R Shiny Investigative Tool into GASTech Personnel Disappearance

Aryah Umralkar Chopra
School of Computing and Information Systems,
Singapore Management University
aryahc.2020@mitb.smu.edu.sg

Rhoda Tong Min Ting
School of Computing and Information Systems,
Singapore Management University
rhoda.tong.2020@mitb.smu.edu.sg

Davmes Tan Chee Seng
School of Computing and Information Systems,
Singapore Management University
davmes.tan.2020@mitb.smu.edu.sg

ABSTRACT

A fictitious scenario was created as part of VAST Challenge 2021. A group of staff members from GASTech, an oil and gas company situated on an island known as Abila on Kronos, had gone missing mysteriously. A group known as Protectors of Kronos (POK) was the prime suspect into the disappearance. A series of unprocessed data were made available to the law enforcement agencies to investigate on. The data were split across three mini challenges, which our team had undertaken Mini Challenge 1 and 2.

Mini Challenge 1 consists of email correspondences, employee records and resumes, historical documents and news articles. The ojective of this mini challenge is to identify the complex relationships among the people and organisations, and possibly infer the disappearance of GASTech to any individuals/group who might be involved. Mini Challenge 2 consists of gps tracking, transaction and loyalty card records, together with car assignment records - linked to the gps tracking data. The objective of this mini challenge is to discover anomalies and suspicious activites that may require additional investigating.

The objective of this research paper is to share on the methods and models used to develop an online investigative tool where a law enforcement at Kronos and Tethys could use, to piece the raw data into useable information and evidences.

1. INTRODUCTION

A fictitious scenario was created as part of VAST Challenge 2021. A group of staff members from GASTech, an oil and gas company situated on an island known as Abila on Kronos, had gone missing mysteriously. A group known as Protectors of Kronos (POK) was the prime suspect into the disappearance. While Mini Challenge 1 and 2 provided a set of raw data that allow investigators to establish and identify complex relationships among the people and organisations, discover anomalies and suspicious activities, such investigative work may require humongous man hours and effort, without data analytics and visualisation.

We would be using Shiny R to develop an online investigative tool to aid in the analysis into the disappearance of GASTech Personnel, allowing investigators to explore information and inferential statistics derived from the unprocessed data available.

2. MOTIVATION

The motivation of this project would be two-fold. First, the data presented to the investigators were raw and unprocessed, and to link and derive insights from these data would require tremendous man hours and effort. Second, while insights could be derived and useful information could be formed, there would be a need to present the information in a visually appealing format to facilitate information dissemination and to allow quick collective appreciation of events among the investigators.

To this end, we would be looking to develop a R Shiny app based on three principles: (a) informative; (b) intuitive and; (c) interactive. Our 3Is principles would have the data undergoing baseline cleaning, making them into suitable formats for subsequent processing for information delivery. The user-interface would be made intuitive so that the investigator would be able to use the application without much references to our user guide. The online investigative tool would also be interactive, such that the investigator would be able to provide varied inputs into the formation towards the final visualisation report.

The R Shiny app would comprise of two main modules: (a) Exploratory Data Analysis allowing investigators to draw information such as transaction records, employee records, email correspondences and such; (b) Inferential Statistics allowing investigators to infer relationship linkages among user-selected employees, possible coded words within email correspondences within an identified group of personnel, their movements towards identified locations and possible anomalies at the locations and transaction analysis using both credit card and loyalty card data.

3. REVIEW AND CRITIC OF PAST WORKS

Taking reference on a submission by students from International Institute of Information Technology Hyderabad, they had derived an interface that uses geotools to show the paths of the cars, moving at a particular time or date. While it was a useful tool to visualise the movement of the vehicles, it was unable to plot the locations where the vehicles may had visited. This would be addressed in one of our module.

While credit card data and loyalty card data were provided for, to the investigators, there are no direct linkages between both set of data, except for the locations and price of item. The timestamp for both data are recorded differently with the credit card indicating the time of transaction while loyalty data indicating date of transactions only. Reviewing past work from members of University of Calgary, they attempted Parallel Coordinates plot by linking the locations, timestamp, price and employees, to establish possible linkages among the four variables. Inspired by their work and the sharing from Prof Kam on using Parallel Coordinates Plot on R, we would attempt to create parallel plots and allow the investigators to decide the variables that they are keen to establish linkages with.

This application and charts, focused on the GPS movements of the vehicle. This provides the investigator with the time and location dimension, however, there will be the question unanswered,'What were they doing there?' This question was addressed in the first mini-challenge data of The Vast Challenge.

Submission from the University of Buenos Aires documented suspicious emails by combining the use of a visual network plot of the participants in an email chain and corresponding a table of emails. Like with most submissions, they were static and only highlighted emails that they deemed as suspicious with the search by employee. The merging of a visual and table representation of the data would be enhanced in our application so that the investigator can not only search by person, but also by key words (or both) that his analysis has deemed suspicious.

In addition to that, in the submission by Tianjin University, their application showed the movement of the email with time between recipients. This was narrowed down between 2 recipients only when certain emails are sent to many people at once or replied to many other people. This inspired us to build a network that shows the entire network of the particular subject.

While the email subjects raise suspicion and the analysis was subsequently based their search, the role that each of the employee plays in the network was not studied in the submissions. This will be built on in our application.

4. DESIGN FRAME

Our R Shiny Investigative Tool's design would be based on our 3Is principle.

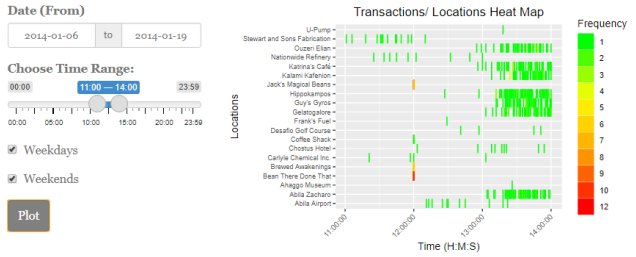
Informative - A series of informative modules would be made available under the Exploratory Data Analysis (EDA) tab, comprising of: (a) Locations Exploration; (b) Transactions Exploration; (c) Cards Exploration; (d) GASTech Employees Information; and (e) Email Correspondences. Details of each modules would be covered under the sub-sections later.

Intuitive - The entire UI design would be simple and intuitive, allowing investigators to use it without frequent references to our user guide. This would be done by ensuring that appropriate input methods were used and that the UI would be kept as clean and simple as possible.

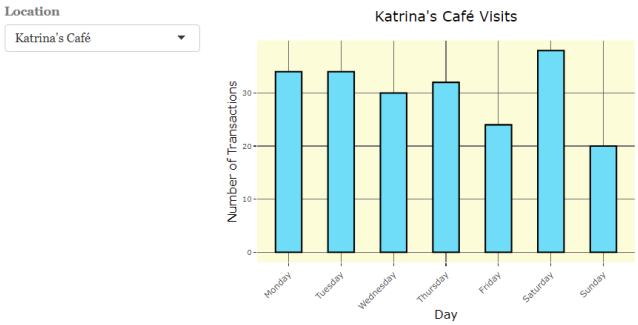
Interactive - The respective modules would require specific inputs from the investigators before a representative visualisation could be produced. It would be made interactive such that the investigators would have the flexibility to select various variables or inputs to create the desired visualisation. Similarly, the investigators would have the option to save the produced visualisation separately.

5. EXPLORATORY DATA ANALYSIS

5.1 Locations - For Exploration!



The intention of this module would to be allow the investigators to explore the popularity of the locations through the analysis of transaction records. It would be conducted in a manner by aggregating the transaction records based on the time of transactions. The transaction counts would be plotted against the aggregated time of transactions and locations using a heatmap to infer the peak periods for the locations. The selected records would be filtered based on the range of date inputs, time range of transactions, with the ability to toggle between weekdays, weekends or both.



The function of this module would be to identify the operating days of the locations through the analysis of credit card transactions, through the aggregation of records based on the day of transactions. The output of this visualisation would suggest the days where the locations would be having transactions, thus inferring their operating days of the week.

5.2 Transactions - For Exploration!

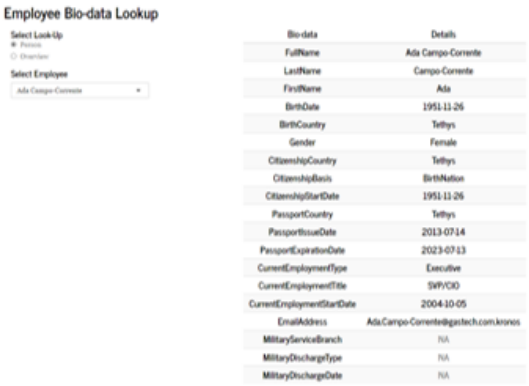
Duis nec purus sed neque porttitor tincidunt vitae quis augue. Donec porttitor aliquam ante, nec convallis nisl ornare eu. Morbi ut purus et justo commodo dignissim et nec nisl. Donec imperdiet tellus dolor, vel dignissim risus venenatis eu. Aliquam tempor imperdiet massa, nec fermentum tellus sollicitudin vulputate. Integer posuere porttitor pharetra. Praesent vehicula elementum diam a suscipit. Morbi viverra velit eget placerat pellentesque. Nunc congue augue non nisi ultrices tempor.

5.3 Individual Card Transactions

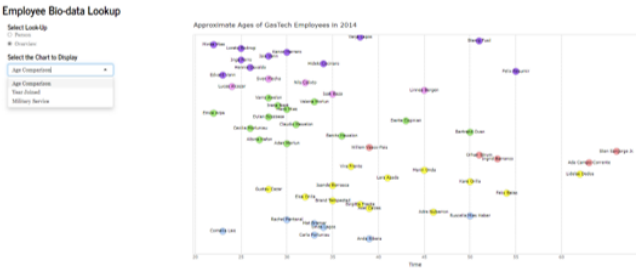
Duis nec purus sed neque porttitor tincidunt vitae quis augue. Donec porttitor aliquam ante, nec convallis nisl ornare eu. Morbi ut purus et justo commodo dignissim et nec nisl. Donec imperdiet tellus dolor, vel dignissim risus venenatis eu. Aliquam tempor imperdiet massa, nec fermentum tellus sollicitudin vulputate. Integer posuere porttitor pharetra. Praesent vehicula elementum diam a suscipit. Morbi viverra velit eget placerat pellentesque. Nunc congue augue non nisi ultrices tempor.

5.4 Employees of GasTech

This module explores the general data of the employees. Like with most investigative portals, this tab will allows the investigator to understand basic bio-data of the employee in question. As shown below, the tab will react to the drop down under *Select Employee* and update to display the respective information.



The tab was extended with three charts. On selection of *Overview*, this allowed the comparison of biodata that is comparable accross, for example, the year joined in the company, ages and military service. This are useful as they allow the investigator to understand possible relationship or acquintances outside of GasTech.



5.5 Email Correspondence

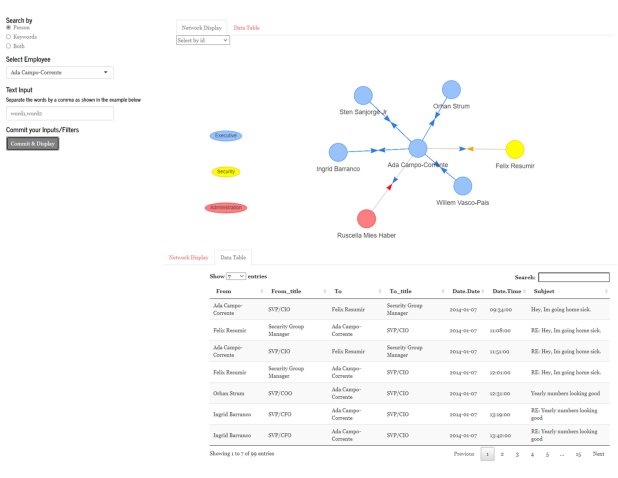
This module is a exploratory part, setting the ground for the Inferential Statistics in the network analysis. Investigators will be able to filter by data, time and employee to understand their email conversation. The main plot will show all emails to and from the employee across the day, while the 2 subplots highlight the unique words shared between filtered employee and their recipients.

The subplots also highlight oddly timed emails and frequency of conversation between the filtered employee and others.

6. INFERENTIAL STATISTICS

6.1 Email Network Analysis

This module will allow the investigator to make searches by person, keyword or both. The search will bring up both the visual representation of the conversation and also the corresponding data table.



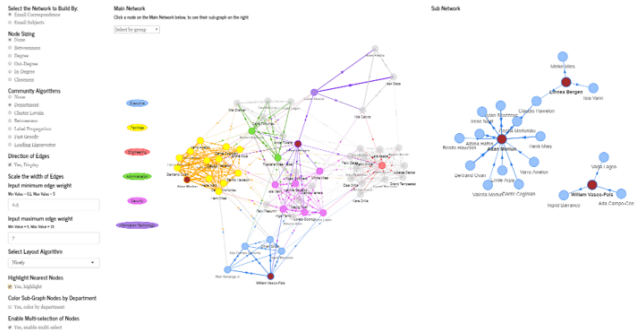
In the submission by University of Abila, the network was colored by suspected POK members too. As this application is meant for analytic deduction by the user, the nodes are colored by department to see inter-department activities. The network is directed. This will highlight nodes of different nature: Authoritative(tend to receive more emails), hubs (tend to receive more emails) and other critical intermediary nodes that facilitate the transfer of information. In addition to that, the network is interactive, the nodes highlight their nearest neighbours on click, showing flow of information. The nodes can be moved around to handle overlap issues.

This tab works in cohesion with the *Networks* tab.

6.2 Networks

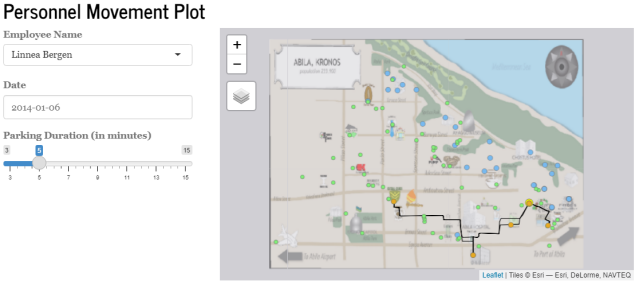
This tab will allow the investigator to see the overall email network at GasTech, and run layout, community detection algorithms and identify key nodes through centrality measures. There are one of 2 options the investigator can see, network by email converstaion or network by content of the email subjects.

The side panel contains all the measures that can be changed, this includes node sizing, community algorithm and layout algorithms.



This tab allows the use of network analysis algorithm to identify key nodes within the network. Similar to previous, the user has freedom to enable multi-select, highlight nearest nodes and also direct interaction with the network.

6.3 Employee Movement Plot



The data preparation for this module would be to identify Points-of-Interests (POI), using the gps records. Since the recording of gps would only occur while the vehicle is in motion, whenever there is a time interval (3g: 3 minutes), would be a good assumption that the vehicle is in parking mode, thus indicating it as a POI. The time interval may be adjusted based on the investigators' tolerance between 3 and 15 minutes. The POIs would be differentiated between homes of employees and other locations, thus allowing tmap to plot the route taken, and the stops taken.

The sub-module would be an extension where the selected gps plot would be sequenced based on the time of activity, and further identified based on marked locations such as their homes, GASTech or prominent locations that were named prior. The subsequent box plot would be an analysis of the duration of stop, by visitors, to the locations.

6.4 Transaction Amount Analysis

Duis nec purus sed neque porttitor tincidunt vitae quis augue. Donec porttitor aliquam ante, nec convallis nisl ornare eu. Morbi ut purus et justo commodo dignissim et nec nisl. Donec imperdiet tellus dolor, vel dignissim risus venenatis eu. Aliquam tempor imperdiet massa, nec fermentum tellus sollicitudin vulputate. Integer posuere porttitor pharetra. Praesent vehicula elementum diam a suscipit. Morbi viverra velit eget placerat pellentesque. Nunc congue augue non nisi ultrices tempor.

7. FUTURE WORKS

For the email network analysis, the data did not provide us with the actual content of the email. While subjects can be useful for the investigator to make reasonable judgments, there could be confirmation biases and logical fallacies with that for e.g. suspected members of the POK using similar subjects for different email content.

This can be addressed with the provision of the email content data. In addition to that, the text network can become more accurately with better data.

(Additional point from Rhoda on how the modules could be improved.)

An improvement towards the Employee Movement Plot could include further analysis on the movement route, in terms of the distance/ speed, which may suggest the vehicle were kept in idle mode, which is currently unable to detected under the current module. In addition, the streets of Abila could be identified based on the movement routes, and thus, and analysis could be conducted whether the vehicle was travelling in a suspicious manner.

8. ACKNOWLEDGEMENTS

The authors wish to thank Professor Kam Tin Seong of Singapore Management University for his extensive guidance and support during this project.

References

[1] Avil, Pilar; Burgos, Valeria; Chikka, Guaymás Amalia 2014. VAST Challenge 2014 Mini Challenge 1. *https://www.cs.umd.edu/hcil/varepository/VAST%20Challenge%202014/challenges/MC1%20-%20Disappearance%20at%20GASTech/entries/University%20of%20Buenos%20Aires%20-%20Avila/*.

[2] Kam, Tin Seong 2020. Hands-On Exercise 8: Creating Parallel Coordinate Plot. *https://rpubs.com/tskam/PCP*.

[3] Pochampally, Yashaswi; Yarrabelly, Navya; Chikka, Veera Raghavendra and Karlapalem, Karlapalem 2014. VAST Challenge 2014 Mini Challenge 2. *http://visualdata.wustl.edu/varepository/VAST%20Challenge%202014/challenges/MC2%20-%20Patterns%20of%20Life%20Analysis/entries/International%20Institute%20of%20Information%20Technology%20Hyderabad/*.

[4] Sahaf, Zahra et al. 2014. VAST Challenge 2014 Mini Challenge 2. *http://visualdata.wustl.edu/varepository/VAST%20Challenge%202014/challenges/MC2%20-%20Patterns%20of%20Life%20Analysis/entries/University%20of%20Calgary/*.

[5] Zhuang, Cai; Mengyao Chen;, Hanqing Zhao; Ying Zhao; Fangfang Zhou; Kang Zhang; 2014. VAST Challenge 2014 Mini Challenge 1. *https://www.cs.umd.edu/hcil/varepository/VAST%20Challenge%202014/challenges/MC1%20-%20Disappearance%20at%20GASTech/entries/Tianjin%20University%20-%20Cai/*.