Group 1 ShinyVA - A Shimy Application for Crime Detection

Li Nan
Singapore Management University
nan.li.2020@mitb.smu.edu.sg

Li Yueting
Singapore Management University
ytli.2020@mitb.smu.edu.sg

ABSTRACT
This Paper described the work of analysing Mini Challenge 1 and Mini Challenge 2 in VAST Challenge 2021 and the related shiny app created.

1. INTRODUCTION
This paper is based on analytics for the Mini-Challenge 1 & Mini-Challenge 2 of the VAST Challenge 2021.
In a fiction scenario,a gas-production company named Tethys-based GAStech has been operating a natural gas production site in the island country of Kronos and it has produced remarkable profits and developed strong relationships with the government of Kronos. However, GAStech has not been as successful in demonstrating environmental stewardship. And in January, 2014, the leaders of GAStech are celebrating their new-found fortune, but in the midst of this celebration, several employees of GAStech went missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, therefore analysis are conducted and an application is created related to the analysis. For this fiction scenario, three Mini-challenges are provided by VAST Challenge 2021 and the focus of this paper will be on mini-challenge1 and mini-challenge2.

i) Mini-challenge1:

Mini-challenge1 provides mainly three kinds of data, News report details data related to the kidnapping events from various of Medias, and the email exchanges data of the company GAStech's employees. The data also consist of each employees' detailed information which enable analysts to do further analytics. A community detection can be processed in order to distinguish the relationship between employees. Moreover, time stamps of news as well as the email transctions can be denoted using these data.

ii) Mini-challenge2

Mini-challenge2 provides 4 datasets to data analysts for exploration. These datasets describe information about Transaction, GPS and Car Assignment. Through data manipulation and visualization, it can be detected anomalies that appear in unmatched transaction records between loyalty card and debit/credit card and GPS data also shows the unfamiliar movement for some employees. Through the background stated, this paper develops interactive visualization approaches to provide evidence and suspicious behaviors of GAStech employees.

The introduction of this paper is followed by an explanation of our motivation and objectives in Section 2, then followed by Section 3 which details the data used and methodology selected. And Section 4 provides a visual overview of the final application and finally section 5 provides the conclusion and insights for this paper.

2. MOTIVATION AND OBJECTIVES
2.1 MOTIVATION
This project was motivated by a desire to: i) identify the complex relationships among all of these people and organizations.
ii) track data for the two weeks leading up to the disappearance, as well as credit card transactions and loyalty card usage data.

2.2 OBJECTIVES
In order to reach the goal, interactive tools are developed to addresses the following requirements:

1)What is the frequency pattern pattern in news reported that related to the event? The answer is to visualize the pattern a heatmap can be adopted in order to see the timeline and when the event draws high media attention.

2)What is the network between employees? is there any suspicious connections between them? A network graph will be presented to demonstrate the connections between employees. Also, in order to find out the most frequent email connections, the number of email engagement will be treated as the weight of the graph, and by altering the weight people may clearly realize the level of connections of communities in GAStech company.

3)What is the suspicious pattern in company cars for employees' personal and professional use? To utilize geographical visualization techniques to create car/truck routes of

GAStech employees based on GPS dataset and also label parking points of car (parking more than 5 minutes) to dig out suspicious movement patterns.

4)How to utilize GPS data to match card owners and the debit/credit card owners? To create interactive transaction data table to track unusual credit/debit card transactions records of GAStech employees and compare them with car parking points shown in visualization map.

3. METHODOLOGY

The paper utilize data in three aspects:

i) Find relationships between different data tables and try to use different data manipulation methods (e.g. group and filter data) to deepen understandings of data story.
ii) Based on the understanding of data, find explicit and effective data visualization methods to deal with varied data types (e.g. data table – currency data; spatial data – geospatial visualization).
iii) Point out and try to give reasonable explanations and insights of data anomalies based on data visualization tools. To enable readers to have deeper understanding about data manipulation in this project, an explanation of the different methods used follows.

3.1 Data Manipulation Flow

The raw data manipulation is followed in the following 4 steps by using tidyverse package in R (shown in Figure 1):

i) Step 1 is to read varied dataset (e.g. csv file, spatial data set and rds file) by using readr package.
ii) Step 2 is focus on data preparation which was conducted in R using base R and the dplyr package to narrow the scope of the data by summarizing data (e.g. extract transaction data by grouping location and day), convert data type (e.g. character type to factor type), and fix garbled character (e.g. Karterina's Cafe).
iii) Step 3 is to use different visualization packages to create visualization graphs and tools. In this project, tmap is used for geospatial visualization and shiny is intended for interactive tools.
iv) Step 4 is to write reproducible report by using r markdown and create poster by using posterdown package.

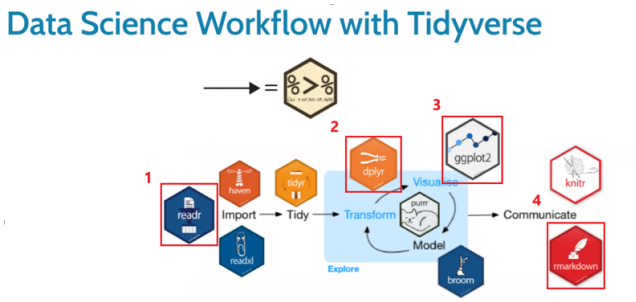

Figure 1 - Workflow with Tidyverse

3.2 Shiny Architecture

Shiny is an R package that makes it easy to build interactive web apps straight from R. Users can host standalone apps on a webpage or embed them in R Markdown documents or build dashboards. The reason why choosing Shiny as interactive tool is as follows:

i) Input values can be changed by the user at any time, through interaction with customizable widgets.
ii) Output values and graphs react to changes in inputs, with the resulting outputs being reflected immediately.
iii) Table and graph can be interactive by setting logic in server code. The output values react to changes in inputs, with the resulting outputs being reflected immediately.

The design of the Shiny web app flowed from the key business questions of the data:

i) First tab- Network Analysis to form email relationships between GAStech employees.
ii) Second tab - Heat map visualization to help user find the news details and the time variation of event heat
iii) Third tab - Geospatial visualization and interactive data table help the user understand and match the relationship between card owners and car drivers and also provide insights of unusual movements of car drivers to detect suspicious persons.
iv) Forth tab - backup table for network visualization

To give better understandings and guidance from users, we included a user guide to accompany each analysis technique featured in the app.

3.3 Analysis Techniques

3.3.1 Data Table

To create an interactive transaction data table, DT package in R is used. The R package DT provides an R interface to the JavaScript library DataTables. In this paper, loyalty dataset and credit/debit card dataset used by DT package is displayed as tables on HTML pages, and through argument settings, this data table provides filtering and sorting features.

Transaction Data

last4ccnum	loyaltynum	location	price	Day	hour
All	All	All	All	All	All
1286	L3572/L3288	Brew've Been Served	14.97	06	08
1286	L3572/L3288	Abila Zacharo	50.14	06	13
1286	L3572/L3288	Brew've Been Served	11.92	07	07
1286	L3572/L3288	Kalami Kafenion	45.05	07	13
1286	L3572/L3288	Brew've Been Served	7.26	08	08

Showing 1 to 5 of 1,490 entries Previous 1 2 3 4 5 … 298 Next

Figure 2 - DataTables Overview

3.3.2 Route Map

Thematic maps are geographical maps in which spatial data distributions are visualized. This package tmap offers a flexible, layer-based, and easy to use approach to create thematic maps, such as route map. Shiny package takes a fresh, interactive approach to telling the data story Another package used in this map creation, sf, supports for simple features, a standardized way to encode spatial vector data.

Figure 3 - Route Map Overview

3.3.3 News heatmap

A heatmap (or heat map) is efficient way to visualize hierarchical clustering. It allow us to simultaneously visualize clusters of news for different time and for different media. A color scheme is applied for the blocks to display 'high' and 'low' frequency values. Package ggplot2 is efficient in presenting this visualization.

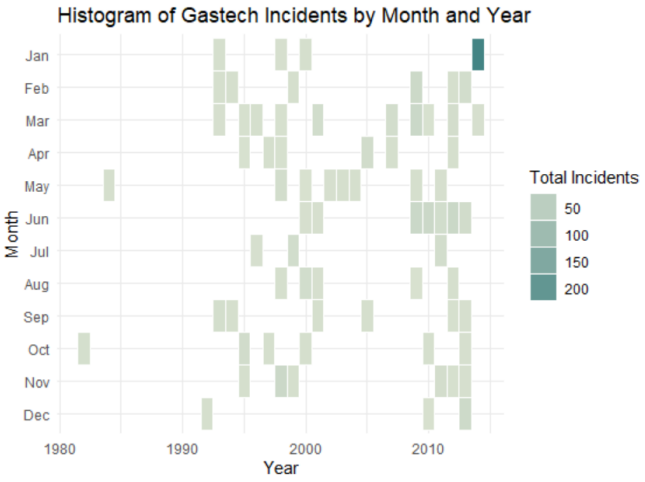

Figure 4 - Heat Map Overview

3.3.4 Email Network

Network is a group of people or organizations that are closely connected and that work with each other. The email connections between each employee of the GAStech formed a natural network. Each of the employee is a node, the email connections between them are the edges, and the frequency of their connect can be represented as the weight of the edges. The network visualization can be achieved using igraph, tidygraph and ggraph.

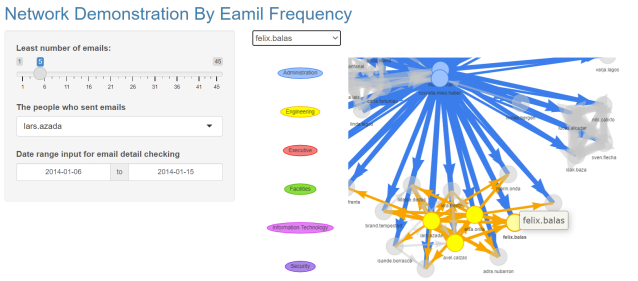

Figure 5 - Network Visualization Overview

4. DELIVERATES AND FINDINGS

The final interactive graphs in the published web application, and an illustration of the potentially wide range of insights that can be gleaned from the application are briefly described below. ## What info can gain from Interactive route map 1) Residence: Here we can identify the living place by changing the input of "Tracking Day," and we can see every day Car ID2(Azada_Las) will appear in this location, so we can identify its living place is here

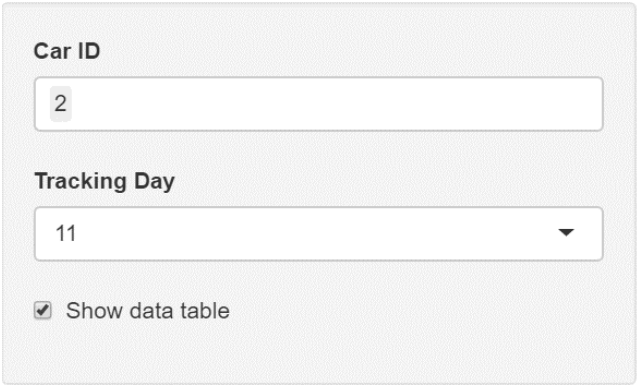

Figure 6 - ID2_Day11_Starting point

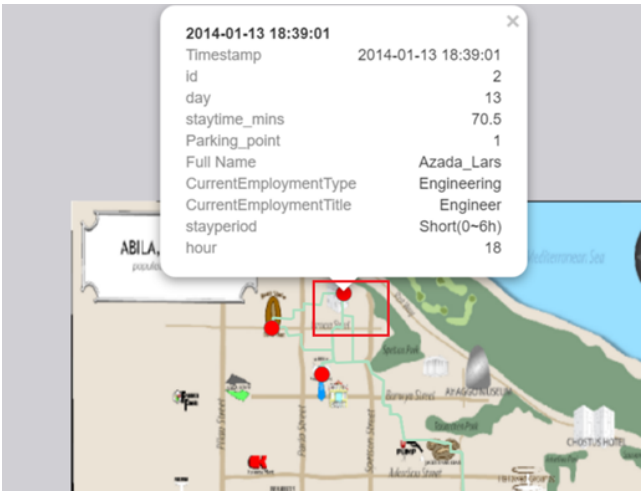

Figure 7 - ID2_Day13

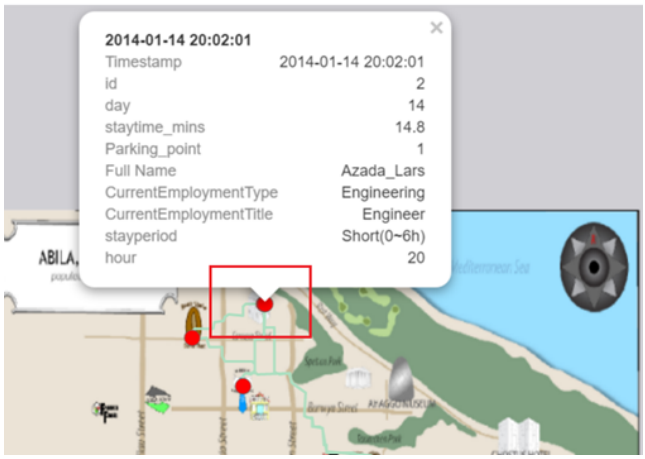

Figure 8 - ID2_Day14

2) Match credit card owner and car owner

Through the stopping points which show info about day and hour and the corresponding transaction record, users can easily see the relationships between car owners and credit card owner, which provides strong support for future detecting work for unusual transaction data.

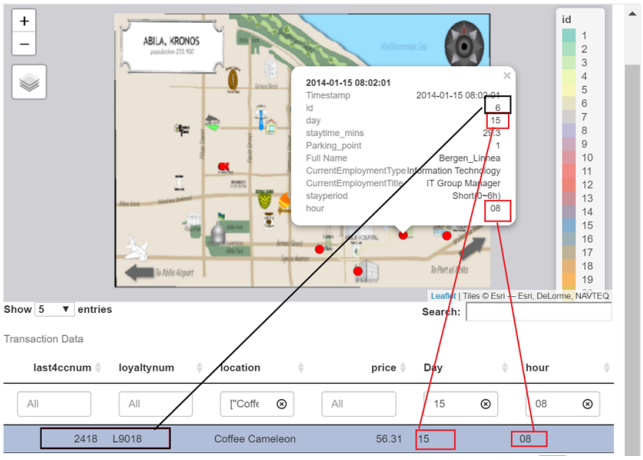

Figure 9 - Car Owner Matching

3) Can detect people who gather together in one day

If we finally notice some suspicious people we can see their behaviors by clicking multiple Car IDs, we can detect in one day people who gather together in one place.

Figure 10 - Select multiple IDs

4.1 What info can gain from Interactive network

1) Can set the threshold of least number/frequency of emails

For any intended investigation time period, users may select by manually to see how the network changing.

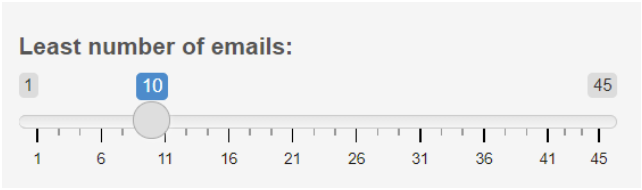

Figure 11 - Choosing email frequency threshold

2) Can specify the people of interest and their network

For any person of interest, users may select their id manually to see the email correlation between this person and his/her colleagues.

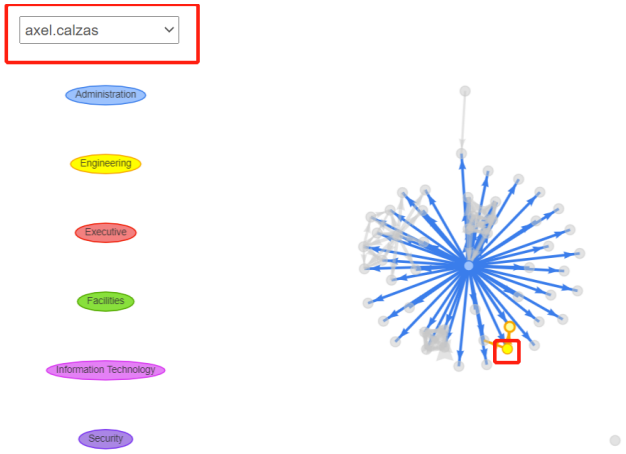

Figure 11 - ID selection for network visualization

3) Can detect people's email activity within self denoted time range

For any intended investigation time period, users may select by manually to see how the network changing.

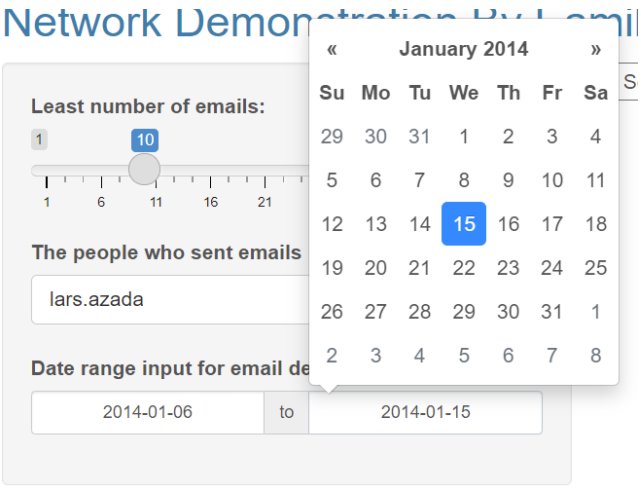

Figure 12 - Time range selection

5. CONCLUSION AND FUTURE IMPROVEMENT

This paper set out the development of a web application targeted at providing an interactive visualization tool to help minimize time to have a thorough understanding of person's social connection and background (e.g. car owner/credit card owner/daily movement/daily spend) and detect suspicious patterns.

Also By observing the timeline, it helps people quickly locate the time point of the incident for interactive comparison. And the Network visualization demonstrate the personal & official connections between people, which enable analysts to quickly understand a person's social network and identify possible suspects and insiders.

However, shiny app also have some demerits that needs to be improved in the future.

For the route map app:

- Consider add "select all" action button. This will help to detect each day all the driver's movement.
- Instead of using Car ID as input, consider using Employee title and Employee department as input, which can help to find relationship between different departments.
- Consider Add starting time and end time in the tooltip, which can be more helpful for detecting suspicious gathering groups.

For the Network Visualization:

- There are only nodes that represent GAStech employees in network, however, there are more than emails within GAStech company provided in the dataset, it might be useful to also include nodes outside the company for a vaster social connection evaluation
- In order to investigate the time distribution of email sending, an comparison graph that gives the email activity related to weekday might be helpful

6. ACKNOWLEDGMENTS

The authors thank Associate Professor, KAM Tin Seong, Singapore Management University for his support and guidance, Classmate Liu Yang Guang for giving advice and suggestions for data visualization and Classmate Syed Ahmad Zaki whose assignment expand ideas and logics.

Memo: If you want to explore the application, please click *Visual Analytics App*