1. ***Description of your software solution***

***1) How do you process the given dataset?***

I used both python and d3js for processing the dataset. Python was used to create matrices that had data related to the frequency of occurrences among any two given variables such as Person, Organization, etc., These matrices were stored as comma separated files (csv) and later were uploaded on github. The libraries used in python include numpy, pandas, itertools, json libraries.

I used an online website to convert the given xml data to a json string and stored it in a variable in the file data.js. This json data is directly imported into the project and processed using d3js.

To display the documents on the webpage, I have used the wrap function for text by following a thread on stack overflow. This function was entirely written by [Mike Bostock](https://stackoverflow.com/questions/24784302/wrapping-text-in-d3/24785497). Secondly, I developed the chord diagram by referring to the examples on the d3js website’s [gallery](https://www.d3-graph-gallery.com/chord). However, I have only taken the template and the rest of the implementation is my own.

***2) Which visualizations are included in your software?***

Two visualizations were used in this project, first is a matrix visualization and the second is a chord diagram.

***i. For each visualization, what marks and channels are used?***

1. Matrix - marks : squares, channels : color.
2. Chord - marks : arcs, channels : color and width.

***ii. Why do you choose to use these marks and channels?***

1. Matrix : In the matrix representation, squares were easy to understand over other shapes such as circles or triangles. To encode the frequency of occurrence of the intersection of each node color was used as a channel. This color ranges from pink to blue where blue represents the highest value. Although the color is not the best channel to encode a quantitative measure, it seemed suitable for this visualization. To show the frequency more clearly, I attempted to show the number on top of each square, but it was hard to interpret so it was removed.
2. Chord diagram : I used a basic chord diagram therefore I used arcs as the markers. To encode each node, a different color is used and the quantitative relation between each node is represented by the radius of the arc. This chord diagram has a limit of showing five nodes. The purpose of the diagram was to show connections between few suspected nodes and an analysis of many nodes was not required to solve this problem so I allocated space only to show fewer nodes at a time.

**3) Instructions on running your software**

* Initially the software starts with a blank screen and few buttons. Among these buttons, the first two provide options to select an attribute to represent along the row and columns. Then the draw button should be clicked to create the matrix.
* Once the matrix is visible on the screen, the user may start the exploration by selecting different attributes to be represented among rows and columns and redrawing the matrix.
* The matrix forms heterogeneous relationships among two different attributes. To see the homogeneous relationships, the chord diagram can be used.
* To view the chord diagram, the user has to click on 2-5 squares on the matrix and click the make chord button.
* Clicking on make chord button would then create two chord diagrams.
* The left and right chord diagrams represent the nodes on rows and columns respectively. This will give some room to see if there are any associations among the few suspected nodes.
* At any time, the user may choose to read the reports associated with nodes shown on the rows and columns of the matrix. This can be easily done by hovering over the names. This will then show the document names on top of the matrix.
* To read any report, the user has to click on the desired document and the text corresponding to the document will be shown on the top of the screen.

***b) Your findings for the given dataset***

***1) What are your findings?***

I have explained in the report finding section.

***2) How does your software help you to get these findings?***

The matrix definitely provides an overview of the connections. It mostly encodes the frequency of occurrences between nodes of two different attributes. Sometimes this could be useful to identify the most active suspect but a user has to make sense and navigate based on the reports. In my findings, I saw that even a person with lower frequency was a suspect, perhaps it could also be the reason that various alias names were used. The chord diagram can be used to quikly explore the relations between few homogeneous nodes.

***i. For each reported finding, you need to describe the process that you take to get it.***

1. Select Person as an attribute for rows and Organization for columns. Here first five nodes along the rows are selected and the draw chord function is used. This shows that Muktar Galab and Sheila Watson have many connections.
2. Upon hovering over Sheila Watson and reading her reports shows that she is not a suspect and she only made a mistake in mentioning her address.
3. Then hovering over Muktar Galab and read few reports. This node showed many connections with other people as even seen in reports. (Abdul Ramazi, Faysal Goba, Hans Pakes, Yasein Mosed).
4. I drew a chord diagram for all the above names and it shows some association Yasein Mosed and Faysal Goba, where both of them went to a University. Also first class tickets were booked for these people who didn’t even attend any classes in their registered University. This seems suspicious.
5. Also checking on Hans Pakes shows that it is an alias name used by Bhagawat Dhaliwal who worked in some explosives department.
6. As it is an alias name, I checked Bhagawat Dhaliwal’s record and found that he is associated with another person with an alias name. Bhagawat Dhaliwal was also an alias to the name Sahim Albakri.
7. After reading a few reports, it can be seen that Sahim Alkbari made deposits of $13000 to the students and orchestrated the train attack.
8. There were numerous mentions about different locations, but this was one of the quickest routes that I have taken in solving the problem. I have plotted matrix b using location and phone numbers as well but this seemed more straightforward.

***ii. You need to provide detailed evidence from your software (e.g., screenshots of your software, or screen recordings to demo the usage of your software) that has been captured in your analysis process to explain how your software is used in your analysis process.***

Figures for each step are shown below. Figure number corresponds to the step number.

Figure1)

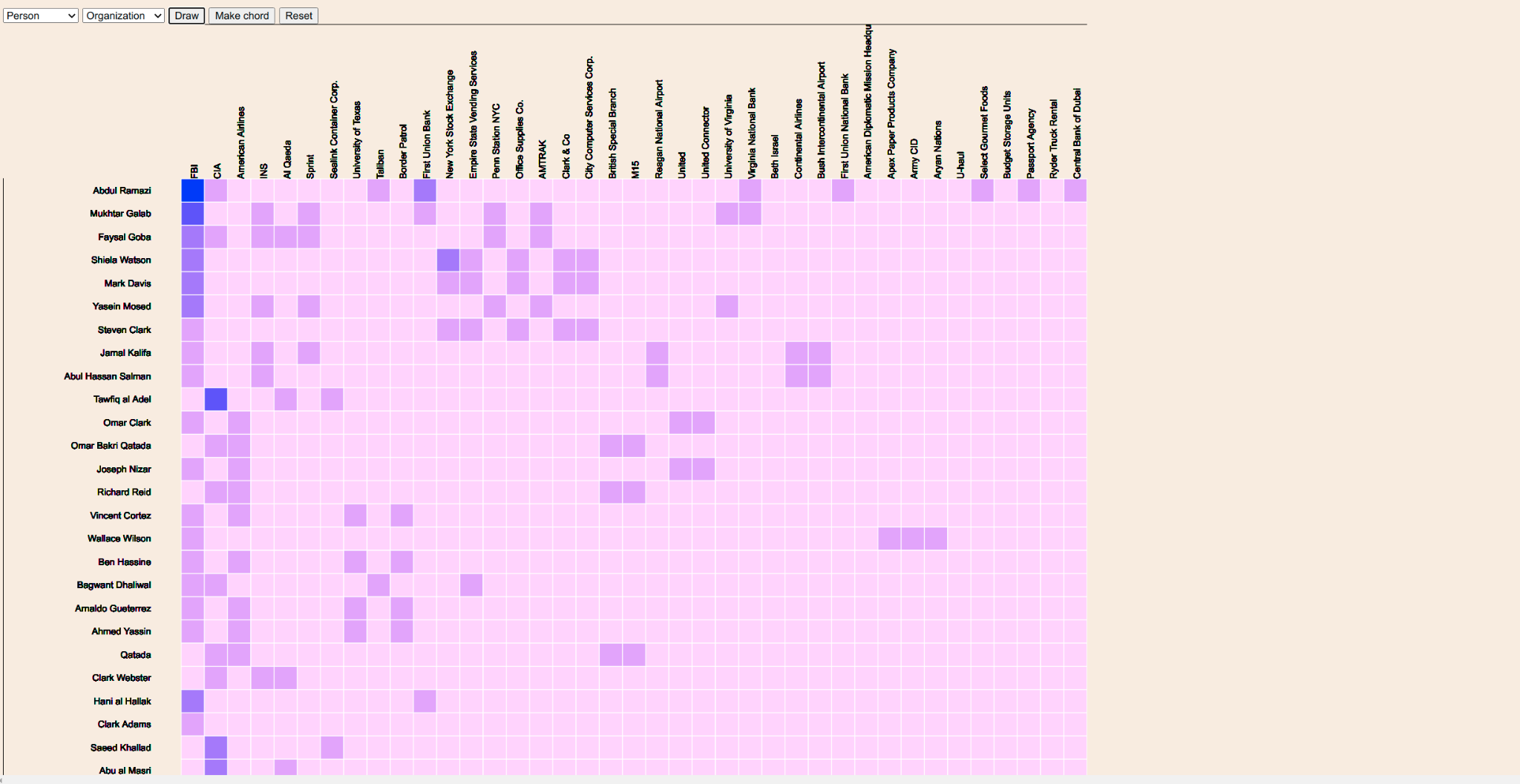


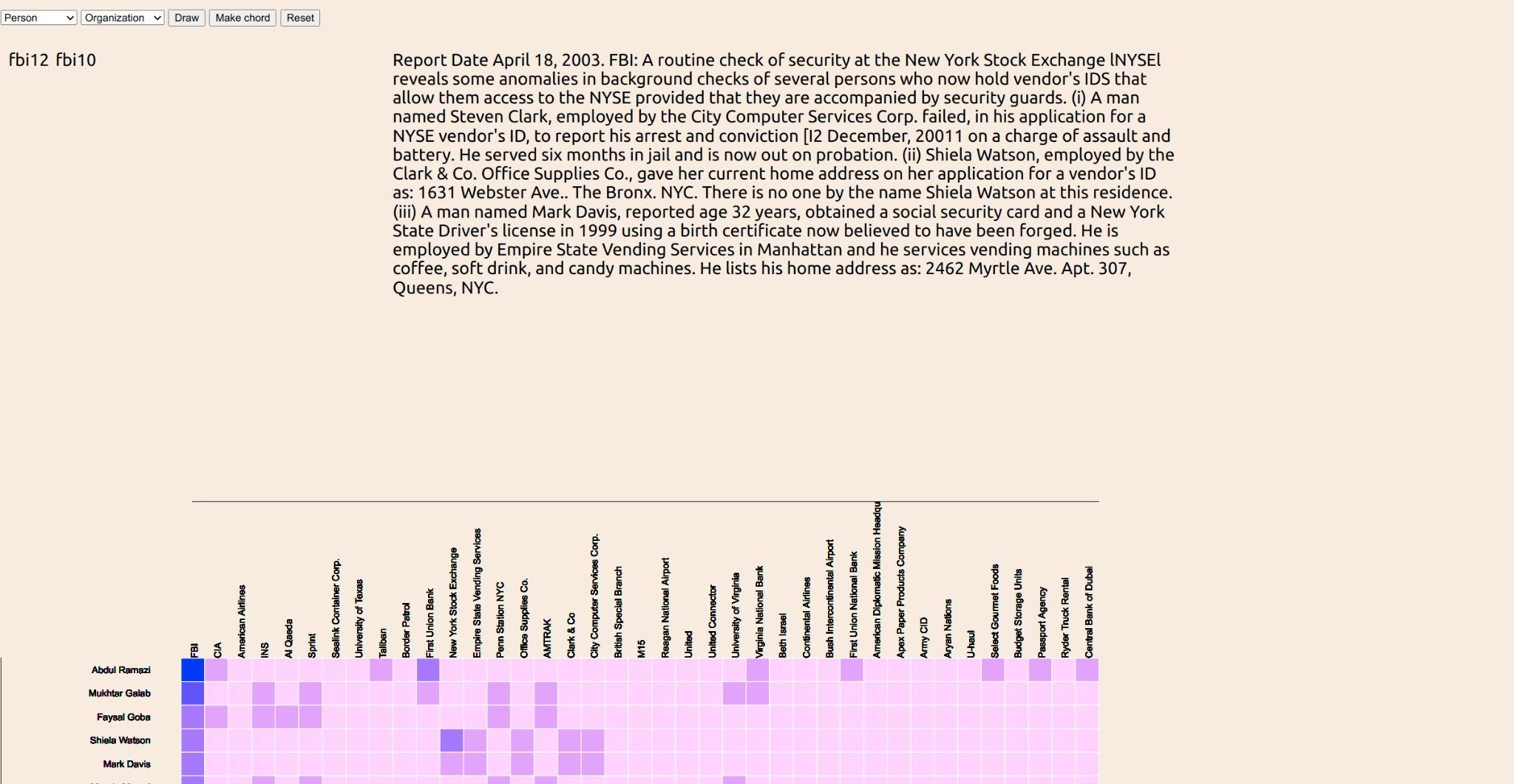
Figure2)

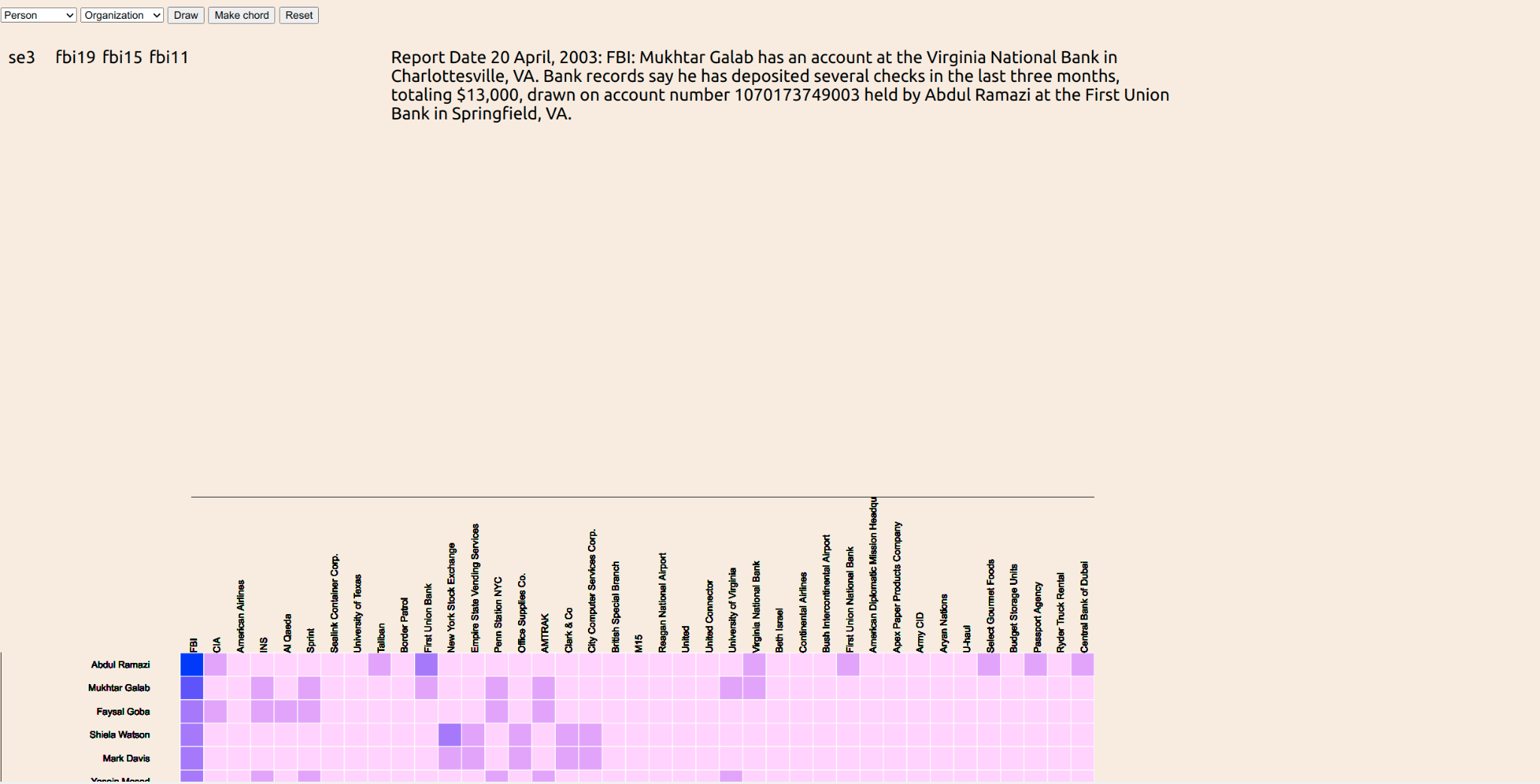
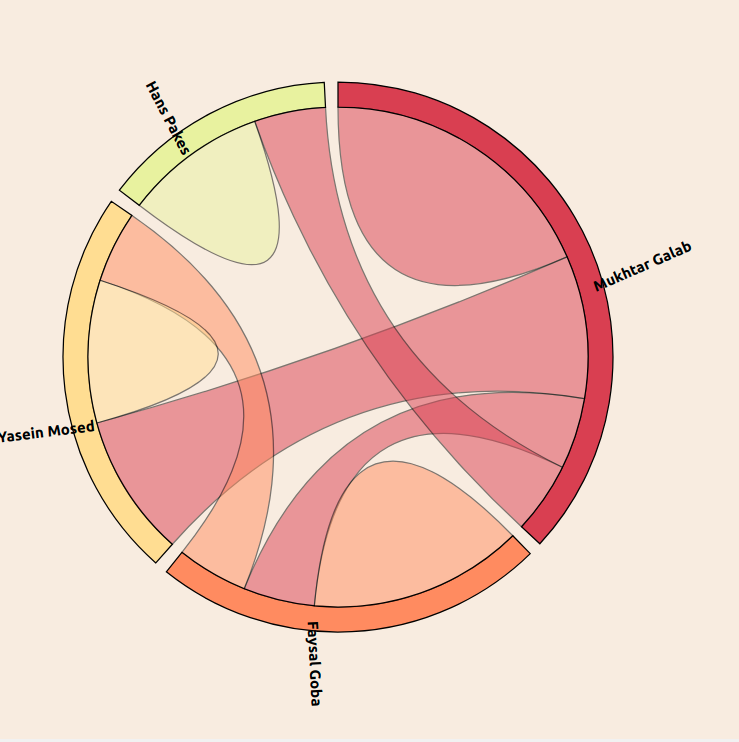
Figure3)

Figure4)

The visualizations in Figure 5,6,7 are similar to above figures and the description is given in the steps taken.