Who are the suspicious employees in GASTech?

Joanna Lian
Singapore Management University
joannalianyz@gmail.com

Ng Yen Ngee
Singapore Management University
ngyenngee@hotmaill.com

ABSTRACT

Through our project, we aim to use a range of data visualisation techniques to develop an R Shiny application to maximise the insights obtained from studying the different aspects of an individual's daily life e.g.credit card transactions and email communications in order to successfully profile an individual. Given the setting, the application aims to help the user determine if an individual in the company is suspicious or not through the analytics tools developed in the application. The analytics and design choices made in the development of the application and initial findings and future work are discussed.

1. INTRODUCTION

The goal of the annual Institute of Electrical and Electronics Engineers (IEEE) Visual Analytics Science and Technology (VAST) Challenge is to advance the field of visual analytics through competition. The 2021 IEEE VAST Challenge brings back a classic challenge from 2014 to see how approaches and techniques have developed since the original release of the challenge. The background of the challenge is as below:

In January 2014, the leaders of GAStech are celebrating their new-found fortune as a result of the initial public offering of their very successful company. In the midst of this celebration, several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem. It appears that certain employees of GAStech may be involved in the disappearance.

Aligning with our intent to create an application that aims to aid the profiling process, our application will serve as an interactive tool for users to visually investigate and identify which employee profile is indicative of suspicious behaviour. This paper documents our approach to designing and developing the interactive application targeted at crime investigators. This introduction is followed in Section 2 by an explanation of our motivation and objectives. Section 3 details the data used and methodology selected. Section 4 provides a visual overview of the final product. Section 5 concludes the report and offers ideas for further development.

2. LITERATURE REVIEW

Our project is a continuous and combination of previous analysis works from Lian (2021) and Ng (2021). The questions that we answered lead us to want to understand more about profiling of criminals. In a paper by Andrew P. Wheeler entitled "Crime Data Visualization for the Future", the author discusses that data visualization is necessary to be able to understand complex, quantitative information and he illustrates his point by using criminal data available in Dallas, Texas. This data includes homicide rates, timing at which burglaries occurred, GPS data from police cars etc. The main point the author would like to emphasize is that visualization methods help us to develop a deep understanding of the information we are analyzing along with being cognizant of how to effectively present that information in informative summaries.

Another paper, "Crime Pattern Analysis, Visualization And Prediction Using Data Mining" shared by Tushar Sonawanev and Shirin Shaikh explores using different analysis and visualization tools to better illustrate criminal patterns as they believe that crime mapping will help the law enforcement bodies to plan strategies for prevention of crime. The authors believe that an ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action. However, the reality at present is that there are many challenges faced in using the current tools available.

Inspired by Wong et al. (2020) application design, we are aligned with the sentiments of the two papers mentioned above and some of the suitable designs as a basis to form our project. The motivation and objectives of our project are elaborated on in the next section.

3. MOTIVATION & OBJECTIVES

Criminal profiling is used as a technique used to identify the perpetrator of a crime by identifying the personality and behavioral characteristics of the offender based upon an analysis of the crime committed. The sensational and dramatic elements of profiling is often portrayed in movies, television shows, and books but how does it fare as an investigative tool against crime in real life? Well, more often than not, there has been doubts within the crime-fighting community about the viability of profiling as an investigative tool. Why? This is because there are views that a criminal profile only gives a broad indication of the type of person who may have committed the crime. It does not indicate a specific individual who happens to fit the profile. The profiler is therefore unable to say whether it is more probable than not that a specific offender did, in fact, commit the crime. Nevertheless, criminal profiling has proven to be helpful in solving some criminal cases.

This is a fascinating topic and thus this project was formed to discover suspicious activities within GAStech itself which were worth investigating. We aim to create a data analytics applications to visualize these suspicious activities and relationships for users to decipher for themselves what defines suspicious behaviour and how can we make use of everyday data to raise red flags on suspicious behaviours, leading to identifying the suspicious people in GASTech. On a larger scale, this could assist in profiling which can be used for generic crime investigations.

We do this by exploring the following:

- Conducting exploration data analysis and inferential data analysis on
  - Credit card expenditure data
  - Email headers data
  - Employee categories
- Delivering a R-Shiny app that achieve the following through an interactive user interface design:
  - Identifying any anomalous or suspicious behavior.
  - Identifying formal (work-related) or informal (non-work related) relationships.
  - Discover any associations based on common interest given in the data.
  - Discover relationships between CC expenditure, email headers and employee records.
  - Decide who are the suspicious GASTech employees
  - Obtaining a holistic profile on these suspicious employees.

4. METHODOLOGY

Our methodology is as below:

- Data preparation using dplyr and other R packages.
- Analysis of VAST 21 data set with background research using some of the following methods:
  - Exploratory Data Analysis (EDA) methods in R.
  - Inferential Analysis methods in R.
  - Network Analysis in R.
- Creating a R shiny dashboard showing our findings/insights and conclusions:
  - R Markdown development for functionality checks
  - R-Shiny app development for user interactivity

4.1 Data

We'll be using data sets from Mini-Challenge 1 [MC1] and Mini-Challenge 2 [MC2].

From [MC1], we will be using the email headers and employee records data. From [MC2], we will be extracting insights based on credit card transactions data.

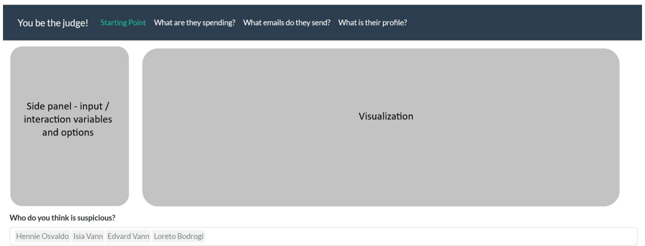
We will also be joining the two data sets based on individuals to analyzing attributes by features across data sets.

4.2 Shiny Architecture

Development of the interactive tool was done on Shiny, and R packages were used to build interactive web apps. Shiny is widely adopted in the data analytics industry because:

- It has a framework that makes it user-friendly to collect input values from a web page, with R code written as output values back to the web page.
- The input values can be modified by the user at any time, through interaction with customisable widgets.
- The output values react to changes in input values, with the resulting outputs being reflected immediately.

The main layout of our Shiny web app will consist of 3 main parts:



Who do you think is suspicious?

Hennie Osvaldo Isia Vann Edvard Vann Loreto Bodrogi

1. A top navigation bar that will help the user to navigate to the various modules that will be present in the blue area.
2. The main space where the visualization will be.
3. An input bar where user can select the names of the GASTech employees whom they decided are suspicious.

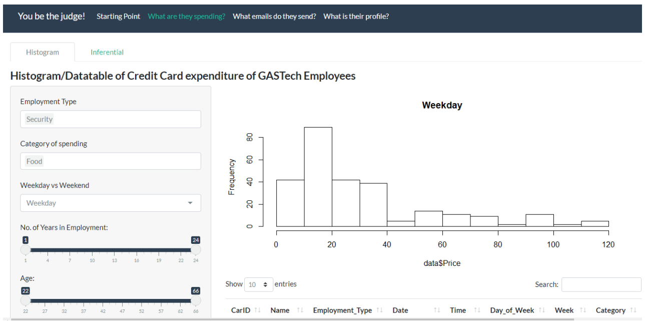
The objective of this layout is when the user switch from one tab to another, based on the visualization and their own analysis, they are able to add or remove the names of the their most suspected employees. Then at the end (in `Module 3`), we would see the profile of the user's decision of their most suspected employees.

Here are our Modules:

4.2.1 *Module 1 – What are the employees spending on?*

We will conduct EDA and inferential analysis on the credit card transaction data. On the top of the main visualization area, we will 2 tabs -

EDA (a histogram and a data table)



Histogram/Datatable of Credit Card expenditure of GASTech Employees

Employment Type: Security

Category of spending: Food

Weekday vs Weekend: Weekday

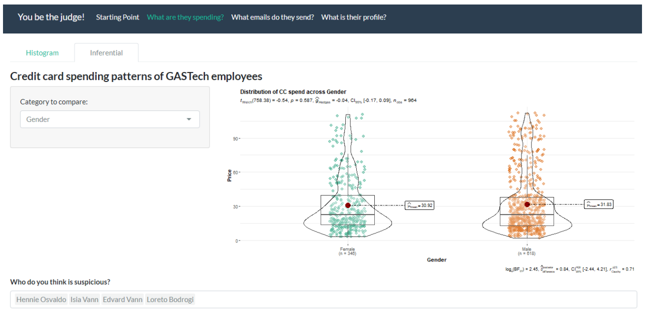
No. of Years in Employment: 1 – 24

Age: 22 – 66

Show 10 entries   Search:

CarID	Name	Employment_Type	Date	Time	Day_of_Week	Week	Category

Inferential analysis (violin plot and ANOVA statistical analysis results)


You be the judge!
Starting Point
What are they spending?
What emails do they send?
What is their profile?
Histogram
Inferential
Distribution of CC spend across Gender
$t_{Welch}(758.38) = -0.54$, $p = 0.587$, $\widehat{g}_{Hedges} = -0.04$, $CI_{95\%}$ [-0.17, 0.09], $n_{obs} = 964$
Price
0
30
60
90
$\widehat{\mu}_{mean} = 30.92$
$\widehat{\mu}_{mean} = 31.83$
Female
(n = 346)
Male
(n = 618)
Gender
$\log_e(BF_{01}) = 2.45$, $\widehat{\delta}^{posterior}_{difference} = 0.84$, $CI^{HDI}_{95\%}$ [-2.44, 4.21], $r^{JZS}_{Cauchy} = 0.71$

Credit card spending patterns of GASTech employees

Category to compare: Gender

Who do you think is suspicious? Hennie Osvaldo Isia Vann Edvard Vann Loreto Bodrogi

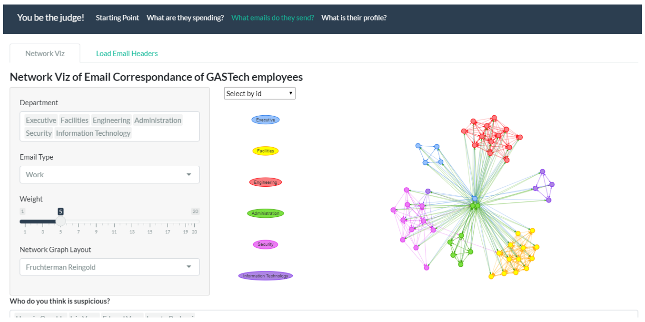
For each of the tabs, we will include options and filters for user interaction.

Depending on our selection of the various filters, the visualization in the main panel will be automatically filtered. From the controls on the side panel, we can explore

4.2.2 Module 2 – What emails do they send?

We utilize a Network Graph to visualize the connection between GASTech Employees via their email correspondence. On the top of the main visualization area, we have 2 tabs -

Network Visualization



Network Viz of Email Correspondance of GASTech employees

Department: Executive Facilities Engineering Administration Security Information Technology

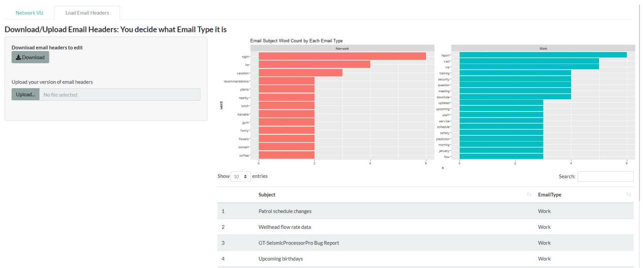
Email Type: Work

Weight: 5 (1 – 20)

Network Graph Layout: Fruchterman Reingold

Who do you think is suspicious?

Upload Email Headers



Download/Upload Email Headers: You decide what Email Type it is

Download email headers to edit

Upload your version of email headers: No file selected

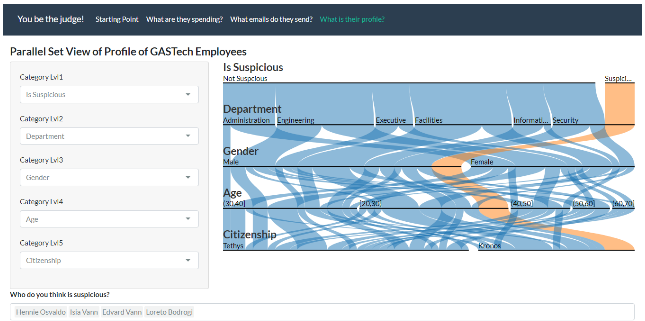
Show 10 entries   Search:

	Subject	EmailType
1	Patrol schedule changes	Work
2	Wellhead flow rate data	Work
3	GT-SeismicProcessorPro Bug Report	Work
4	Upcoming birthdays	Work

For the Network Visualization, we included filters for user interaction. The user is also able to download a template of the email headers and key in the email type category themselves to categorize the email headers if they wish to. The Network Visualization would change accordingly from the user's categorization if they uploaded their own version. Otherwise, the app will by default use the categories that has been pre-defined.

4.2.3 Module 3 - What is their profile?

We utilise a Parallel Set Plot to profile the employees with 5 degrees of interaction as shown below.



Parallel Set View of Profile of GASTech Employees

Category Lvl1: Is Suspicious

Category Lvl2: Department

Category Lvl3: Gender

Category Lvl4: Age

Category Lvl5: Citizenship

Who do you think is suspicious? Hennie Osvaldo Isia Vann Edvard Vann Loreto Bodrogi

To facilitate exploration by users, we have created a user guide that runs through the App step by step.

4.3 Analysis Techniques

4.3.1 Histogram

A histogram is a plot that illustrates the underlying frequency distribution of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution). In particular it allows us to spot outliers and skewness which will help identify suspicious transactions.

4.3.2 Violin Plot

A violin plot depicts distributions of numeric data for one or more groups using density curves. The width of each curve corresponds with the approximate frequency of data points in each region. Densities are frequently accompanied by an overlaid chart type, such as box plot, to provide additional information.

4.3.3 Network Graph

Network graphs show interconnections between a set of entities. Each entity is represented by a Node (or vertice). Connections between nodes are represented through links (or edges). In the context of our shiny app, employees represent the nodes while the links represent the email correspondences. It allows us to see the connections between identified suspicious people and find more suspicious employees.

4.3.4 *Parallel Set Plot*

Parallel Set Plots (ParSet) are a visualization method for the exploration of categorical data. They focus on the data frequencies instead of individual data points. The technique is built on the axis layout of parallel coordinates, with boxes representing the categories of data (e.g. gender, department, spending patterns etc.) and parallelograms between the axes showing the relations between categories. Our app enables the user to interactively remap the data to up to five levels of categorization.

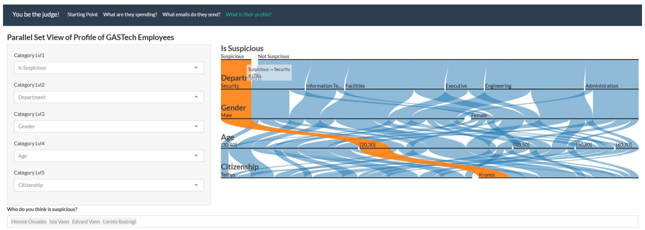
5. DESCRIPTION OF PRODUCT & FINDINGS

5.1 First Look

From our previous analysis, we know that these four GAStech employees have familial relationships with members of POK and hence are the most suspicious people in GAStech.

- Hennie Osvaldo
- Isia Vann
- Edvard Vann
- Loreto Bodrogi

Hence, this is our starting point. We use the ParSet view to see their profile:



You be the judge! Starting Point What are they spending? What emails do they send? What is their profile?

Parallel Set View of Profile of GASTech Employees

Category Lvl1: Is Suspicious

Category Lvl2: Department

Category Lvl3: Gender

Category Lvl4: Age

Category Lvl5: Citizenship

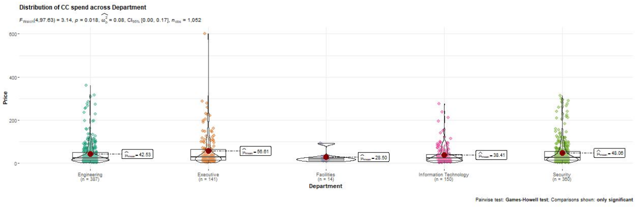
Who do you think is suspicious? Hennie Osvaldo, Isia Vann, Edvard Vann, Loreto Bodrogi

Suspicious → Security
4 (7%)

From here we can observe that these four people have many similarities. They are all males in their twenties and all from the security department, probably with a lot of drive and impulse to possibly plot an internal kidnapping. In our subsequent analysis, we will scrutinize closely the data with these characteristics.

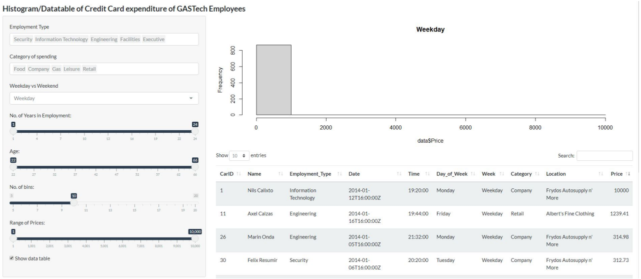
5.2 Credit Card Expenditure

Now looking at credit card transactions, we added all of the options so that the data is not filtered and then we sort the data table by price. We can immediately see that there are a few outliers. To ensure that the outliers are worth investigating, we select the `inferential` tab to visualize statistical analysis.



When we exclude the outliers, the analysis tells us that there is insufficient statistical evidence to reject the null hypothesis that mean of the transactions between the departments are the same. In fact, the means of the transactions hover between $25-60 which suggests the outliers that we spotted is worth investigating into.

Histogram for Weekday:



Histogram/Datatable of Credit Card expenditure of GASTech Employees

Employment Type: Security, Information Technology, Engineering, Facilities, Executive

Category of spending: Food, Company, Gas, Leisure, Retail

Weekday vs Weekend: Weekday

No. of Years in Employment: 1 – 24

Age: 22 – 66

No. of bins: 10

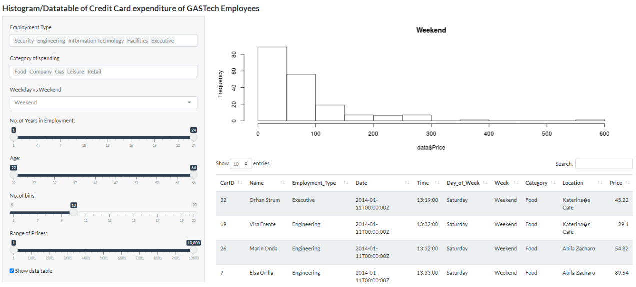
Range of Prices: 1 – 10,000

☑ Show data table

Show 10 entries Search:

CarID	Name	Employment_Type	Date	Time	Day_of_Week	Week	Category	Location	Price
1	Nils Calixto	Information Technology	2014-01-12T16:00:00Z	19:20:00	Monday	Weekday	Company	Frydos Autosupply n' More	10000
11	Axel Calzas	Engineering	2014-01-16T16:00:00Z	19:44:00	Friday	Weekday	Retail	Albert's Fine Clothing	1239.41
26	Marin Onda	Engineering	2014-01-05T16:00:00Z	21:32:00	Monday	Weekday	Company	Frydos Autosupply n' More	314.98
30	Felix Resumir	Security	2014-01-06T16:00:00Z	20:20:00	Tuesday	Weekday	Company	Frydos Autosupply n' More	312.73

Histogram for Weekend:



Histogram/Datatable of Credit Card expenditure of GASTech Employees

Employment Type: Security, Engineering, Information Technology, Facilities, Executive

Category of spending: Food, Company, Gas, Leisure, Retail

Weekday vs Weekend: Weekend

No. of Years in Employment: 1 – 24

Age: 22 – 66

No. of bins: 10

Range of Prices: 1 – 10,000

☑ Show data table

Show 10 entries Search:

CarID	Name	Employment_Type	Date	Time	Day_of_Week	Week	Category	Location	Price
32	Orhan Strum	Executive	2014-01-11T00:00:00Z	13:19:00	Saturday	Weekend	Food	Katerina�s Cafe	45.22
19	Vira Frente	Engineering	2014-01-11T00:00:00Z	13:32:00	Saturday	Weekend	Food	Katerina�s Cafe	29.1
26	Marin Onda	Engineering	2014-01-11T00:00:00Z	13:32:00	Saturday	Weekend	Food	Abila Zacharo	54.82
7	Elsa Orilla	Engineering	2014-01-11T00:00:00Z	13:33:00	Saturday	Weekend	Food	Abila Zacharo	89.54

We make the following observations:

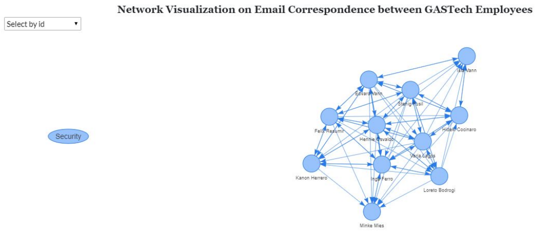
- Nils Calixto
  - Made a purchase of $10,000 at Frydos Autosupply n' More just two days before the kidnapping.
  - Other transactions are charged to the decimal place, but this particular transaction is very neat at $10,000.
  - if we compare to the other purchases within the same cateogory, most transactions fall below $500.
  - This single transaction make Nils Calixto very suspicious.
- Axel Calzas
  - Made a purchase of $1239.41 at Albert's Fine Clothing just 3 days before the kidnapping.
  - Most purchases made are around $300. It could be that Axel helped pick up for 3-4 person.
  - We consider Axel Calzas as suspicious first but we should look into his relationships in more detail.

- Sten Sanjorge Jr
  - Made a purchase of $600 at the Chostus Hotel just 3 nights before the kidnapping incident.
  - Sanjorge Jr was initially counted as one of the kidnapped victims but it turns out that he was just returning to Tethys. (add ref)
  - The combination of this news and this transaction makes the CEO a suspect as well.

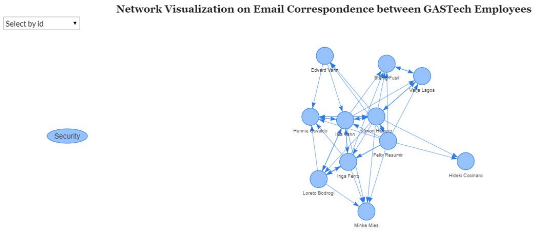
5.3 Email Correspondence

We focused straight into the correspondence within the security team and compare the differences between Work and Non Work below:

Network Visualization for Security Department - Work Correspondence:

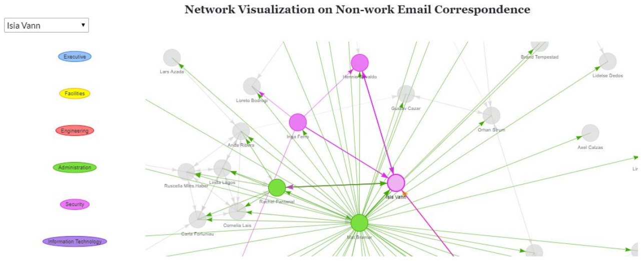


Network Visualization for Security Department - Non-Work Correspondence:



If we observe the network of Work Correspondence, we observe that the emails sent are sent rather consistently throughout the team, however if we look at the other visualization we can see quite obviously the imbalance. Isia Vann and Hennie Osvaldo seems to be in the center of all these correspondence. There are some members who only receive emails while some members where there are exchange of emails.

Let us scrutinize the non work email correspondence of Isia Vann with the rest of the organization:

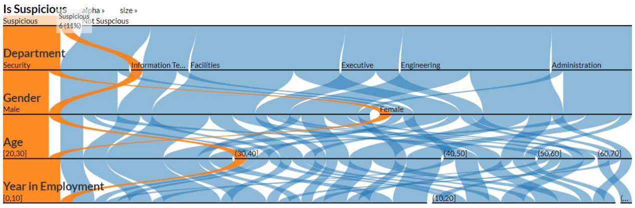


We found another member closely related to our current suspicious candidates - Inga Ferro who is in correspondance with at least 3 of the original suspicious candidates which makes this person suspicious as well.

When we select Axel Calzas, we realize that this person do not have non-work correspondence, and any work related emails are department of company based. This eliminates the suspicious off and we removed the name.

5.4 Parset Final

Here is the final profile of our judgement of suspicious employees in GASTech:



6. CONCLUSION & FURTHER IDEAS

This paper set out the development of a web application targeted at exploring how we could use visually motivated tools to facilitate the profiling process. The project was motivated by our realization through the VAST 2021 project that there were suspicious behaviors demonstrated by some individuals which were worth investigating.

The application was developed using the Shiny architecture on R, supported with a range of statistical packages to provide users with a whole range of techniques to derive insights from the data. Definitely though, the range of techniques and ideas executed in this project is non-exhaustive and We suggest that the below can be explored in the future development of the app:

- Additional user interaction features within and between each tab in the app to speed up the detection of suspicious individuals. E.g. users could be given the option to highlight certain transactions in the violin plot and that simultaneously triggers the network visualization tab to showcase any connections between the individual who made that transaction.
- The option of filtering transactions of identified suspicious employees to better understand their purchasing behaviour.
- Additional analysis on email headers based on who sent and who received could be visualized. E.g. when selecting non-work, there was a network from Mat Bramar who is the assistant to CEO to everyone, this means that the particular email header might have been identified wrongly even though it seemed that the email header is non-work related.
- The GPS dataset can be utilized further. An additional tab exploring the daily commutes of the employees could be used to further illustrate the daily routiness and the relationships between the employees.

7. REFERENCES

Hendricksen, J. (2013). *Deriving a digital profile from open source information.* Radboud Universitieit Nijimegen.

Lian, J. (2021, July 18). *Visual Analytics: Assignment - MC2.* https://isss608vizanalysis.netlify.app/posts/2021-07-18-assignment-mc2/

Ng, Y. N. (2021, July 25). *MC1: Findings and Discoveries.* https://yenngee-dataviz.netlify.app/post/2021-07-16-mc1-findings/#3-connections-and-relationships

Srikanth, H. S. T. (2021). Crime Pattern Analysis, Visualization and Pridiction Using Data Mining. *International Journal for Research in Applied Science and Engineering Technology*, 9(VIII), 397–401. https://doi.org/10.22214/ijraset.2021.37323

Wheeler, A. P. (2017). *Crime Data Visualization for the Future.* School of Economic, Political, and Policy Sciences – Criminology Program.

Wong, C., Dunsford, D. R., & Nair, V. (2020). *A Visual Exploration of Media Consumption in Singapore.* Https://Wiki.Smu.Edu.Sg/1920t2isss608. https://wiki.smu.edu.sg/1920t2isss608/Group02_research_paper