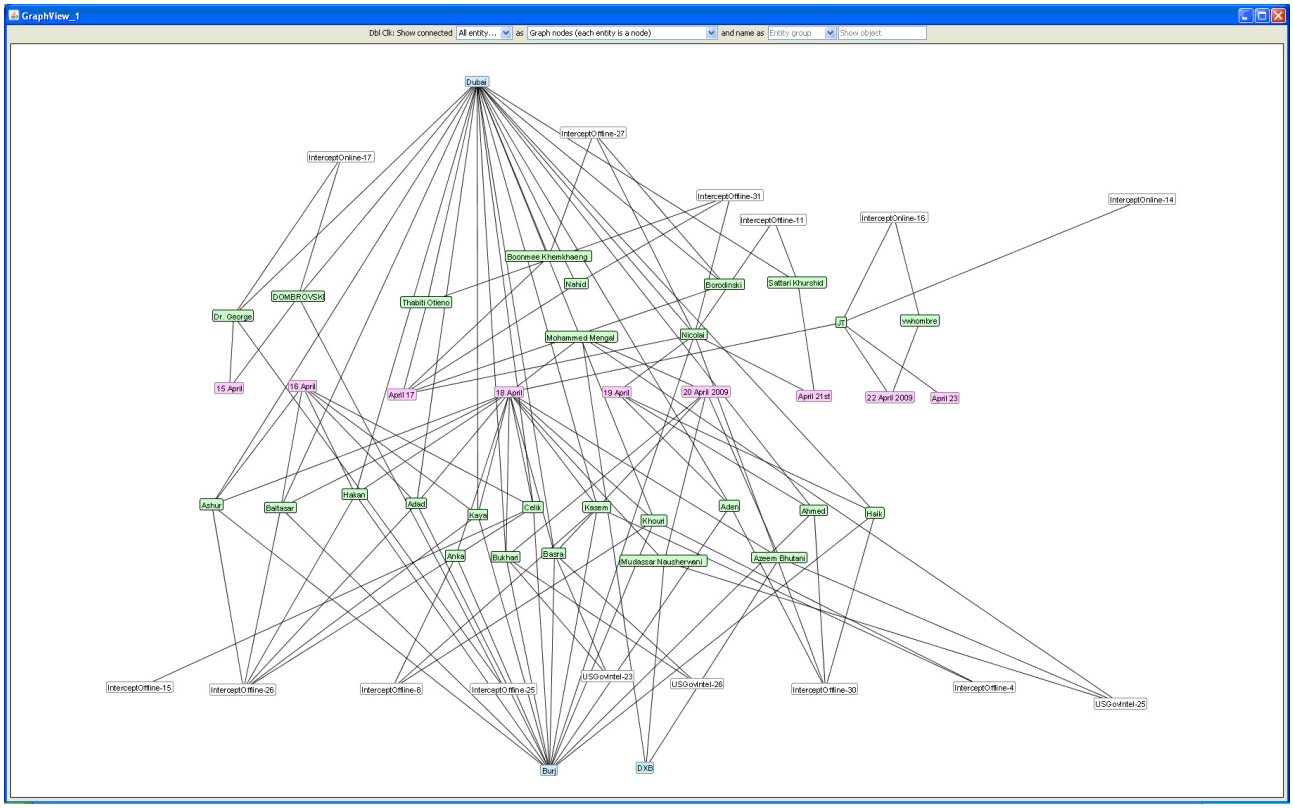

Figure 2.12: A CZSaw graph view created for the VAST 2010 Mini challenge 1 showing the social network of key players. [Image extracted from Chen et al. [2010b]. Copyright remains with original authors. Used under the terms of Austrian copyright law: §42f]

2.5.2 Features

In addition to the features inspired by Jigsaw, CZSaw offers some noteworthy features:

- A visual action history.
- A transaction script records all performed actions in a scripting language which can be used to program an analysis process as well.
- A dependency graph models all relevant views and transactions and creates a big picture of the work done within the tool (the analysis process).
- Multiple linked views allow for interconnected analysis of various dimensions.
- A highly dynamic graph view of documents (see figure 2.12) supporting automated and manual clustering and semantic zoom for clusters (see figure 2.13).

2.5.3 Limitations

The main limitation of CZSaw is the sole focus on text documents and document collections. As investigative analysis often deals with heterogeneous data, this can be crippling.
When working with text documents in the context of investigative analysis, entity extraction is a vital feature. However, CZSaw does not support automated entity extraction, but relies on external tools for it. This is especially frustrating when considering that most of the functionality of CZSaw strongly depends on entities.

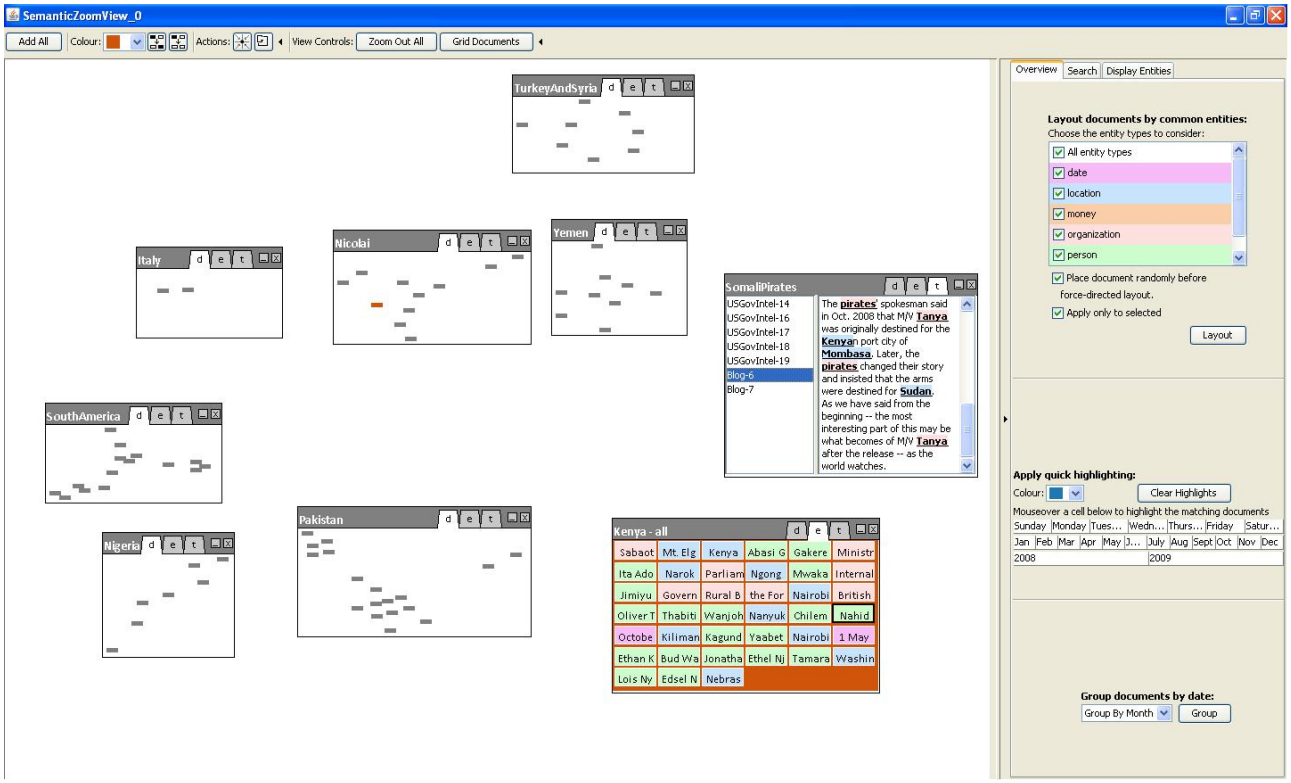

Figure 2.13: A CZSaw semantic zoom view created for the VAST 2010 Mini challenge 1 showing manually labeled clusters with different levels of semantic zoom. [Image extracted from Chen et al. [2010b]. Copyright remains with original authors. Used under the terms of Austrian copyright law: §42f]

2.5.4 Subjective Assessment

It is a shame that CZSaw is not available anymore as it seems to be a well thought-through tool. Its basic features are inspired by Jigsaw, which is known to be a powerful tool when it comes to textual analysis. However, the feature that could very well elevate it above Jigsaw is the focus on the analysis process.

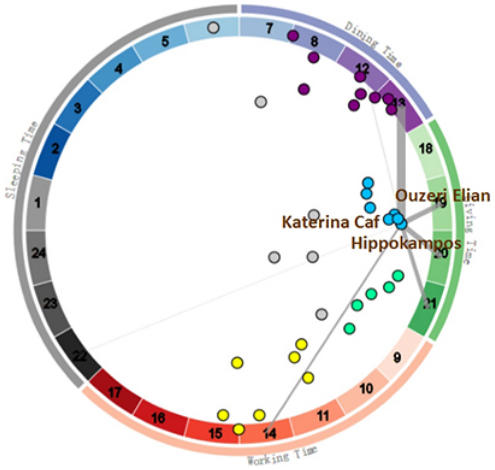

Figure 2.14: RadViz is used to categorize the unknown location called "Hippokampos" as a restaurant due to its similar neighbors which are known to be restaurants. [Image extracted from Zhao et al. [2014] under © 2014 IEEE. Used under the terms of Austrian copyright law: §42f]

2.6 RadViz, PMViz, SGGViz

In 2014 the team around Zhao et al. [2014] from Central South University, Tianjin University, and University of Texas at Dallas collaborated to participate in the VAST 2014 challenge [SEMVAST, 2014]. The result are three intertwined tools: PMViz, RadViz, and SGGViz. Due to the nature of the VAST 2014 Mini challenge 2 [SEMVAST, 2014], the three tools focus mainly on trajectory data (for example as provided by GPS tracking), transactions (for example credit card transactions) and plain text.

The tools were developed as web applications using mainly D3, a JavaScript library for visualization developed by Bostock et al. [2011].

2.6.1 Licensing

Only a few traces of the tools can be found on the web. All of them lead to dead ends. One can only assume that the three tools were either never published or taken offline soon after publication. Therefore, no statement about possible licenses or whether they are open source can be made.

2.6.2 Components

The three components are meant to be used in a pipeline and will be described in more detail in the following sections. RadViz is used for preprocessing, PMViz for filtering and statistical analysis, and SGGViz for finally connecting all relevant dimensions to build hypotheses for spatio-temporal characteristics.

RadViz

RadViz was used for preprocessing in the VAST 2014 Mini challenge 2. The main purpose of the tool is to perform clustering and matrix reordering for inconsistent datasets for which they use techniques as proposed by Daniels et al. [2012]. Through clustering they can discover employees with similar habits and discover locations not present or named consistently in the dataset. Figure 2.14 demonstrates the categorization of a location called Hippokampos which is likely a restaurant (due to its similar known neighbors).

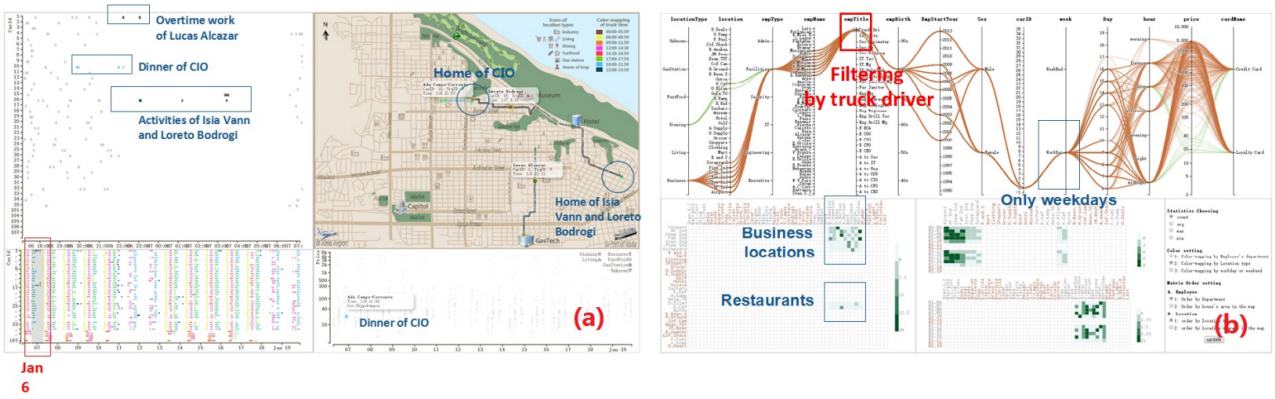

Figure 2.15: (a) shows the SGGViz tool as used to detect anomalies on January 6^{th}. (b) shows the PMViz tool as used to filter out transactions of truck drivers. [Image extracted from Zhao et al. [2014] under © 2014 IEEE. Used under the terms of Austrian copyright law: §42f]

PMViz

The PMViz tool is used for working with transaction data and uses a dynamic parallel coordinates view as its main visualization technique. Additionally, some matrix views are used. This visualization in combination with statistical functions helps to identify interesting data points. An example can be seen in figure 2.15(b).

SGGViz

SGGViz consists of multiple views (including scatter views and a map) to analyze the connection between the different data types: temporal, geographical and transaction data. The different types are combined to allow the user to follow transactions and people in time and space. Figure 2.15(a) shows the discovery of anomalies in the data using SGGViz.

2.6.3 Subjective Assessment

Due to the limited resources available for RadViz, PMViz and SGGViz no sound assessment can be made. However, it can be concluded that the tools are used for a rather limited scope of problems. It would be nice to see the (fairly good looking) tools applied to other problems to draw more meaningful conclusions.

2.7 Analyst's Notebook

Analyst's Notebook is a tool which was introduced by IBM [Randjelović and Popović, 2011]. This tool is mostly used by the United States Army program for fraud and terrorism attack detection. It is not possible to test the software, since it is not available. From the pictures and publicly available videos, we can assume this tool uses onthologies and entity detection to visualize the complex relationship between various documents. Since the tool is mostly used by the United States army, it also supports geospatial and multi-dimensional data sets.

2.7.1 Licensing

The tool is not available to the public. We assume it is more business-related and suitable for governments rather personal and normal users.

2.7.2 Limitations

From the documents on the website, Analyst's Notebook seems like a quite powerful tool for any type of investigation analysis. However, since the tool is not available it is hard to judge what kind of limitations the tool exactly has.

2.7.3 Subjective Assessment

Analyst's Notebook is a powerful tool to load data for investigative analysis and visualize it. It supports various views and also various data types like temporal, geospatial, ontologies, and text for event extraction, fraud detection, and social network analysis. Although it supports many formats and different types of data representation, it is hard to further investigate this tool due to its unavailability.

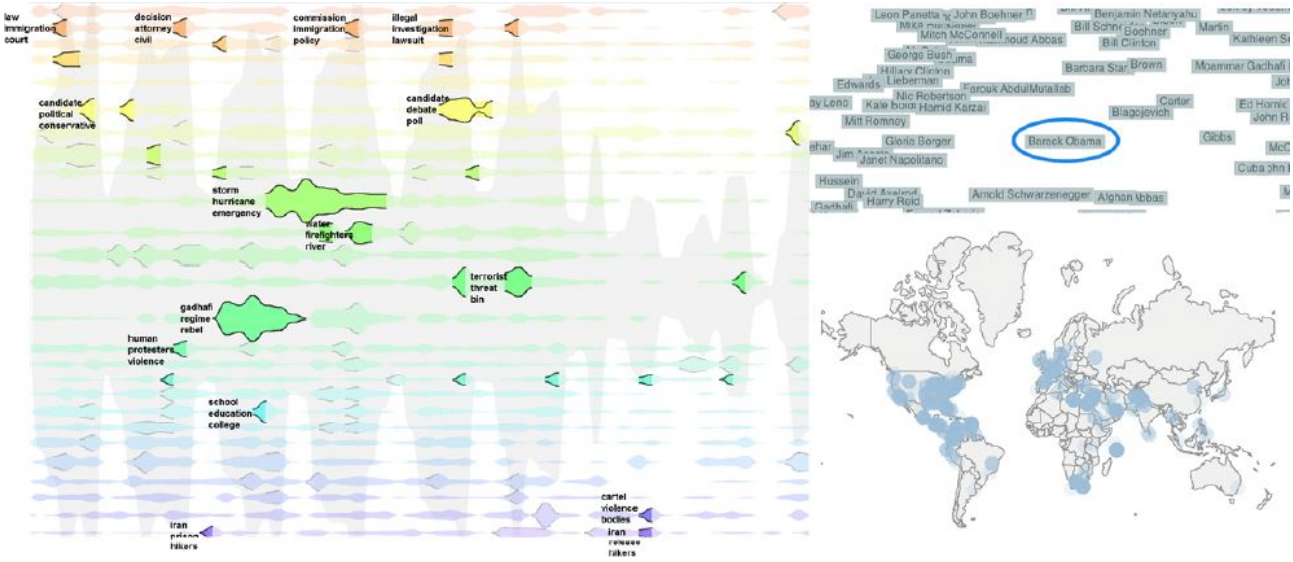

Figure 2.16: Overview of Leadline. In the top right of the figure president Obama is selected. On the left side, entities related to president Obama are shown. In the lower right corner of the figure the locations related to him are marked. [Image extracted from Dou et al. [2012]. © LeadLine project 2012. Used under the terms of Austrian copyright law: §42f]

2.8 LeadLine

LeadLine is a tool for visualizing multiple text documents in various views [Dou et al., 2012]. This tool basically uses text documents as an input, then extracts entities like locations, persons, and organizations. It also extracts events that are related to them. Later, these entities will be shown in different views. The user can select an entity or event in one of the views and then see the relation of the chosen event to other events or entities in the system. For example, figure 2.16 displays the addition of a news feed to the system. President Obama was chosen as an input for whom the system shows various locations and also different events related to this person.

2.8.1 Licensing

LeadLine was developed in the course of research for a paper in 2012 [Dou et al., 2012]. The authors of the paper did not release the software for other developers or users to test. Since there is no further information about this project on the author's web page, it can be assumed that the project is discontinued.

2.8.2 Features

The system consists of three major parts: A timeline, a world map, and a text cloud. After an user adds text files to the system, the engine automatically extracts events and other entities. Afterwards, a text cloud composed of entities will be shown on the screen. It is possible to select each of the events to visualize them in a timeline while also displaying the world map and correlation of the events.

2.8.3 Limitations

The tool is limited to analysis and visualization of text data. Also, some of the views used to visualize data lack further functionality. Although entity extraction seems to be the most important feature of this tool, the way the tool attempts visualization of data could be improved. Due to its unavailability, it is hard to make further assumptions.

2.8.4 Subjective Assessment

LeadLine is used to extract events from text data, which is the main contribution of this tool to the published paper. It also nicely links the events with entities. Since the tool is not publicly available, it is hard to evaluate it in a more detailed way.

2.9 Palantir

While participating in VAST contests in 2008, 2009, and 2010, Palantir as a company has not made the used tool available to the public [SEMVAST, 2017]. Nowadays, Palantir focuses on providing services for customers in need of investigative analysis as well as two separate tools called "Palantir Gotham" and "Palantir Metropolis" [Palantir, 2017]. The description of the different aspects such as features and limitations will be mainly conducted in reference to the version of the tool used in the VAST challenges. Palantir's more recent software platforms are heavily based on those older versions and also have the same main functionalities.

2.9.1 Licensing

The services and software offered by Palantir are not freely accessible. Pricing for these products is also not officially available and has to be inquired via direct contact.

2.9.2 Features

When looking at one of the older versions of Palantir's tool it is obvious that the main focus of this platform lies on visualizing data using graphs and maps [Payne, Solomon, Sankar and McGrew, 2008]. However, importing data from multiple sources seems to be supported. In addition to that, information can be filtered using a timeline. These two aspects are displayed in figure 2.17. This tool also introduces dynamic display of information flows in given graphs and maps. Displaying such flows can be helpful to gain an overview of the strength and intensity of given relations. When talking about preprocessing and importing data, Palantir seems to also offer simple text entity extraction mechanisms as well as management of basic ontologies. Furthermore, the company claims to provide extension points which enable the possibility of customization.

2.9.3 Limitations

When evaluating the given resources and promotional videos, there were no apparent downsides identified. This lack of documented limitations obviously stems from the fact that it was not possible to perform an evaluation using the software platform in practical.

2.9.4 Subjective Assessment

Unfortunately, Palantir offers no possibility of testing their tools or services. The company however provides some video demos on its official Youtube channel, for example about "Palantir Gotham", which is one of its newer platforms [Palantir, 2014]. From the looks of both older and newer platforms, the tools Palantir offers seem like well-developed software for investigative analysis. Especially the feature of displaying information flows in a graph or map represents a feature which most other tools do not provide.

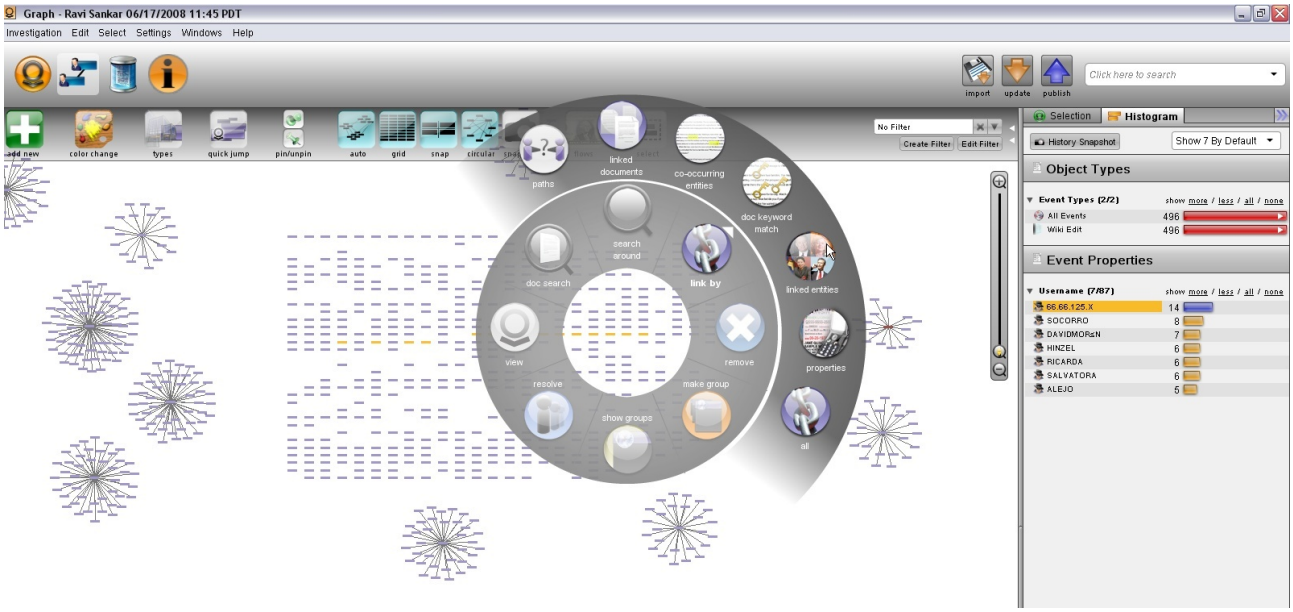

Figure 2.17: The main screen for visualizing graphs in Palantir. An UI-element used to interact with the given nodes can also be seen. [Image extracted from Payne, Solomon and Sankar [2008] under © 2008 Palantir Technologies, Inc. Used under the terms of Austrian copyright law: §42f]

2.10 Analyst's Workspace

Analyst's Workspace (AW) is an analytics tool which is primarily used on large-scale screens with high resolution. It has been developed by Virginia Tech and tries to mimic an analogous workspace where it is possible to freely drag documents around as desired. Enormous screen sizes also enhance the users' capability "to rapidly access information through physical navigation (eye, head, and body movement), while maintaining a strong sense of the layout of the space", as C. Andrews, Hossain et al. [2011] state.

The main goal of the tool is to create a working environment where "foraging and synthesis activities can be integrated into a single, fluid investigative process." AW mainly focuses on documents where the complete text is directly visible for the user. The already mentioned feature to drag elements around provides manual clustering of those in existence in order to manually gain an overview [C. Andrews and North, 2012], [C. Andrews, Hossain et al., 2011].

2.10.1 Availability

Trying to gather information on whether AW is available for personal use or not turned out to be more difficult than beforehand expected. There is no official website for the tool and the few detectable papers which deal as an information source for this report did not mention this topic at all. However, the non-existing online presence indicates that the tool is supposedly unavailable. Therefore the question of licensing is obsolete.

2.10.2 Features

AW provides an useful range of interaction tools, such as familiar click-and-drag, selecting rectangles, and multi-click selections. Moreover, information organization facilities, such as a graph layout and temporal ordering are included. Due to all these operations being local, this functions as a kind of sandbox for the user [Hossain et al., 2011].

An user's task to explore data is facilitated by showing direct visual connections between certain objects in all opened documents on the large screen. This interlinking is represented by coloured underlining of homogeneous elements, as illustrated in figure 2.18.

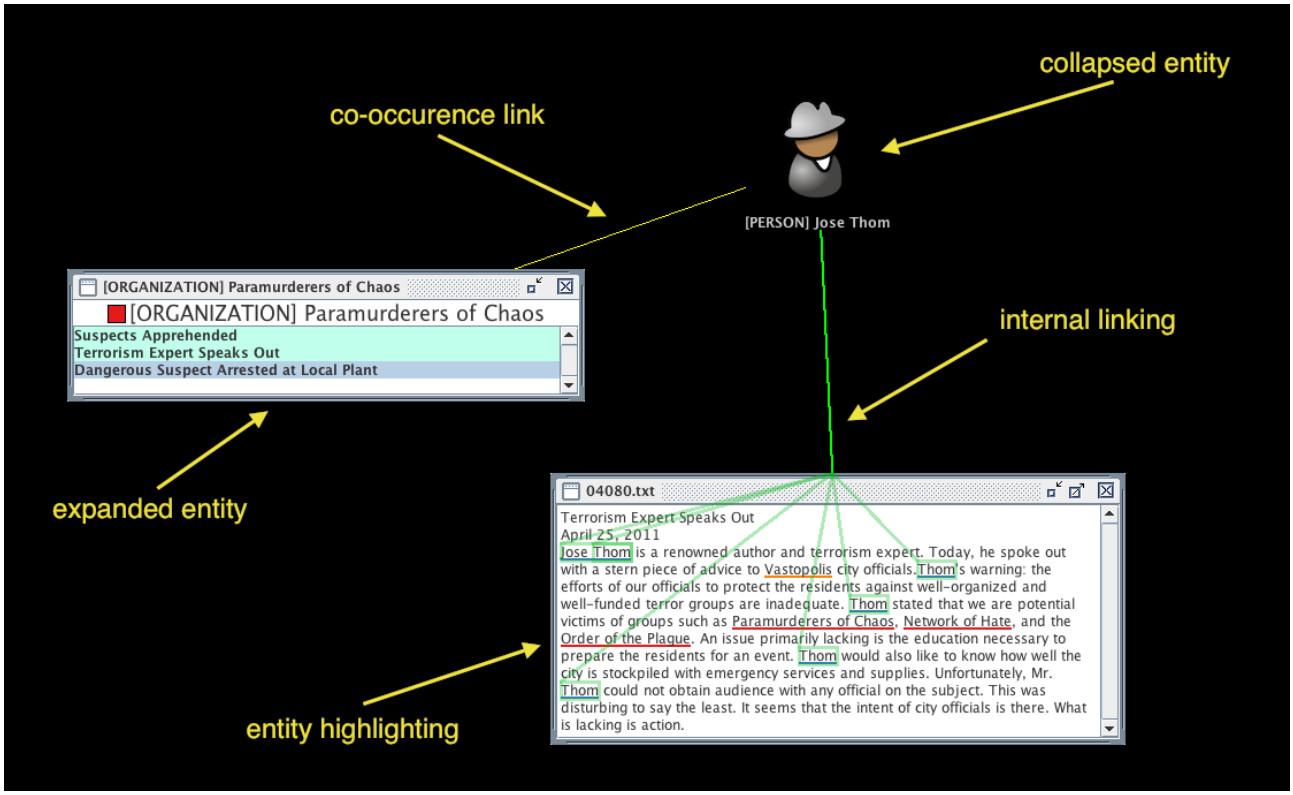

Figure 2.18: Demonstration of interlinked connections between entities in Analyst's Workspace (AW). [Image extracted from C. Andrews, Hossain et al. [2011]. © 2017 Virginia Tech. Used under the terms of Austrian copyright law: §42f]

Not only document-entity exploration, but also entity-entity and document-document exploration are possible in AW through graph-traversing. Material for finding links between these mentioned elements are provided by AW when drawing a connection within the tool. For example, similarity search of different documents can be conducted by using a neighbourhood tool which is based on a vector-space model [C. Andrews, Hossain et al., 2011].

2.10.3 Limitations

Entities may be sorted by frequency within AW, but it may be a difficult task to find a suitable entry point for the exploration. Moreover, the tool does not automatically supply the user with an overview of the data set. Another issue which arises is the lack of a handy extraction method after having worked with AW. There appears to be no way of transforming the final workspace into presentable form. [C. Andrews and North, 2012]

Due to a lack of current information, it cannot be well and truly said whether the limitations will ever be tackled or the development of the tool will be or has already been ceased.

2.10.4 AW and VAST

Virginia Tech has successfully participated with its tool in the year 2011. They were able to receive a "Novel Use of Large Screen Workspace to Support Analysis" award for solving the third mini challenge.

2.10.5 Subjective Assessment

Using multiple large screens as a basis for an investigative analysis environment is an interesting concept. However, not much can be said about AW as a whole due to its unavailability.

Chapter 3

Tool Comparison

To provide a quick and useful overview, the tools will be compared in this chapter. As the amount of resources available for each tool varies greatly and not every tool could be tested, the basis for comparison is not the best. Another problem is presented by the different focuses of the tools. They are each built for particular purposes and therefore not always strictly comparable.
Nevertheless, table 3.1 shows some criteria useful for comparing the tools. Note that not all the criteria can be answered objectively.

Based on our research and experience with the tools we can identify some personal favorites we would like to mention.
nSpace appears to be the most complete tool and has a lot to offer. Being able to work collaboratively with multiple people is a huge bonus for any investigation. The only real downside is the license fee.
When it comes to pure textual analysis Jigsaw is the best tool. It is easy to use and its power is featured by the many times it was used in VAST challenges throughout the years. However, the focus on text documents and document collections makes it fairly limited.
Another tool which is worth mentioning is CZSaw for its focus on the analysis process, a notion we only found in this tool. Incorporating recording and scripts into the analysis process is a valuable idea which can easily boost productivity and creativity in an investigation.
If a company would be looking for a tool which is also partly available as open source, Visallo might be an interesting alternative. However, its strong focus on ontologies could be a negative factor for possible users.

Name	nSpace	JigSaw	Visallo	Tulip	CZSaw	PMViz	Palantir	Leadline	Analyst's workspace	Analyst's notebook
Focus	Completeness	Text	Visual	Graphs	Analysis process	Geo + temporal	Completeness	Events	Documents	Geo and temporal
Availability	Fee	Free	Free	Free	✗	✗	Fee	✗	✗	✗
Open Source	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
Platform	Web	Multi/Web	Web	Multi	Multi	Web	?	Multi	?	Windows
Data structure	Anything	Text	Ontology	Relational	Text	?	Anything	Text	Text	Anything
Ease of use	?	Easy	Hard	Easy	?	?	?	?	?	?
Entity extraction	✓	✓	Manual	✗	✗	✓	✓	✓	✗	✓
Extensibility	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗

Table 3.1: Tool Comparison with Some Selected Criteria. A Question Mark Depicts an Unknown Fact.

Chapter 4

Conclusion

This survey provided an overview of the most prominent tools for visual investigative analysis used throughout the last years. While the VAST benchmark repository [SEMVAST, 2017] offers a great deal of information and promotes the most promising tools through various awards, it also makes the lack of well supported tools apparent.
In the early years of the VAST challenges - starting from 2006 - nSpace set itself apart from other tools as it offered the most complete and sophisticated approach. Another recurring winner was Jigsaw, which was introduced in 2007. Since then, generic award-winning tools were scarce as solutions became strongly tailored to specific tasks and no tool could emerge from the contenders like nSpace and Jigsaw did.
From this, two possibilities can be derived: Either nSpace and Jigsaw already perfected the task, or the developers of high quality tools are not interested in participating in the VAST challenges. Since the first statement seems highly unlikely, the latter seems to be true. And indeed, applications like IBM's Analyst's Notebook are powerful (but expensive) tools which do not partake in any VAST challenges.
In addition to the lack of supported (and affordable) tools another development can be identified which is best showcased by Palantir. While Palantir was a single tool once, recent information suggests that they have moved away from selling the tool but are now selling the service of performing analyses instead in addition to offering to distinct tools for very specific purposes. Visallo takes a hybrid-approach: their tool is free and open source, but their financial model seems to rely on providing supporting services for the analysis process.

It can be concluded that some big players like nSpace and IBM have established their tools very well, while recent participants of the VAST challenges have mainly academic interests and offer specialized, tailored solutions which cannot be applied to a wide range of problems and therefore cannot contend with already existing alternatives. In the spirit of progress, we can only hope that a free and powerful tool will emerge to rekindle some of the competition in the affordable market.

Bibliography

Andrews, Christopher, M. Shahriar Hossain, Samah Gad, Naren Ramakrishnan and Chris North [2011]. *Virginia Tech — Analyst's Workspace. VAST 2011 Challenge: Mini-Challenge 3 — Investigation into Terrorist Activity*. VAST challenge 2011. 2011. http://www.cs.umd.edu/hcil/VASTchallenge2011/entries/148-VT-AnalystsWorkspace-MC3/ (cited on pages 28–29).

Andrews, Christopher and Chris North [2012]. "Analyst's Workspace: An Embodied Sensemaking Environment for Large, High-Resolution Displays". In: *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. Oct 2012, pages 123–131. doi:10.1109/VAST.2012.6400559. http://ieeexplore.ieee.org/document/6400559/ (cited on pages 28–29).

Andrews, Keith [2012]. *Writing a Thesis: Guidelines for Writing a Master's Thesis in Computer Science*. Graz University of Technology, Austria. 22nd Oct 2012. http://ftp.iicm.edu/pub/keith/thesis/ (cited on page iii).

Andrews, Keith [2017]. *Information Visualisation. Course Notes*. Keith Andrews. 24th Mar 2017. http://courses.iicm.tugraz.at/ivis/ivis.pdf (cited on page 1).

Apache Software Foundation [2004]. *Apache License, Version 2.0*. Apache. Jan 2004. apache.org/licenses/LICENSE-2.0 (cited on page 12).

Auber, David [2004]. "Tulip — A Huge Graph Visualization Framework". In: *Graph Drawing Software*. Edited by Michael Jünger and Petra Mutzel. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pages 105–126. ISBN 978-3-642-18638-7. doi:10.1007/978-3-642-18638-7_5. http://dx.doi.org/10.1007/978-3-642-18638-7_5 (cited on page 17).

Bostock, Michael, Vadim Ogievetsky and Jeffrey Heer [2011]. "D3: Data-Driven Documents". *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* [2011]. http://vis.stanford.edu/papers/d3 (cited on page 22).

Chen, Victor, Dustin Dunsmuir, Saba Alimadadi, Eric Lee, Jeffrey Guenther, John Dill, Cheryl Qian, Chris D Shaw, Maureen Stone and Robert Woodbury [2010a]. "Model Based Interactive Analysis of Interwoven, Imprecise Narratives: VAST 2010 Mini Challenge 1 Award: Outstanding Interaction Model". In: *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE. 2010, pages 275–276. doi:10.1109/VAST.2010.5653060. http://ieeexplore.ieee.org/document/5653060/ (cited on page 19).

Chen, Victor, Dustin Dunsmuir, Saba Alimadadi, Eric Lee, Jeffrey Guenther, John Dill, Cheryl Qian, Chris D Shaw, Maureen Stone and Robert Woodbury [2010b]. *VAST 2010 Mini Challenge 1 Results from Simon Fraser University*. VAST challenge 2010. 2010. http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202010/challenges/MC1%20-%20Investigations%20into%20Arms%20Dealing/entries/Simon%20Fraser%20University/ (cited on pages 20–21).

Chien, Lynn, Annie Tat, Pascale Proulx, Adeel Khamisa and William Wright [2008]. "Grand Challenge Award 2008: Support for Diverse Analytic Techniques-nSpace2 and Geotime Visual Analytics". In: *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. IEEE. 2008, pages 199–200. doi:10.1109/VAST.2008.4677385. http://ieeexplore.ieee.org/document/4677385/ (cited on page 4).

Daniels, Karen, Georges Grinstein, Adam Russell and Mason Glidden [2012]. "Properties of Normalized Radial Visualizations". *Information Visualization* 11.4 [2012], pages 273–300. doi:10.1177/1473871612439357. http://journals.sagepub.com/doi/pdf/10.1177/1473871612439357 (cited on page 22).

Dou, Wenwen, Xiaoyu Wang, Drew Skau, William Ribarsky and Michelle X Zhou [2012]. "Leadline: Interactive Visual Analysis of Text Data through Event Identification and Exploration". In: *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE. 2012, pages 93–102. doi:10.1109/VAST.2012.6400485. http://ieeexplore.ieee.org/document/6400485/ (cited on page 25).

Foundation, Free Software [2007]. *GNU Lesser General Public License v3*. Free Software Foundation, GNU. 29th Jun 2007. https://gnu.org/copyleft/lesser (cited on page 17).

Gorg, Carsten, Zhicheng Liu, Neel Parekh, Kanupriyah Singhal and John T Stasko [2007]. "Jigsaw meets Blue Iguanodon-The VAST 2007 Contest." *IEEE VAST* 7 [2007], pages 235–236. doi:10.1109/VAST.2007.4389034. http://ieeexplore.ieee.org/document/4389034/ (cited on page 9).

Grinstein, Georges, Theresa O'Connell, Sharon Laskowski, Catherine Plaisant, Jean Scholtz and Mark Whiting [2006]. "VAST 2006 Contest - A Tale of Alderwood". In: *2006 IEEE Symposium On Visual Analytics Science And Technology*. Oct 2006, pages 215–216. doi:10.1109/VAST.2006.261420. http://ieeexplore.ieee.org/document/4035768/ (cited on page 2).

Haack, Jereme, Carrie Varley, Mark Whiting and Katie Wolf [2006]. *IEEE VAST 2006 Contest. Dataset and Tasks*. VAST repository. 2006. http://www.cs.umd.edu/hcil/VASTcontest06/dataset.htm (cited on page 2).

Hossain, Mahmud Shahriar, Christopher Andrews, Naren Ramakrishnan and Chris North [2011]. "Helping intelligence analysts make connections". In: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. 2011. http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/download/3937/4323 (cited on page 28).

IEEE VAST symposium [2006]. *IEEE Symposium on Visual Analytics Science and Technology 2006*. IEEE. 2006. http://conferences.computer.org/vast/vast2006/ (cited on page 1).

Jonker, David, William Wright, David Schroh, Pascale Proulx, Brian Cort et al. [2005]. "Information Triage with TRIST". In: *2005 International Conference on Intelligence Analysis*. 2005, pages 2–4. http://hcil2.cs.umd.edu/newvarepository/InfoVis%20Contest%202007/challenges/InfoVis%20Goes%20to%20the%20Movies/entries/OculusInfo-nSpaceAndGeoTime/Index_files/ref%20papers/Oculus_TRIST_Final_Distrib.pdf (cited on pages 4–5).

Kadivar, Nazanin, Victor Chen, Dustin Dunsmuir, Eric Lee, Cheryl Qian, John Dill, Christopher Shaw and Robert Woodbury [2009]. "Capturing and Supporting the Analysis Process". In: *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE. 2009, pages 131–138. doi:10.1109/VAST.2009.5333020. http://ieeexplore.ieee.org/document/5333020/ (cited on page 19).

Kang, Youn-ah, Carsten Gorg and John Stasko [2011]. "How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation". *IEEE Transactions on Visualization and Computer Graphics* 17.5 [May 2011], pages 570–583. ISSN 1077-2626. doi:10.1109/TVCG.2010.84. http://ieeexplore.ieee.org/abstract/document/5482577/ (cited on page 1).

Palantir [2014]. *Investigating the Illicit Ivory Trade with Palantir Gotham*. Youtube. 20th Nov 2014. https://youtu.be/yMv3TBxulu4 (cited on page 27).

Palantir [2017]. *Palantir products*. Palantir. 14th May 2017. https://palantir.com/products/ (cited on page 27).

Payne, Jason, Jake Solomon and Ravi Sankar [2008]. *Palantir VAST 2008 Challenge*. VAST benchmark repository. 2008. http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202008/challenges/Grand%20Challenge%202008/entries/Palantir%20Technologies/ (cited on page 28).

Payne, Jason, Jake Solomon, Ravi Sankar and Bob McGrew [2008]. "Grand Challenge Award: Interactive Visual Analytics – Palantir: The Future of Analysis". In: *Proc. IEEE Symposium on Visual Analytics Science and Technology*. IEEE Computer Society. Oct 2008, pages 1–2. doi:10.1109/VAST.2008.4677386. http://ieeexplore.ieee.org/document/4677386/ (cited on page 27).

Proulx, Pascale and Casey Canfield [2015]. "The beneficial role of the VAST Challenges in the evolution of GeoTime and nSpace2". *Information Visualization* 14.1 [2015], pages 3–9. doi:10.1177/1473871613487090. http://journals.sagepub.com/doi/pdf/10.1177/1473871613487090 (cited on page 4).

Randjelović, Dragan and Brankica Popović [2011]. "Visual analytics tools and theirs application in social networks analysis". In: *Telecommunications Forum (TELFOR), 2011 19th*. IEEE. 2011, pages 1340–1343. doi:10.1109/TELFOR.2011.6143801. http://ieeexplore.ieee.org/document/6143801/ (cited on page 24).

SEMVAST [2010a]. *VAST 2010 Mini Challenge 1*. VAST benchmark repository. 2010. http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202010/challenges/MC1%20-%20Investigations%20into%20Arms%20Dealing/ (cited on page 19).

SEMVAST [2010b]. *VAST Challenge 2006*. VAST benchmark repository. 2010. http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202006/challenges/2006%20Contest/ (cited on pages 15–16).

SEMVAST [2014]. *VAST 2014 Mini Challenge 2*. VAST benchmark repository. 2014. http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202014/challenges/MC2%20-%20Patterns%20of%20Life%20Analysis/ (cited on pages 17, 22).

SEMVAST [2017]. *Visual Analytics Benchmark Repository*. VAST benchmark repository. 11th May 2017. http://hcil2.cs.umd.edu/newvarepository/benchmarks.php (cited on pages 2, 4, 8, 27, 33).

Uncharted Software [2017]. *nSpace Overview*. Uncharted Software. 11th May 2017. https://www-prev.uncharted.software/nspace/ (cited on pages 6–7).

Visallo [2015]. *Insider Threat Demo*. Youtube. 9th Aug 2015. https://youtu.be/Vov0_BrIH6E (cited on page 12).

Visallo [2017a]. *Visallo Developer Documentation*. Website. 14th May 2017. http://docs.visallo.org/ (cited on pages 12, 16).

Visallo [2017b]. *Visallo, GitHub Repository*. GitHub. 14th May 2017. https://github.com/visallo/visallo (cited on page 12).

W3C [2017]. *Semantic Web - Ontologies*. World-Wide Web Consortium. 14th May 2017. https://w3.org/standards/semanticweb/ontology (cited on page 12).

Wright, William, David Schroh, Pascale Proulx, Alex Skaburskis and Brian Cort [2006]. "The Sandbox for Analysis: Concepts and Methods". In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pages 801–810. doi:10.1145/1124772.1124890. http://dl.acm.org/citation.cfm?id=1124890 (cited on page 5).

Zhao, Ying, Yanni Peng, Wei Huang, Yong Li, Fangfang Zhou, Zhifang Liao and Kang Zhang [2014]. "A collaborative visual analytics of trajectory and transaction data for digital forensics: VAST 2014 Mini-Challenge 2: Award for outstanding visualization and analysis". In: *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE. 2014, pages 371–372. doi:10.1109/VAST.2014.7042571. http://ieeexplore.ieee.org/document/7042571/ (cited on pages 22–23).