2016 PRESIDENTIAL ELECTION ANALYSIS

---- *Final Report of BIA 658*

Instructor: Prof. Rong Duan

Dun Wang, Ran Huan, Liye Pan, Tailun Song, Xianqiao Li

Project Statement

2016 United States Presidential Election

As one of the most influential events in today's mainstream world, the presidential election of the United States of America attracts attention than ever before. The 45th president of the United States will be elected on November 8, 2016.

The two main parties in the US election are republic and democrat. Currently there are 11 republicans running for the election and only 3 democratic candidates running.

Social Media and Politics

Nowadays, it is inevitable that social media has taken over the traditional media to become the mainstream to impact politics. Ever since Obama started using Facebook to engage with his supporters in 2008 and made a success, there has been more and more politic figures utilizing social media to take the upper handle of this game. From 2010 to 2014, the registered voters who follow political figures on social media has double.

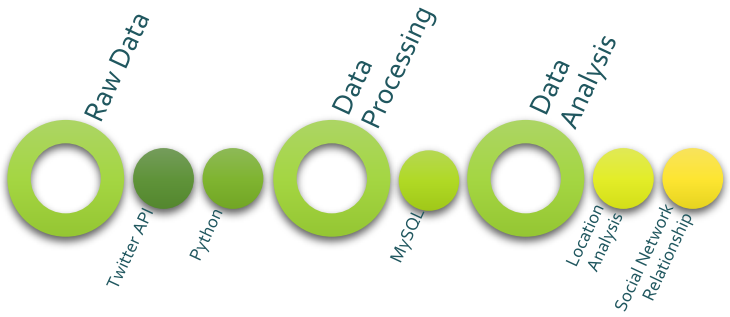
Project Purposes

•Detect the most influential person in a particular community

•Detect relationships between users in a specific geographical location (i.e. States, County, Town etc.)

•Unlock any interesting phenomena

Project Process



Raw Data Preparation----twitter API and Python

Two major twitter API been used:
GET followers/ids

https://dev.twitter.com/rest/reference/get/followers/ids
GET statuses
https://dev.twitter.com/rest/reference/get/statuses/mentions_timeline

For Python, two major packages been used in this project: tweepy and pymysql.
The basic process is, first, politician's ID and their respective followers' IDs and location information are extracted by using Twitter API and Python code. We use MySQL to store the data we collected. And put the database server on the Amazon Web Service. Last but not least, because the number of data is huge, more than 15 million records in total, we have decided to use 1% of the data collected to do the analysis.

Data Processing----SQL

a) Radom selection
   i. Premise
      1. Use random group sampling method to select 1% samples from the population.
      2. Remove duplicate samples
      3. Compare the differences.

Random samples	Remove duplicate samples	Difference	Rate
156863	155911	952	0.61%
156856	155822	1034	0.66%
157602	156632	970	0.62%

   ii. Premise
      1. Randomly select 1% samples from Hillary's database and Trump's database
      2. Remove duplicate samples
      3. Compare the differences

Random samples	Remove duplicate samples	Difference	Rate
89181	89139	42	0.047%
89458	89414	44	0.049%
88754	88703	51	0.057%

b) Comparison
   i. Premise
      1. Compare Hillary's followers with Trump's followers
      2. Remove the duplicate followers
      3. Compare the difference

Hillary	Trump	Total Number	Difference	Rate
4,742,474	4,181,164	8,923,638	445,909	5%

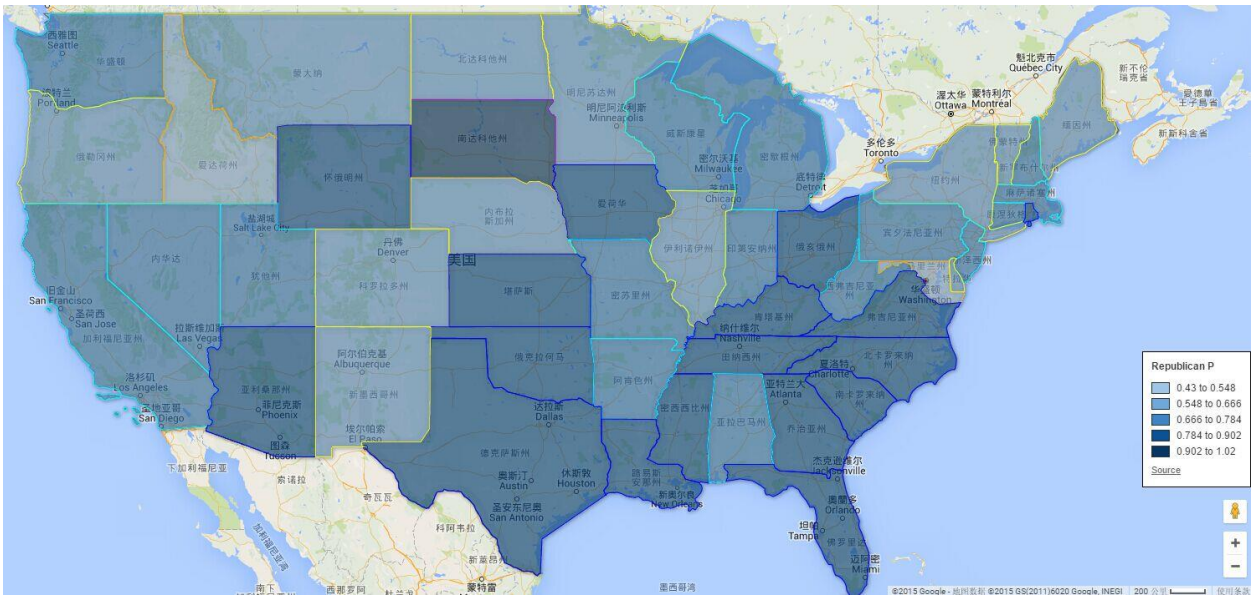
Based on the data processing, we can find that only a small amount of users follow more than one candidate. And most users' political preferences are clear.

Location Analytics----Map Visualization in Google Fusion Table

We capture the location information of the 1% size sample and the final dataset returned contains 64846 records.

The distribution of followers in state level is recorded in the following link and here is the overall screenshot.

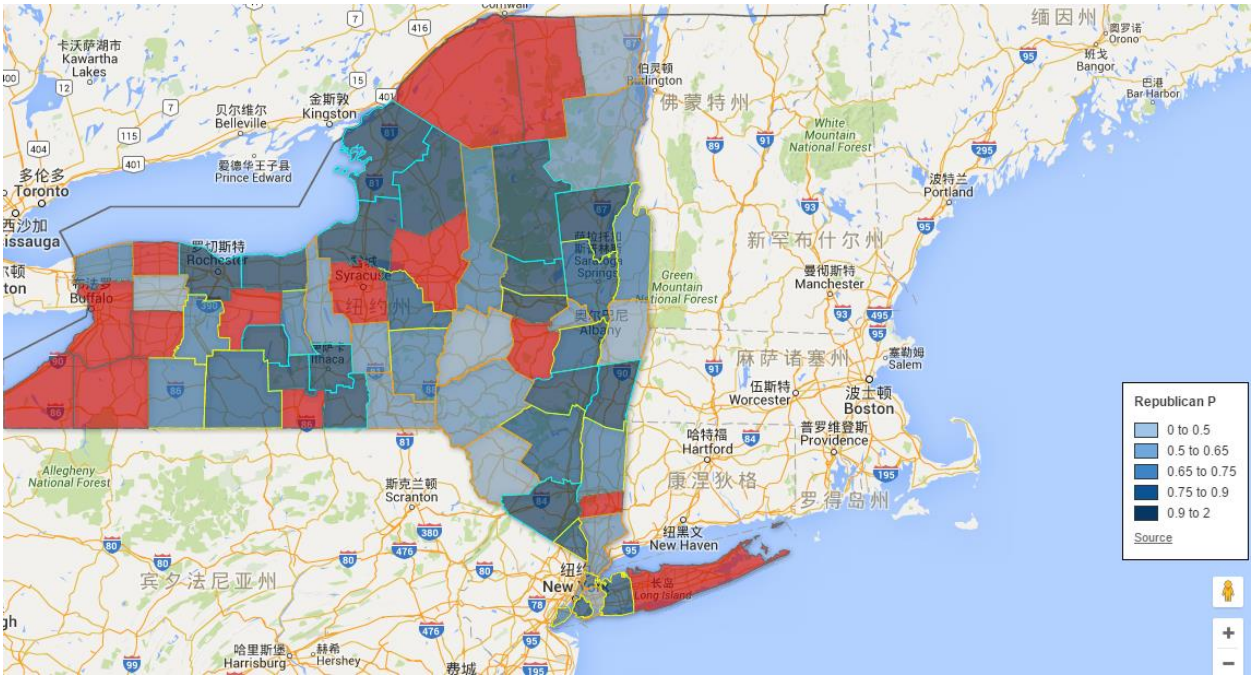
https://www.google.com/fusiontables/DataSource?docid=1qdaSGMHckgmVaZMu1P2_nGxyhcONFYapab8cDkDw



People in Southeast have a higher preference for Republican Party.

Then we look into New York State in county level:

www.google.com/fusiontables/DataSource?docid=10Hk9tv9zSI-V5nCFxC5A55Wg5cz-BlGwvVLfnWuy



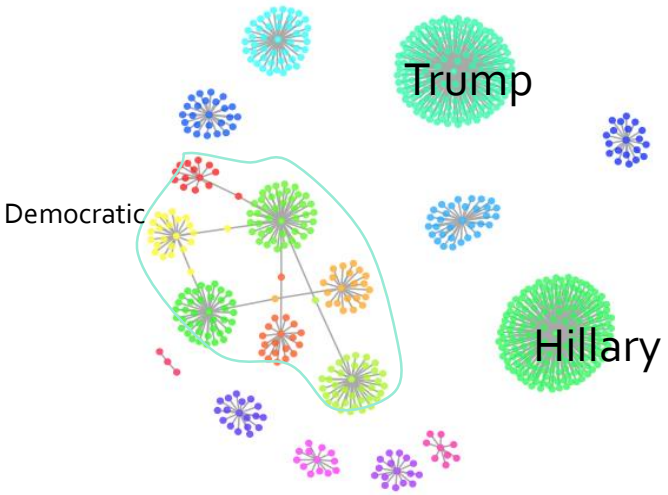
Social Network Relationship----R

Because we have large amount of data, we only choose the followers' data in NY and CA to make the social network relationship graph.

Data includes twitter_id, candidate_id and party.

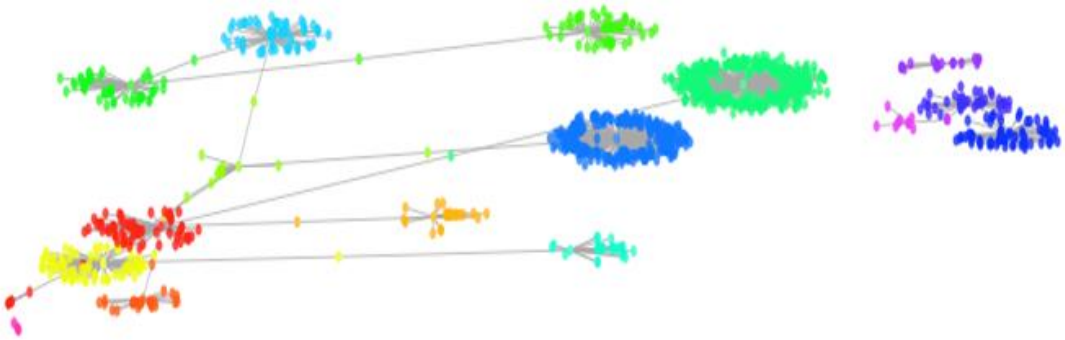
Party is a tag we set, which means we consider the party the follower follows more as the party he supports.

Social Network Relationship in NY



There are 724 accounts in this dataset. And we use igraph to plot this graph. In order to make a more intuitive graph, we also use walkstrap.community in the igraph library to implement community detection. And we set different color for different communities. In this graph, we can see that only a few accounts follow more than one candidates in Democratic Party.

Social network relationship in CA.



We have 1148 accounts in this dataset. We also set different color for different communities. In this graph, you can see some account follow more than one candidates.

Conclusions

- People in Southeast have a higher preference for Republican Party. People's attitudes are diverse inner New York State.
- Though people's political preference differs, most users' political preferences are clear.
- Commonly we think that Democratic Party is supposed to win the 2016 Election because of Hilary, but the data from twitter shows that Republican Party still have a chance to beat the Democratic and people on twitter seem to like Republican better.

Challenges and Limitations

- Over 15 million records extracted
- More than 50 hours spent
- Continuous improvement on coding and its efficiency
- Differences in languages used (i.e. Software incompatible with certain languages)
- Gather followers' other attributes and determine whether a particular account is inactive
- For example we can choose two attributes:
    - Friend density = friends_count/registered days
    - Active level = tweets_account/registered days
- Since these attributes are independent, Naïve Bayes classifier can be used to judge whether the account is fake