Visual Analytics Tool in Assistance of Disappearance at GAStech Investigation

Thiam Wai Chua*

ABSTRACT

In nowadays, a good visualization speaks thousand words for allowing to display large amounts of data while ensuring fast understanding of the information. An effective and meaningful visual analytics (VA) tool makes it easier to identify trends and patterns.

This report proposes a design of an interactive VA tool to assist the local authorities or local law enforcement from Kronos and Tethys to assess the situation and figure out the disappearance of employees at Thetys-based GAStech through the dataset of communication of emails between GAStech employees.

The visualizations in the designed interactive VA tool consists of t-SNE, heatmap, dynamic network graph (parallel coordinates) and it is built upon JavaScript libraries: react.js, D3.js, papaparser.js and spacy.js.

Index Terms: Interactive visual analytics tool—t-SNE—heatmap—Dynamic network graph—JavaScript

1 INTRODUCTION

This project focuses on defining a visual analytics (VA) problem, proposing a VA tool to support finding a solution to the problem, and then demonstration of the designed VA tool.

For approximately twenty years, Tethys-based GAStech operated a natural gas production site in the island country of Kronos, having produced remarkable profits as well as strong relationships with the government of Kronos. However, GAStech was not as successful in demonstrating environmental stewardship. In January 2014, the leaders of GAStech celebrated their new-found fortune, resulting from the initial public offering of their successful company. Along this celebration, several employees of GAStech go missing. There is an organization, the Protectors of Kronos (POK), which needs to be verified involvement in this case.

The challenge is based on a text–based data collection concerning the disappearance of GAStech employees and the analysis is focused on the assessment of the situation regarding of disappearance of these employees. The social movement group POK is suspected in the disappearance, but this needs to be corroborated. A visual analytics tool is designed to assist the local authorities or local law enforcement from Kronos and Tethys to assess the situation through analyzing the communication of emails.

2 BACKGROUND

The computational analysis of document text has been pointed as an approach to this type of problems, including text mining [3, 8]. In this section, we discuss the dimensionality reduction technique (t-SNE) used in the VA tool in Section 2.1 and text mining technique for preprocessing of the dataset in Section 2.2.

2.1 t-SNE

Dimensionality reduction (DR) methods can be used to create low dimensional (typically 2-dimensional) representations to visualize and explore patterns and dominant structure of high-dimensional datasets. Non-linear dimensionality reduction methods are particularly powerful for being capable to preserve local structures in the embedding while showing global information, for instance clusters existing at several scales. The popular methods are stochastic neighbour embedding (SNE) [9], t-distributed stochastic neighbour embedding (t-SNE) [15], LargeVis [13], and Uniform Manifold Approximation and Projection (UMAP) [10]. The most frequently performed dimensionality reduction method is t-SNE because of excelling to reveal local structures in high-dimensional data.

Firstly, the t-SNE constructs a probability distribution on pairs in higher dimensions such as similar objects are assigned a high probability and dissimilar objects are assigned lower probability. Then, t-SNE replicates the same probability distribution on lower dimensions iteratively until Kullback-Leibler (KL) divergence is minimized. KL divergence [5] is a measure of difference between probability distributions in higher dimension and lower dimensions and it is given as the expected value of the algorithm of the difference of these probability distributions.

Mathematically, in higher dimensions, the input dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is taken to define a probability distribution for a random variable e, of which the value domain indexed by all pairs (i, j) of indices $i, j \in [1..n]$ with $i \neq j$. This distribution is determined by specifying probabilities $0 \leq p_{i,j} \leq 1$ such that $\sum_{i \neq j} p_{ij} = 1$ which is equal to the probability that $e = (i, j)$. The distribution $\boldsymbol{p}_{ij}$ is defined as follows:

$$p_{ij} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(-\left\|x_k - x_l\right\|^2 / 2\sigma^2)}. \tag{1}$$

The goal of a t-SNE is to find another embedding $\boldsymbol{Y} \in \mathbb{R}^{n \times d'}$ in lower dimensions, from which another probability distribution q_{ij} is derived as follows:

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq l} (1 + \left\|y_k - y_l\right\|^2)^{-1}}. \tag{2}$$

An embedding $\boldsymbol{Y}$ is deemed better if the distance between these two probability distributions is smaller, as quantified by the KL divergence:

$$\mathrm{KL}(p||q) = \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right). \tag{3}$$

2.2 Text Mining

A possible use of text mining [6] is to identify words or expressions that repeatedly occur across the set of documents under analysis. Then, the resulting list will additionally include a small overview with rectangles representing each of the documents with small marks identifying term positions, with graphs of appearance count across documents as well as textual views with terms highlighted. The company then used a follow-on system, POSVis [12], to perform over the same documents a word-based part-of-speech analysis, and then displaying the results through pixel-based overviews, word clouds, and network diagrams.

*e-mail: t.w.chua1985@gmail.com

3 Data and Task Abstraction

3.1 Data abstraction

The datasets contain current and historic news articles regarding GAStech, resumes of the company's employees, email headers from two weeks of internal GAStech company emails and employee records. They consist of a text–based data collection of 10 days of data. In this project, we mainly focus on visualization of the two datasets which are email headers and employee records in developing the VA tool.

Email headers (Figure 1) is a dataset that contains email subject throughout two weeks of internal GAStech company communication. Employee records (Figure 2) contain personal records of GAStech's employees such as date of birth, gender, citizenship, job, email and military information, etc. Both datasets are stored in comma-separated values (CSV) format. This information is helpful to assess the employees' disappearance through the email network.

For the email headers data, the main attributes are From, To, Date and Subject, The From and To columns are the item attribute (emails). The Date column is the temporal attribute and in sequential order, whereas the Subject (email header) column belongs to item attribute type. The key attribute in the table is the Subject which is used to construct the distance matrix and then the matrix is to compute the embedded data through t-SNE algorithm. After that, the item attribute in the date column is then encoded in the t-SNE visualization.

For the employee records data, the main attributes are LastName, FirstName, CurrentEmploymentType, CurrentEmploymentTitle and EmailAddress. All these main attributes are considered as key attributes in the table and they are used as a look-up item attribute in the table and then encoded in a data visualization. The availability of both datasets is static because the entire datasets are available all at once in a period between 6 January 2014 and 17 January 2014.

	A	B	C	D	E	F	G	H
1	From	To	Date	Subject				
2	Sven.Flecha@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kron	1/6/2014 8:39	GT-SeismicProcessorPro Bug Report				
3	Kanon.Herrero@gastech.com.kronos	Felix.Resumir@gastech.com.kronos, Hideki.Cocinaro@gastech.c	1/6/2014 8:58	Inspection request for site				
4	Bertrand.Ovan@gastech.com.kronos	Emile.Arpa@gastech.com.kronos, Varro.Awelon@gastech.com.k	1/6/2014 9:28	New refueling policies - Effective February 1				
5	Valeria.Morlun@gastech.com.kronos	Dante.Coginian@gastech.com.kronos, Albina.Hafon@gastech.co	1/6/2014 9:38	Route suggestion for next shift				
6	Mat.Bramar@gastech.com.kronos	Rachel.Pantanal@gastech.com.kronos, Lars.Azada@gastech.com.	1/6/2014 9:49	Upcoming birthdays				
7	Bertrand.Ovan@gastech.com.kronos	Emile.Arpa@gastech.com.kronos, Varro.Awelon@gastech.com.k	1/6/2014 10:01	Don't text and drive!				
8	Orhan.Strum@gastech.com.kronos	Stenig.Fusil@gastech.com.kronos	1/6/2014 10:25	Service anniversary				
9	Varja.Lagos@gastech.com.kronos	Varja.Lagos@gastech.com.kronos, Hennie.Osvaldo@gastech.com	1/6/2014 10:28	Patrol schedule changes				
10	Brand.Tempestad@gastech.com.kronos	Birgitta.Frente@gastech.com.kronos, Lars.Azada@gastech.com.k	1/6/2014 10:35	Wellhead flow rate data				
11	Isak.Baza@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kron	1/6/2014 10:43	RE: GT-SeismicProcessorPro Bug Report				
12	Lucas.Alcazar@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kron	1/6/2014 10:50	RE: GT-SeismicProcessorPro Bug Report				
13	Linnea.Bergen@gastech.com.kronos	Rachel.Pantanal@gastech.com.kronos, Lars.Azada@gastech.com.	1/6/2014 11:00	RE: Upcoming birthdays				
14	Loreto.Bodrogi@gastech.com.kronos	Ingrid.Barranco@gastech.com.kronos	1/6/2014 11:04	Plants				
15	Linnea.Bergen@gastech.com.kronos	Rachel.Pantanal@gastech.com.kronos, Lars.Azada@gastech.com.	1/6/2014 11:16	RE: Upcoming birthdays				
16	Lucas.Alcazar@gastech.com.kronos	Nils.Calixto@gastech.com.kronos, Sven.Flecha@gastech.com.kro	1/6/2014 11:23	GT-SeismicProcessorPro Bug Report				
17	Isak.Baza@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kron	1/6/2014 11:34	RE: GT-SeismicProcessorPro Bug Report				
18	Lucas.Alcazar@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kron	1/6/2014 11:40	RE: GT-SeismicProcessorPro Bug Report				
19	Sten.Sanjorge Jr.@gastech.com.tethys	Orhan.Strum@gastech.com.kronos, Willem.Vasco-Pais@gastech.	1/6/2014 11:53	IPO				
20	Mat.Bramar@gastech.com.kronos	Mat.Bramar@gastech.com.kronos, Anda.Ribera@gastech.com.kr	1/6/2014 12:05	Craft night				
21	Lucas.Alcazar@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kron	1/6/2014 12:05	RE: GT-SeismicProcessorPro Bug Report				
22	Ingrid.Barranco@gastech.com.kronos	Loreto.Bodrogi@gastech.com.kronos	1/6/2014 12:15	RE: Plants				
23	Axel.Calzas@gastech.com.kronos	Felix.Balas@gastech.com.kronos, Lars.Azada@gastech.com.kron	1/6/2014 12:15	Wellhead flow rate data				

Figure 1: Screenshot of part of the email headers file.

3.2 Task abstraction

The main goal of this challenge is to develop a VA tool to assist the local authorities and local law enforcement from Kronos and Tethys to discover the suspicious employees connected to the employees' disappearance through the email communication throughout the 10-day period between 6 January 2014 and 17 January 2014. This the visual analytics problem.

Thus, this analysis aims to explore the patterns and insight of emailing between employees, as well as to answer following research questions:

1. How does the communication network change over time?
2. Which people communicate with each other?
3. Does the communication between people change between the first and second week?

Hypothetically, the local law enforcement personnel would acquire the data of the email and the employee records from GAStech,

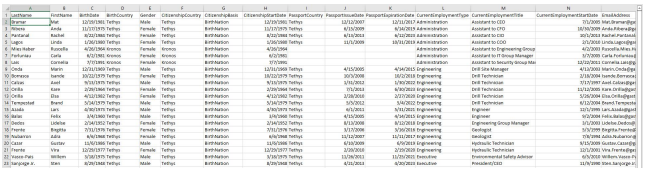

Figure 2: Screenshot of part of the employee records file.

then make sure these datasets are stored in CSV format. He/She could then turn to this VA tool to study patterns of the email communication network. The t-SNE provides all email headers which are clustered semantically, more information about a selected cluster can be selected and visualized. The law enforcement personnel could use his/her domain knowledge to direct this analysis and perform the assessment of the employees' disappearance. Then, detect the suspected employees.

According to Munzner [11], three levels of actions define the user's goals. The highest-level action describes how the visualizations are used to analyzed by either *consuming* or *producing* data. In this project, the VA tool needs to do both tasks. On the other hand, the dataset is consumed for computation. This can be further distinguished into the need to *discover* the pattern of the emailing network. This is used for answering the abovementioned questions.

Next, the end user (local law enforcement personnel in this case) needs to *annotate*, *record* and *derive* information from the visualizations in the VA tool. The *annotate* refers to adding graphical or textual details, such as adding circles to a selection in the t-SNE in this case. The *record* refers refers to capturing information and artifact about the visualization session, which can be achieved by saving the visualizations in their current condition as a screenshot, for example. Next, the *derive* refers to producing new data elements based on current data items, in this VA tool, the t-SNE plot is derived from the dataset and the dynamic network graph is derived from the selected circles in the t-SNE.

The mid-level actions are classified as *search* action, it is subdivided into two parameters (*identity* and *location* of the items of interest) with two values (*known* and *unknown*) each. The end users of this VA tool know where they are going to find what they need, hence the *location* in *known*. Also, the *identity* of what exactly they are looking for is *known*, since the range of information to be discovered has been constrained in the t-SNE.

Now that the search action has been defined, we can move to the lowest-level action, *quey*. This level is subdivided into three scopes: *identification*, *comparison* and *summarizing*. Only *identification* and *summarizing* apply to our VA tool. The user needs to *identify* the suspected employees through *summarize* the overview of the email network to be able to evaluate the situation.

Based on the Munzner [11], two of the three high-level targets are used in this VA tool, which are *trends* and *features*. The pattern of emailing network is shown in the heatmap, t-SNE and dynamic network graph. The features of interest of the datasets can be identified in the senders' list, senders' employment title, senders' email subject and date. Understanding the trends, distribution and correlations of these attributes. the emailing network is important for the local law enforcement personnel to answer the questions shown in this Section 3.2.

4 Visual Analytics Tool

Our VA tool is composed of three different views which give different perspectives on the GAStech dataset challenge. The overall tool is based on a web-app built using the React Framework and D3.js as a visualization library. In this section, the preprocessing of the dataset, design decisions and visual encoding of the dataset using the d3 library are discussed.

4.1 Preprocessing

The preprocessing of the dataset for plotting the heatmap and t-SNE on daily basis is performed in two steps. In the first step, as the dataset is static as discussed in section 3.1, thus we can divide and filter the data manually by exporting only the useful columns from the datasets. After that, we used an open-source natural processing language (NLP) classifier, spaCy, on the selected data to compute the similarity score for each employee based on the semantics of each of its email subject header from the set of sent emails. The score was done using the cosine similarity using an average of word vectors. This preprocessing of the data was done in a separate Jupyter Notebook and then exported as CSV format.

For plotting the t-SNE and dynamic network graph for the total period, a square distance matrix of all email subjects is constructed based on the text semantic similarity using cosine similarity in JavaScript and HTML [4, 14]. The dimension of the square distance matrix is $m \times m$, where m is the number of tuples in the dataset (email headers file in Figure 1). The element in the matrix range from 0.0 to 1.0, where 0.0 and 1.0 means low and high similarity between email headers, respectively. After that, this distance matrix is used to compute the t-SNE visualization that embeds the matrix in 2 dimensions.

4.2 Design Decisions

This section presents the visualization design of the tasks based on the research questions in Section 3.2. For every task abstraction, the choices for visual encodings and interaction design are explained. Furthermore, it illustrates how the design enables users to perform the tasks and answer the research questions. This is related to the third level of the nested four-level model for visualization design and evaluation as identified by Munzner [11].

First, we discuss briefly the VA tool framework design decision. The React.js [2] is chosen because it is a versatile graphical user interface (GUI) framework and the variety of user interface (UI) elements and components helped us to speed up building of the actual UI. Thus, we can spend more time on the actual visualizations. D3.js [1], this visualization library is chosen because it can be implemented and integrated easily within the React.js framework.

4.3 Visual Encoding and Interaction

In this section we discussed the chosen visual encoding and the interaction in each visualization. The visual encodings are divided into two main parts, which are visualization on daily basis and throughout total period. The dashboard of the VA tool is shown in Figure 3.

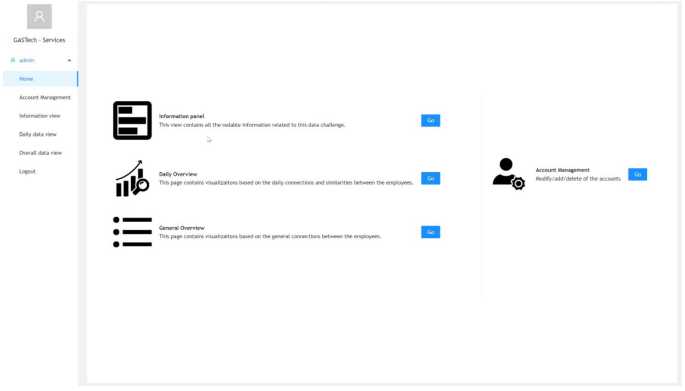

Figure 3: Dashboard

4.3.1 Visualize the articles

- **Visual Encoding:** Table idiom that show the article as shown in Figure 4 published from 2012 till 2014 and all the GAStech employee records. This provides the user to quickly look up the employee record and the articles in the tool. And then investigate the connection between employee, email, and the article by utilizing the domain knowledge of the user.

- **Interaction:** The user can use the scroll bar to locate the employee in the table and use the slider to locate the article for a specific date.

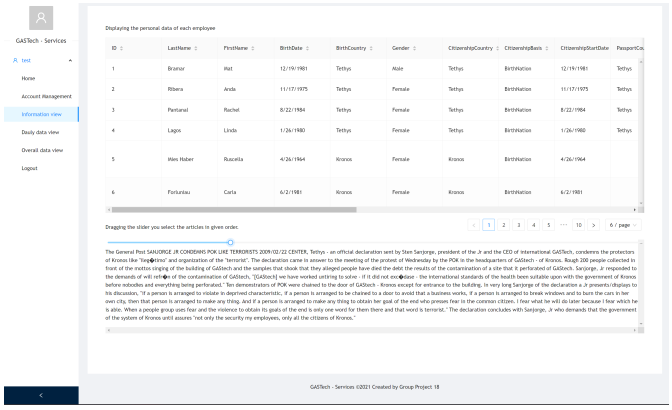

Figure 4: Articles published from 2012 to 2014 and the employee records.

4.3.2 Identify the similarity between employees defined by the email subjects on daily basis

- **Visual Encoding:** The heatmap idiom (left side of Figure 5) consists of two axes and cells with a value. The x-axis is represented by the ID of the employee (the ID of the employee is given during the preprocessing of the dataset), and the y-axis is represented by the string value of all the possible email subject headers. Each cell represents the similarity score of the set of emails of the specific employee that contains the corresponding email header. If two employees have the same cell score for a specific email header, it means that their set of email headers contain the exact header. This means that these employees most-likely have sent the same set of emails. Furthermore, it provides an overview of the data on a daily basis in order to reduce the information clutter. These values are colored using the palette from yellow to red. The cells with closer to yellow means the similarity score is low, whereas the cells with closer to red means the similarity score is high. This way the user able to identify the cells with high similarity score (red color) instantly.

- **Interaction:** The user can hover the mouse over each cell to view its similarity score.

4.3.3 Identify the clusters of employees defined by the similarity score on daily basis

- **Visual Encoding:** t-SNE idiom (right side of Figure 5) consists of clusters of employees with the same set similarity score where each dot represents an employee. A cluster of dots represents employees with the similar similarity score. The color of the dots is colored using the color HEX coded in the program because the ID of the employee is a categorical attribute.

- **Interaction:** The user can hover over each dot to pop up a tooltip with its specific ID. Then, the user can check its scores in the heatmap. Furthermore, the user can replot the t-SNE by resetting the t-SNE parameters (perplexity, step and epsilon). The definition of these parameters are explained in Section 4.3.4.

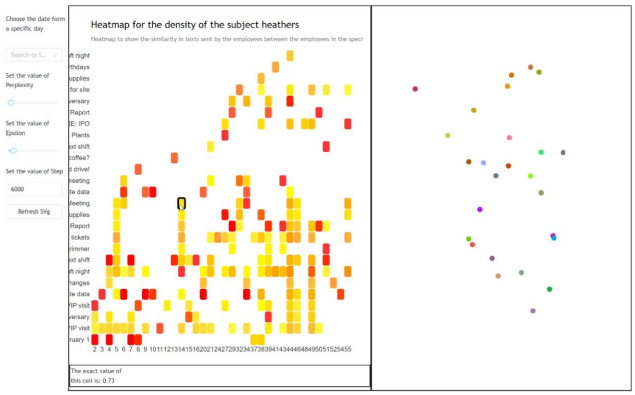

Figure 5: Heatmap on the daily data basis

4.3.4 Identify the clusters of employees defined by the email subjects semantically throughout total period

- **Visual Encoding:** t-SNE idiom that shows the clusters of the semantically similar email headers where each dot is colored by the value of the date. These values are colored using the palette Viridis (yellow to green and then to purple) because the email headers dataset is stored in sequential order in date. The dots with color closer to yellow mean its date is closer to the beginning of the period, dots with green mean in the middle of the period and the dots with purple mean they are closer to the end of the period. This way the user able to identify the date in the t-SNE easily. The t-SNE plot is shown in Figure 6.

- **Interaction:** The user can replot the t-SNE by adjusting the t-SNE parameters, they are perplexity, step and epsilon. The perplexity can be interpreted as a smooth measure of the effective number of neighbors. The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50 [15]. The step is the iteration number to compute the t-SNE, the solution gets better with higher steps. The epsilon is the learning rate for the gradient descent optimization. A too small learning rate leads to slow convergence, and a too large learning rate could lead to possible divergence. The dots in the t-SNE can be selected with the box selection for plotting the dynamic network graph, as well as display the senders' email subject and senders' email.

4.3.5 Identify the clusters of employees defined by the email subjects semantically throughout total period

- **Visual Encoding:** The dynamics network graph idiom is shown by using the parallel coordinates. The colored lines in the parallel coordinates represent the email communication network from the senders to the receivers throughout one or multiple days after selecting the dots in the t-SNE. The parallel vertical axes represent the sender and receivers. The space between the parallel vertical axes is the email's sending date. The name of the senders and receivers are placed in a separated column to avoid the blind spot in the visualization. The color of the lines is colored using the color HEX coded in the program because the email is the categorical attribute. Besides that, the dashboad also show senders' name and senders' email subject of the selected dots in the t-SNE. The color of the senders' name and email subject is identical with the line color in the parallel coordinates for easy identification.

- **Interaction:** The user can hover a line in the parallel coordinates to look up the pattern of the email network for a sender, i.e. only the hovered lines are highlighted. The user can also update the parallel coordinates, senders' name and email subject by re-selecting the dots in the t-SNE.

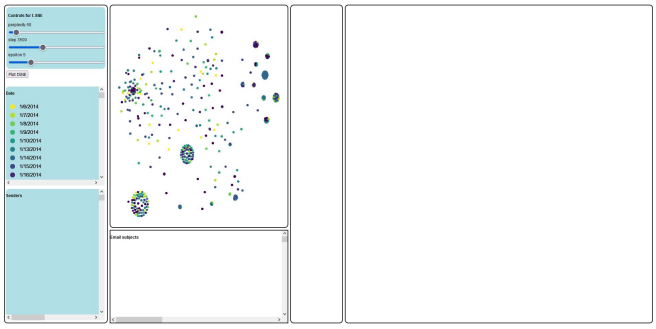

Figure 6: t-SNE visualization with some settings (perplexity, step and epsilon).

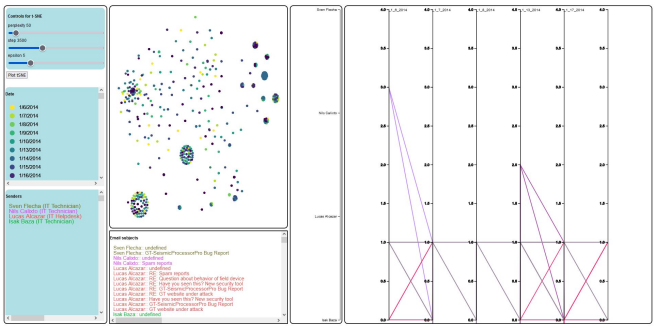

Figure 7: Dynamic network graph by using the parallel coordinates and the displayed senders' name and email subject after selecting the dots in the t-SNE.

5 Implementation

This section describes the use of languages and libraries to create the VA tool.

5.0.1 Languages

The main programming language is JavaScript because the framework is written in JavaScript, and it has many packages and libraries, for designing the user interface (UI).

The secondary programming language is Python for preprocessing and precomputing the similarity scores of the dataset. This language is chosen for computing the similarity scores using the spaCy library [7].

5.0.2 Libraries

- **React.js:** this is used as the main framework, working on the Node server. Every state changing and the logic of the visualization tool was done using the state interaction in React thus providing a more fluid experience.

- **Ant Design:** this is used as the main UI library. Every UI component was done using components and subcomponents of the ANt-Design with their full functionality.
- **D3.js:** this is used as the main visualization library. All visualizations are done using D3 and sub-libraries of D3.
- **Papaparse.js :** this is used in order to read the data from the files in a sync manner in order to reduce the delay of the pages when loading.
- **spaCy:** this is used for the NLP part of the data in order to extract the main features of the email subject headers and compute the similarity scores.

6 Results and Evaluation

In this section, we evaluate the effectiveness of the VA tool to perform the tasks in Section 3.2.

6.1 Changes of communication network over time

This VA tool primarily to help the local authorities from Kronos and Tethys to assess the situation of the disappearance of GAStech employees by analyzing the communication of emails. Initially, a user will therefore be interested in the semantic clusters of the email subjects throughout total period (10 days in this case). The user simply provides the data of emails headers and employee records and set the t-SNE parameters (primarily perplexity and step) to plot the semantic clusters of emails. After that, the parallel coordinates can be plotted after selecting the dots in t-SNE to visualize the changes of communication network over time.

The accuracy of the interpretation of the t-SNE is however in the hands of the user. Figure 6 shows the t-SNE with the setting of perplexity equals to 50 (as suggested by [15]), step equals to 3500 and epsilon equals to 5. There are two large clusters and approximately nine small clusters. One of the large clusters is related to the travel and training based on its email subjects (e.g. *Babysitting recommendations*, *Can someone cover for me next week?*, *Question - new travel forms*, *Training opportunity*, *Training question*, etc.). Whereas, another large cluster is related to general business of GAStech according to its email subjects (e.g. *Equipment audit approaching*, *Daily morning announcements*, *Updated Safety Policies*, *All staff announcement*, *Employee of the month*, etc.). From the color of dots in these two large clusters, they generally cover the total period of 10 days. Therefore, we can infer that this communication pattern happen in daily basis, and they might not be of interest to the local authorities.

On the other hands, the small clusters might be of interest to the local authorities because they are related to the information technology (IT) of GAStech, for example, the emails in one of the small clusters are *Have you seen this? New security tool*, *GT website under attack*, *Question about behavior of field device*, *Spam reports*, etc. The GAStech under attack happened on 6 January, 7 January, 8 January, 13 January, 17 January and reported by Lucas Alcazar to Isak Baza. Hence, it might be useful for the local authorities to ask Lucas Alcazar and Isak Baza for investigation assistance, especially, the detail of the GAStech website under attack incident. Questions such as *How is behind the GAStech website attacks?*, *What is the motive of the website attack?*, *Is this linked to the disappearance of the GAStech employees?* etc. can be asked.

6.2 Communication network between people

The user have discovered the email communication pattern global view (throughout the total period) and have found out the insight which like to investigate in detail. The user can study the communication network pattern between people on daily basis from the heatmap (Figure 5). For example, in the drop-down list, the user can select a specific date, and then the heatmap visual encoding will be plotted. The user can study the heatmap row by row. From Figure 5, we can see that the fifth row from the bottom has many cells which are close to red color, i.e. this semantically similar email subject is sent by many employees on this date. After that, the user can request these emails from GAStech for further investigation. This increases the efficiency of the investigation by eliminating many not important emails (for example, the second row from the bottom as all the cells in this row are yellowish). Besides selecting the row with most reddish cells, another conservative approach is that the user can also select the rows have more than five reddish cells. So that, none of the possible suspicious emails is ignored.

6.3 Changes of communication between first and second week

The difference of communication network pattern between the first week (between 6 January 2014 and 10 January 2014) and second week (between 13 January 2014 and 17 January 2014) can be observed from the t-SNE (Figure 8). We can see that most of the 'first week' dots (yellow to green) are scattered individually. The two large clusters contain the dots of all 10 days period. However, the dots in the small clusters are mostly 'second week' dots (green to purple). Therefore, it might be not of interest to the user to investigate the scattered and the two large clusters dots.

One the of small clusters only contains the dots of one day (14 January 2014), its email headers are *Take a look at this*, *Seeing strange network activity*, *Man your battlestations!*, *Can someone verify this behavior?*, etc. which suggests the suspicious activities went on in the second week. Hence, the local authorities can investigate the events happened on 14 January 2014. This might be connected to the disappearance of the GAStech employees.

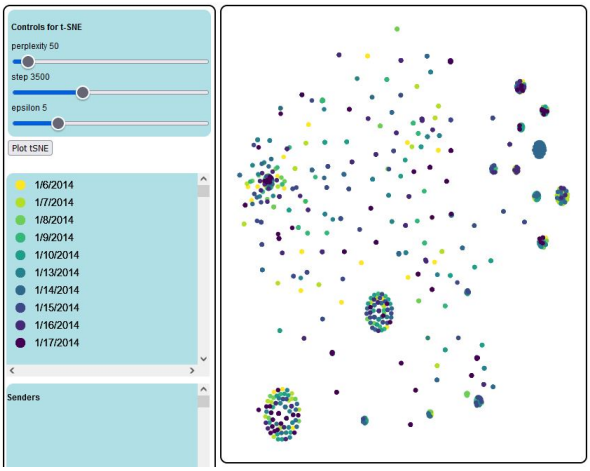

Figure 8: Close up view of the t-SNE from Figure 6.

7 Discussion

Overall, the resulting visualization is an effective VA tool for discovering the email networking pattern to investigate the disappearance of the GAStech employees. However, in the future this VA tool could be improved further. For example, the order of the cells in the heatmap could be sorted for meaningful representation of the data, i.e. the cells can be relocated in the descending order of the similarity score from the upper left corner of the heatmap along the diagonal. Besides that, for further improvement of the heatmap, the threshold of the similarity score and number of cells which have higher than this threshold can be set, after that the emails which meet these requirements will be highlighted. This way, the user can identify the suspicious emails and employees ID instantly.

For the t-SNE, we have tried to integrate the zooming and panning functionalities so that the user could take a closer look at each cluster. To do so, however, the box selection on the zoomed t-SNE gives an incorrect parallel coordinates due to the limitation of the D3.js. The possible solution is to add a button such that the zoomed t-SNE is reploted to original visual encoding after clicking the button, i.e. the user can only perform the box selection on the original visual encoding. Another interesting function is to include the real-time update of the dots' radius through a slider, this is interesting because the dots with later dates (e.g. 16 or 17 January) will cover the dots with earlier dates (e.g. 6 or 7 January) if these dots are proximity. Thus, the dots' radius adjustment enable to disclose the hidden dots.

For the dynamic network graph by using the parallel coordinates idiom which provides the emailing network pattern throughout the period, the lines are highlighted when the user hovered on that line. Besides highlighting the lines, it is interesting to highlight the sender's name displayed in the senders' name Document Object Model (DOM) and show the tooltip with displaying the sender's and receivers' name, unfortunately, the D3.js parallel coordinates library does not provide such functionality.

During the development stage of the VA tool, a few changes to the visualization design have been taken place. Initially, the force-directed network graph was selected as the dynamic network graph. However, this visual encoding does not provide a meaningful picture of changes of email communication network over time due to large number of emails. Besides the force-directed network graph, the radial and arc static networks were implemented in the VA tool. However, both static networks have the limitation to display the dynamic network graph of the emailing communication. Eventually, these visual encodings were removed from the design of the VA tool and parallel coordinates is selected.

Another interesting addition to this VA tool could be to include the sentiment analysis could be on the articles dataset using term frequency-inverse document frequency (TF-IDF), hence a connection between the articles and the email headers data could be investigated. Besides that, besides the t-SNE, other dimensionality reduction techniques such as, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Multi-Dimensional Scaling (MDS) and Uniform Manifold Approximation and Projection (UMAP) could be implemented and then comparison between them can be made. This gives the user flexibility to select the meaningful dimensional reduction visual encoding for further investigation.

Finally, the effectiveness of this visualization tool relies heavily on the domain knowledge of the local authorities. The visualization design assumes that the user is capable of directing the analysis by investigating the right settings for the t-SNE, right interpretation of the t-SNE and the heatmap and then extract the meaningful pattern emailing communication network from the parallel coordinates. He or she is required to be familiar with the t-SNE and parallel coordinates and to be able to interpret them. For example, simply selecting the very low or high perplexity produces a low accuracy t-SNE visual encoding.

8 Conclusion

As the complexity of the investigation for the disappearance of a human rises with the involvement of technology and higher dimensions of the dataset (i.e. large number of tuples and attributes), the authorities (e.g. local law enforcement, police, etc.) do not yet have an appropriate and advance method to assist them with the investigation. Our designed VA tool bridge the gap between the data dimensionality reduction technique used to visualize the complexity of the data in 2 dimensions and the end user (local authorities) who is might not able to interpret this multi-dimensional data. This VA tool is composed of three different views to provide different perspectives on the GAStech dataset such as heatmap, t-SNE and the dynamic network graph by using parallel coordinates. The t-SNE and parallel coordinates are connected using interactions which allow to select different cluster in the t-SNE. The interactions are used to guide the user through this analysis.

With this VA tool, it is possible to perform the visual analysis on the global view basis (i.e. total period) and then zoom in to investigate the suspicious part in the data (i.e. daily basis). It also gives the list of the senders' name and emails so that the user able to pinpoint swiftly the employee who is suspicious or who is able to assist the investigation. And then, the user can use the tool to answer the questions during investigation of cases, human disappearance in this case. The examples of questions are as in Section 3.2. For instance, this tool can be used to identify the pattern of email network over time to the question of how does the communication network changes over time or when do the abnormal activities happened.

While this VA tool is successful in visualizing the email communication network in big picture (t-SNE) and then zoom in to the suspicious part with different visual encoding (e.g. parallel coordinates), there is room for improvement to become a mature case investigation application. Extending the dimensional reduction techniques and introducing more advanced user interactions are necessary to develop this tool. The field of visual analytics is quickly developing and hopefully that this VA application will help the law enforcement authorities to perform the investigation efficiently in the future.

References

[1] D3.js. `https://d3js.org/`. [Online; accessed 19-April-2022].

[2] React.js. `https://reactjs.org/`. [Online; accessed 19-April-2022].

[3] B. W. Bader, M. W. Berry, and M. Browne. Clustering, classification, and retrieval. In *Survey of text mining II*. Springer, 2008.

[4] E. Berga. Building a text similarity checker using cosine similarity in javascript and html, 2020.

[5] O. M. Cliff, M. Prokopenko, and R. Fitch. Minimising the kullback–leibler divergence for model selection in distributed nonlinear systems. *Entropy*, 20(2):51, 2018.

[6] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 213–222, 2007.

[7] Explosion. spaCy. `https://spacy.io/`. [Online; accessed 20-April-2022].

[8] R. Feldman, J. Sanger, et al. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[9] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[10] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap.

[11] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015.

[12] R. V. T. C. C. Plaisant and A. Kumar. What's being said near "martha"? exploring name entities in literary text collections.

[13] J. Tang, J. Liu, M. Zhang, and Q. Mei. Visualizing large-scale and high-dimensional data. *WWW '16: Proceedings of the 25th International Conference on World Wide Web*, 01 2016. doi: 10.1145/2872427.2883041

[14] tomericco. Cosine similarity in javascript. `https://gist.github.com/tomericco/14b5ceac90d6eed6f9ba6cb5305f8fab`, 2019.

[15] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.