Project Report – 2008 VAST challenge

Abhishek Singhal

2008 VAST challenge:

- 4 mini challenge
- 1 grand challenge

Attempted:

- Mini Challenge 1: Wikipedia Edits
- Mini Challenge 2: Boat migration records
- Mini Challenge 3: Call records

A. Mini Challenge 1

Problem:

The Paraiso Wikipedia entry has become another focal point for conflict between Catalano's followers and the general population. We would like to use visual analytics to understand the social relationships and dynamics reflected in the tug of war over this Wikipedia entry.

Approach:

Part 1.Data Transformation

- Some data cleaning was done before it could be used. All the records did not contain all the columns. There were missing values. They were not removed, but NaN was placed. It took considerable time to order data into Pandas DataFrame columns.

Finally, the data was put in columns.

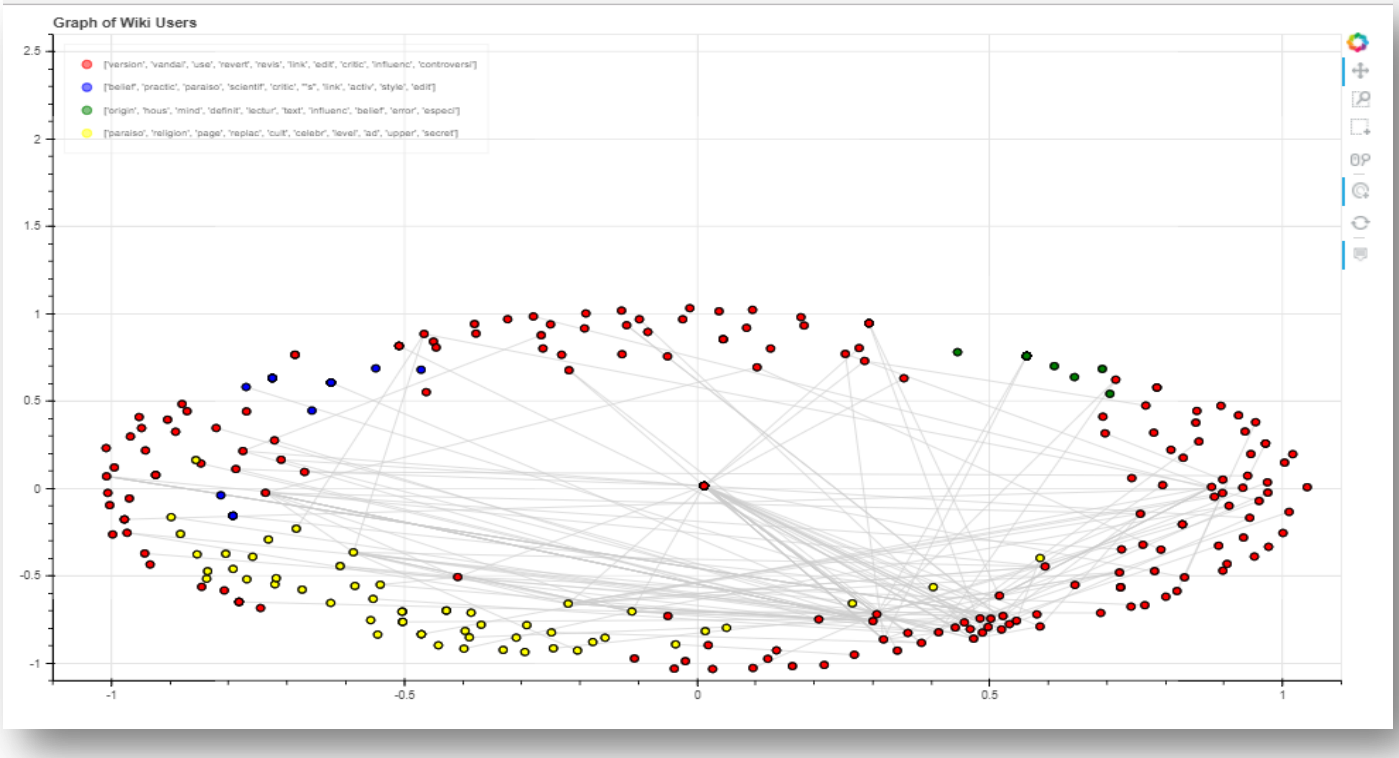
Col	User	Date/Time	Edit Length	Minor Edit	Comments
Dtype	String	Datetime	Numeric	Bool	String

There were total 1009 entries.

- Further, the data was grouped by User. This allowed all the comments clubbed together for analysis. In total, there were 387 total Users.
- Next, the comments were tokenized (nltk.word_tokenize) and Term Frequency-Inverse Document Frequency (TF-IDF) was made.

Part 2. Visualizations

2A) Network Graph



- Nodes: The nodes in the graph were Users. The high dimensional TF-IDF matrix was used to represent features of a user. Hence dimension reduction techniques were applied to get maximum data in two coordinates, 'xs' and 'ys'. We can select between MDS, PCA, and SVD algorithms.
- Edges: Each edge represented use of another Username in one's comment.

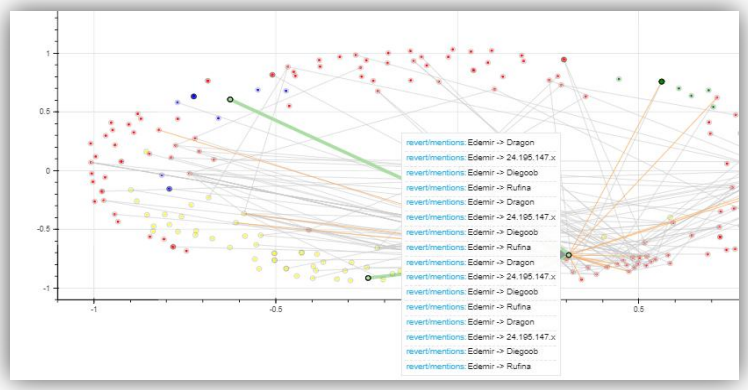
```
timestamp                                  2007-01-13 14:44:00
user                                                   Gustava
minorEdit                                                False
pageLength                                              100959
comment       Undid revision xxxxxxxxx User:Moisescorral Und...
entireEdit    # (cur) (last) 14:44, 13 January 2007 Gustava ...
Name: 5, dtype: object
From Gustava ->To Moisescorral
```

Edge

- Cluster: The most important part was to cluster the Users TF-IDF matrix, and get the grouping. Then the edges will help to analyse further. KMeans was used and the number of clusters specified were 4. This is because there were 4 grouping in previous HWs on another dataset, and this was logical. The nodes were appropriately colored according to cluster.
    - For each cluster centre, main set of feature were also obtained.

```
Red['belief', 'use', 'vandal', 'revert', 'link', 'revis', 'edit', 'practic', 'critic', 'controversi']
Blue['paraiso', 'religion', 'page', 'replac', 'critic', 'belief', 'cult', 'scientif', 'celebr', "'s"]
Green['origin', 'influenc', 'hous', 'mind', 'definit', 'lectur', 'text', 'belief', 'error', 'especi']
Yellow['version', 'agustin', 'rv', 'socorro', 'sara', 'hamsterlopithecus', 'rosalind', 'snakey', 'honoria', 'hispa'
```

- Observations from Graph:
    1. Edimir was user with most edits. But it cannot be judged whose side he is on.

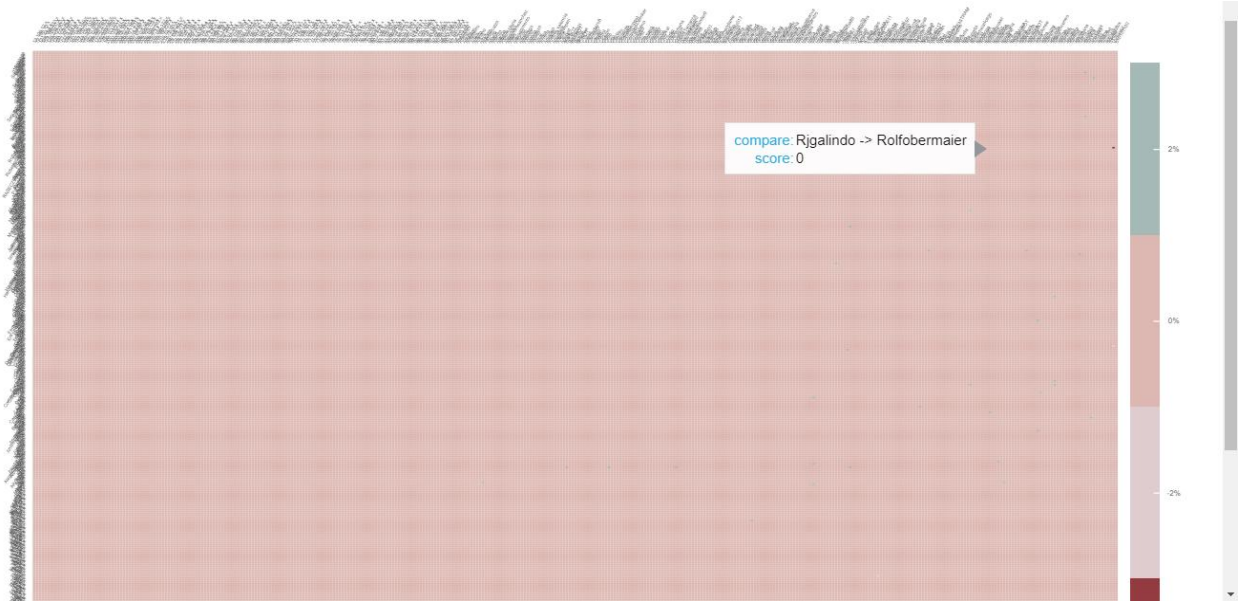


2. VictoriaV, RyogaNica, DailosTamanca, Rm99, Sara, Edemir, Agustin, Amado stand out among different groups.

3. Sara and VictoriaV have edited/revertes DailosTamanca. So possibly they are together. There is no other user in graph which have common reverts.
4. VictoriaV is clearly opposite to Rm99 and Augustin. Sara has edges to Rm99. This possibly confirm hypothesis in above statement that VictoriaV and Sara are together.

2B) Heat Map

- A 'relationshipmatrix' (397,397) was created . It gave a score (positive or negative) based on how a user mentions another user in his/her comment.
- Keywords that were used to check sentiment were 'mention' and 'revert'.
- Based on this matrix, a heat map was visualized.
- Observations:
  1) I got a very good heat map in HW6 (another Wiki-edit dataset). But on this dataset , it did not come good.
  2) The users were far too many, and most of the matrix was sparse.



Part3. Manual Analysis

- The graph gave a very good picture about what is going on.
- We have singled out 7-8 user based on their edits and their relationships with other users.
- There was a further need to re-iterate what the graph was suggesting, and gathering concrete evidence about 'wiki-war' going on.

```
print ((userWiseComments[userWiseComments['user']=='Edemir']['allComments'].tolist()))
```

```
['Reverted good faith edits by Dragon; No need to dampen it; the reference points to a blanket opinion of psychiatry and psycho
logy.. using TW  - Counselor doesn\'t need quotes here. Reverted 1 edit by 24.195.147.x identified as vandalism to last revisio
n by Diegoob. using TW  Paraiso as a cult - More cleanup, more typos, fixed refs, etc. Texts and Lectures - More vandalism clea
nup, and moved a period Origin - Trimming out more vandalism .  - Removing vandalism Minor language - Good job on that pernicio
us sentence, Amado That darn sentence... I\'m not even sure the intro should be this long. Restored language that actually does
reflect the contents of the refs, and removed a ref that didn\'t apply.  Neutral, factually accurate. Says nothing more or less
than that the State Department has such reports, and that Paraisos report descrimination. Previous phrasing is more neutral, bu
t the ref is fine. Introduction - Minor language Reverted 2 edits by Alverio identified as vandalism to last revision by Rufin
a. using TW Reverted 1 edit by Alberto identified as vandalism to last revision by Sara. using TW Reverted good faith edits by
Callas; That more properly belongs in the Angel article, where it is already.. using TW Actually, that\'s not a typo, it\'s rea
lly spelled that way: Deseret Membership - Being bold, removing inaccurate image map per talk page. Influences - Language clean
up, references cleanup  - Undoing because the added ref isn\'t a RS, and the new language wasn\'t grammatical. Training - Clean
ed up grammar and syntax, rephrased for readability, replaced "smart" punctuation, formatted refs, etc... Undid revision xxxxxx
xxx - VictoriaV, it\'s still not relevant, and we\'ve no reason to doubt the source. See talk page. . using [[W Membership - Sp
otted another one. Membership - Cleaned up some typos around refs  The rest is superfluous, but the actual number of interviewe
es is fine -- and more importantly, in the ref. Membership - That\'s OR, as the ref doesn\'t mention that percentage. Besides,
that\'s a perfectly valid sample size for statistical analysis. Reverted 1 edit by 207.61.57.x identified as vandalism to last
revision by Seina. using TW Reverted 1 edit by 209.250.162.x identified as vandalism to last revision by Edemir. using TW Rever
ted to revision xxxxxxxxx by Niermague; Restoring unvandalized version - Anoryat removed punctuation unnecessarily. using TW Re
verted 1 edit by 75.131.224.x identified as vandalism to last revision by Alvaro. using TW Reverted 1 edit by 58.179.241.x iden
tified as vandalism to last revision by BakBOT. using TW Beliefs - Tone reads like an advertisement for the church in some plac
es. Removed extraneous cruft that got re-added in an anti-vandalism revert Typos Paraiso as a state-recognized religion - If re
ferences for these statements aren\'t found post-haste, this paragraph goes. Reverted 1 edit by Oscar identified as vandalism t
o last revision by ErKomandante. using TW Home Health Care - added confrontation of Paraiso members and Dept of Health Home Hea
lth Care - Dept of Health intervention refs Home schooling for girls - Whoops, messed up the punctuation. Sorry.  Home schooli
```

- Edemir has often mentioned VictoriaV, Sara, Amado in negative sense. (Notable edit: addition Scientific criticism of Paraiso\'s beliefs)
- Amado has often reverted Rm99. (Notable edit: there are also Journalists, courts and the governing bodies that are not critical of Paraiso)
- VictoriaV viewed Paraiso as a state-recognized religion.

Solution to MC1:

Pro-Paraiso	Anti-Paraiso
VictoriaV	Rm99
Sara	Augustin
Amado	Edimir

- **Things to be noted** While arriving at solution, visualization was used as a tool to help. And not give final answer. Although clustering gave a very good idea about groups and faction, an assumption was marked while making graph. This is, that all mentions/reverts in the comments by user of another user are in opposite/negative sense. It helped chalk out and visualise the edges. On further inspecting, could I finally conclude the possibilities.
- **The answer arrived was already given but visualization was a bit different. Had the answer been not known, most of the people might have been grouped correctly by Visual Inspection only. Only confusion was Edemir.**

B. Mini challenge 2 (Boat Migration)

Problem:

This data records information collected by the Coast Guard concerning the mass movement of Isla Del Sueño persons departing the island for the United States during 2005 - 2007. This activity was precipitated by the crackdown of the island nation’s government on the Paraiso social and religious movement gaining popularity there.

We would like to use visual analytics to help us understand more about the migration during these years.

Approach:

Part1. Data:

- Most of the data here was cleaned, although there were considerable records missing the launch sites.
- Xml was converted to pandas DataFrame.

Col	Data type
EncounterCoords	String
EncounterDate	Datetime
LaunchCoords	String
NumDeaths	Numeric
Passengers	Numeric
RecordsNotes	String
RecordType	String
USCG_Vessel	String
VesselType	String
Encounterx	Numeric
Encountery	Numeric
Launchx	Numeric
Launchy	Numeric

Part2. Approach

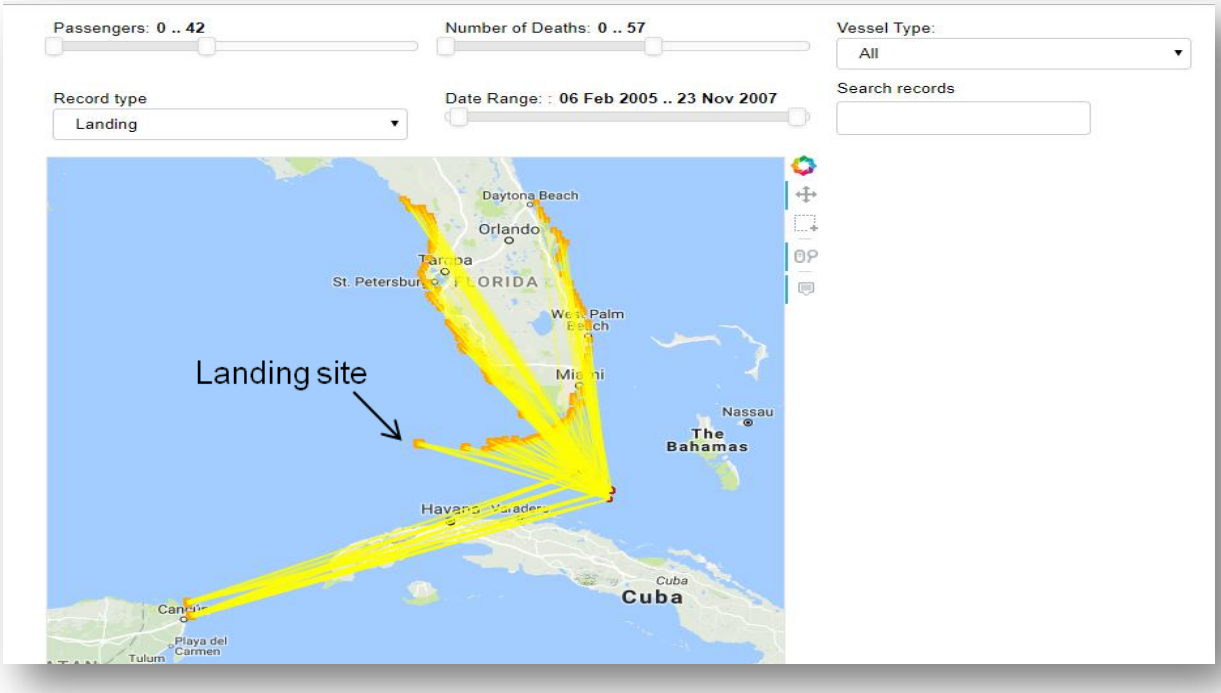
- The approach was completely visualization+analytics. No machine leaning was applied.

Visualization:

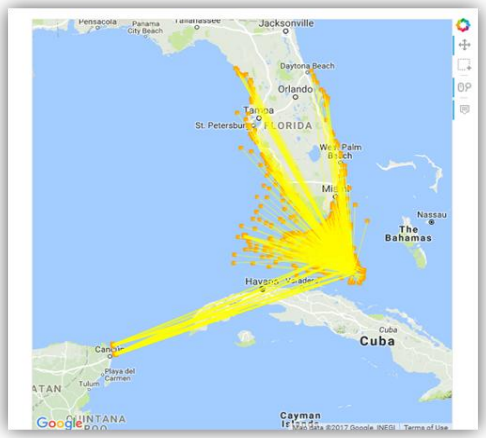
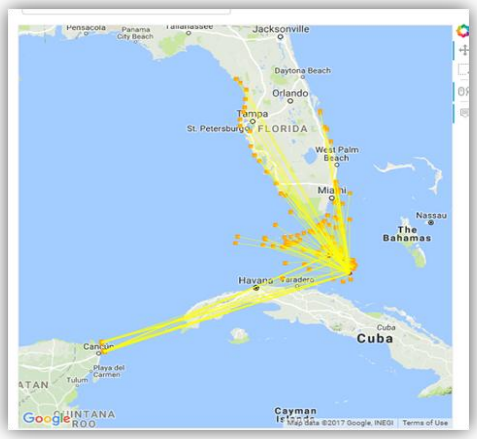
Migration graph was superimposed on Google maps using GMapPlot library in bokeh.

The dashboard had many interaction features to filter the records, observe the trends and find the solution.

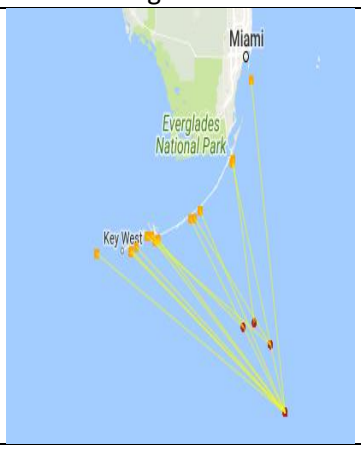
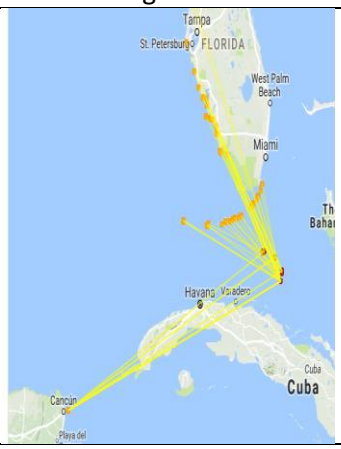
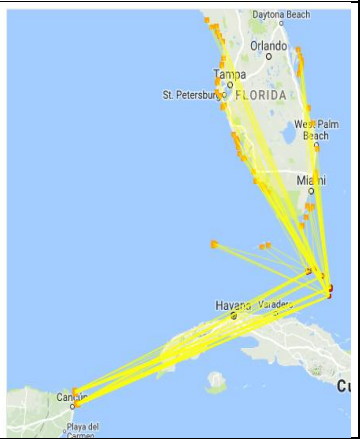
- Number of passenger
- Number of deaths
- Date of Encounter
- Search names through records
- Type of Record (Landing/Interdiction)
- Vessel Type



- Observations:
  - The number of deaths involved in migration is considerable. (1143 records reported one or more death, 7229 records reported no death). We can say that migration was not safe on these boats.

No deaths	1 or more deaths
	

  - The landing pattern changed over 3 years (2005,2006,2007). Mostly heavy landing occur in summer months (June-September). Very little landing attempt was seen during months of December, January.
    In 2005, migrants directly went to south Florida (shortest route).
    In 2006, in addition to previous route, migrants went farther away (west coast Florida) till St. Petersburg even.
    In 2007, they tried on East coast Florida and Mexico.
    (These pictures do not include Interdiction)

2005 Landings	2006 Landings	2007 Landings
		

  - Searching the record notes for Vidro/Catalano family had 3 records. (2 interdictions, 1 landing in Mexico).
  - Interdiction was always close to US coast.

C) Mini Challenge 3 (Cell phone records)

Problem:

These cell phone call records cover a ten day period in June 2006 and should give us some idea about the Catalano social network. We were able to narrow the dataset down to about 400 unique cell phones during this period, and we would like to use visual analytics approaches to help us understand this data. . We have medium confidence that Ferdinando Catalano is identifier 200. We believe Ferdinando would call brother Estaban most frequently. We also believe that David Vidro coordinates high level Paraiso activities.

Approach:

Part1. Data

Data was already cleaned.

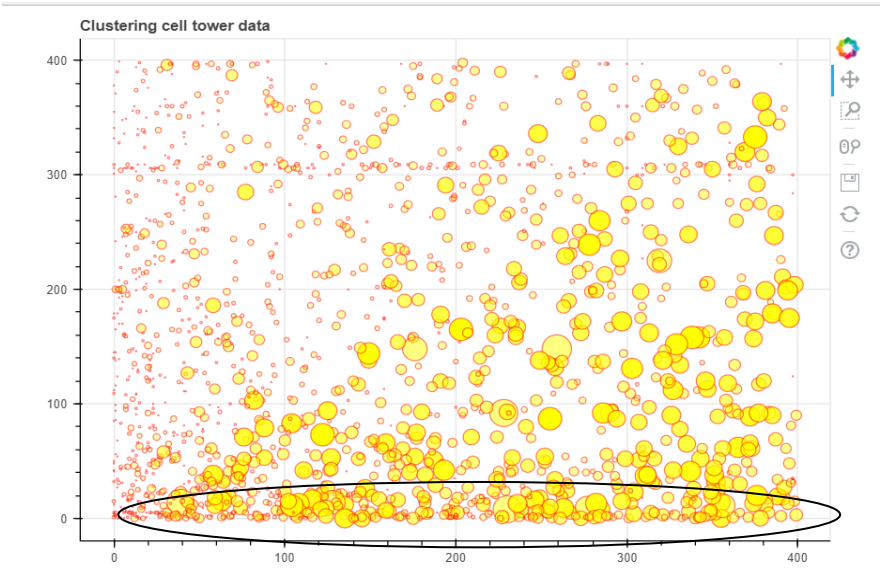
Another dataset was created for all the callers connected to ID 200.

Part2. Approach

Clustering was attempted on individual records keeping IDs,tower and duration data. No significant analysis came out. This idea was dropped.

Visualisation1: Scatter plot

- After clustering, scatter plot was attempted to see overall picture. The size of scatter points and the density of color varied according to duration.



- Observations:
  There is too much clutter in near ID 0. He/she is involved in too many calls.
  Also it gave a look at calling pattern of ID 200.

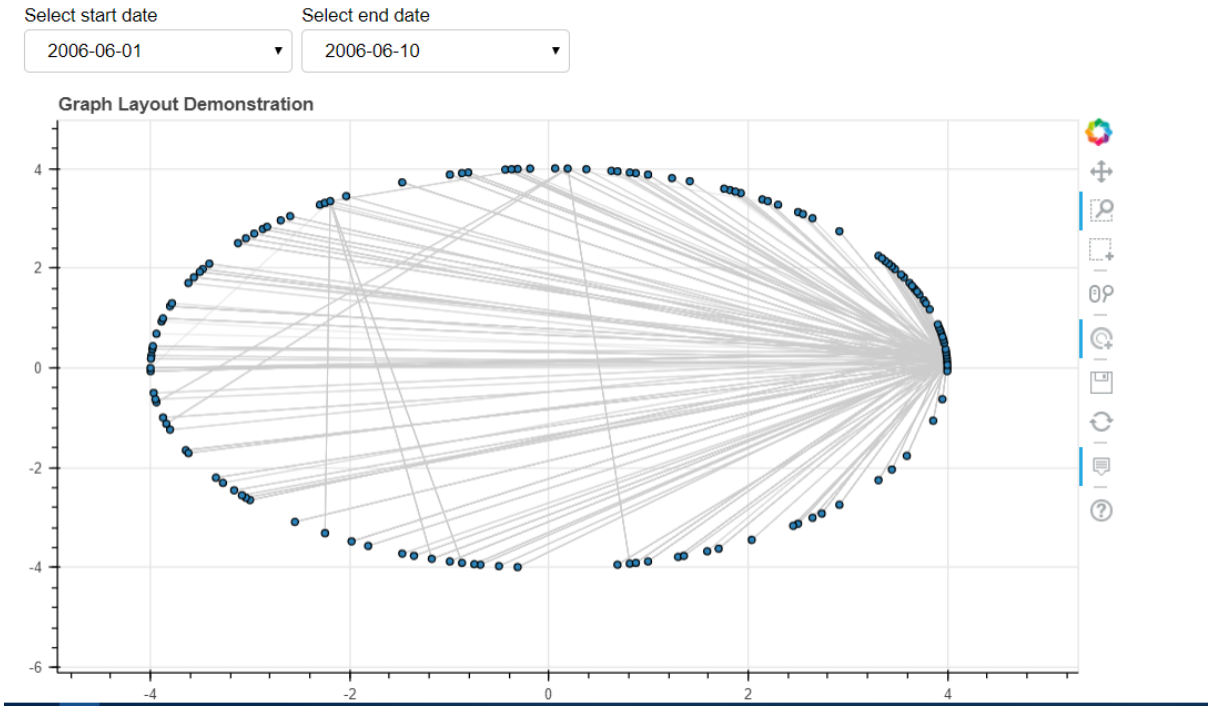
Visualization2: Social graph

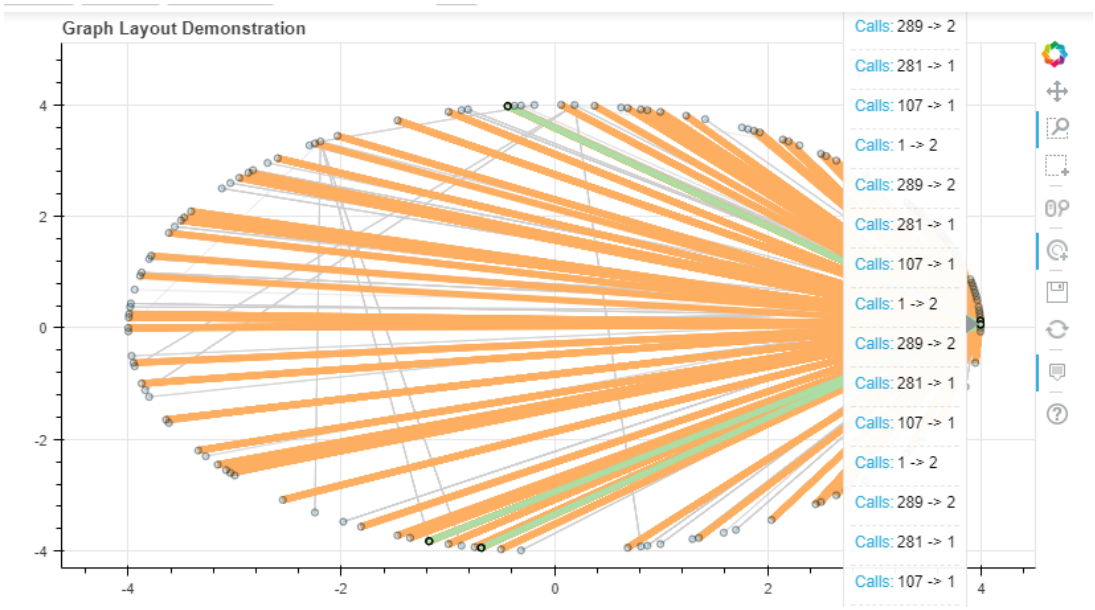
- Another graph was attempted based on callers IDs.

  Nodes: They represented user IDs.

  Edges: A call from user1 to user2. The did not contain any extra information. Although some extra information could have been encoded in graph edges.

- Observation:
  - ID 1 was most active clearly. It took bulk of the graph. He could be David Vidro according to problem statement.
  - ID 200 calls ID 5 frequently. According to problem, ID 5 could be Catalano's brother.





- On 8th day, there was no activities associated to Catalano call network. On 9th and 10th day, there was a little activities, not consistent with previous days.

Solution to MC-3

ID 200 – Ferdinando Catalano (given)

ID 1 – David Vitro (Most active)

ID5 – Estaban Catalano (brother of Ferdinando)

On 8th day onwards, there was very little activity connected to ID 200.

Note: A few more things could be done in this Mini challenge. For example, what happened 8th day onwards. We could easily find out pattern of ID1 from scatter plot and predict other IDs that Vitro might be using, and then plot social graph for that ID. But there was not enough time for a single student.