Sensen Chen
Interactive Visualization Final Report

Figure 1: Live Popularity

Motivation

I started this project with an initial question of how I could measure an individual person's influence. Many news sites have lists claiming that their list of individuals are the most influential people of the year. However, I wondered how these lists come to be? Are there tools that someone can use to track and measure influence? I came across papers that tackled this problem through the use of social media platforms such as Twitter or Facebook. I wanted to create a tool that anyone could use to track their own influence score or the influence scores of others. In the end, I also turned to social media, specifically Twitter, to measure popularity.

Readings

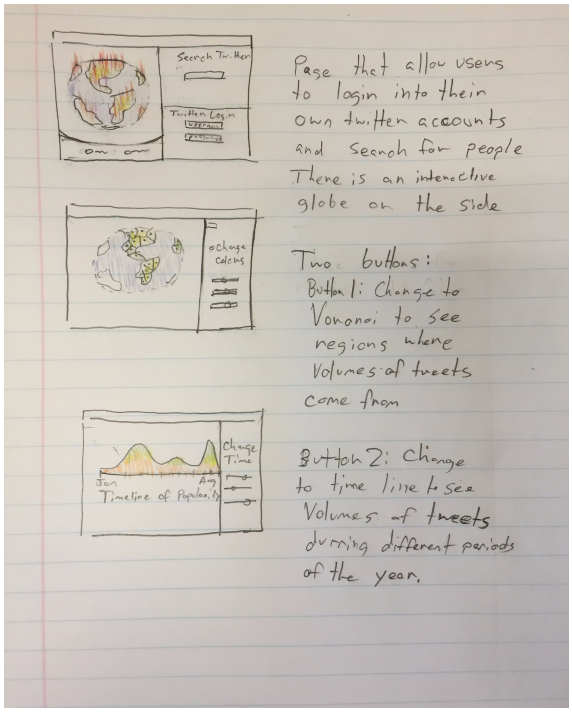
Measuring User Influence in Twitter - The Million Follower Fallacy: This paper tries to tackle the same problem of measuring influence using Twitter. The paper proposes calculating some influence score based on in degree (number of followers), retweets, and mentions. I draw many parallels between this paper and my own project and plan to use twitter in a similar way to analyze follower data.

Real-time Analytics for Fast Evolving Social Graphs: This paper specifically uses the Twitter Streaming API to analyze live data as it comes in the stream. This paper offers a scalable solution (Floe) to handle graph updates. I plan to investigate this tool in order to see if it is possible to analyze live data about a person

A Framework for Visual Analytics of Massive Complex Networks: This paper describes a framework used to visually analyze networks. Specifically, the author describes a system that is both scalable and created for large social networks. The author presents various case studies from the WoS (web of science). The author also notes that they plan on testing their framework with Twitter in their plans for the future. For my purposes, this paper offers many different methods of representing my massive dataset, many of which we have learned in class. I plan to investigate if any of these visualizations will provide a better solution for representing my data.

Sensen Chen
Interactive Visualization Final Report

Initial Storyboard



The first image is of a globe (not on fire) and the lines coming out are regions where followers of a particular person reside (location). The colors of the streaks will vary based on what degree of follower (first, second, etc.). This will help visualize the global reach of the person.

The second image is of a globe with voronoi plot separating the different regions. This helps show if there are concentrated areas where the user has more or less influence.

The third image is of a time plot depicting the the popularity of the user (based on the number of tweets about the user or directed at the user). The color will define which type of tweets and help determine if a certain type of tweet is more popular than others (from followers or news networks).

Initial Plan

My initial plan was to scrape a user's follower list, their follower's follower's list, and subsequent levels of follower's lists. The idea was to measure the size of someone's network up to a certain degree of followers. I was hoping to measure and visualize, on a global level, the audience of a person. The idea in my head was a globe that glowed where the network was (figure. 2).

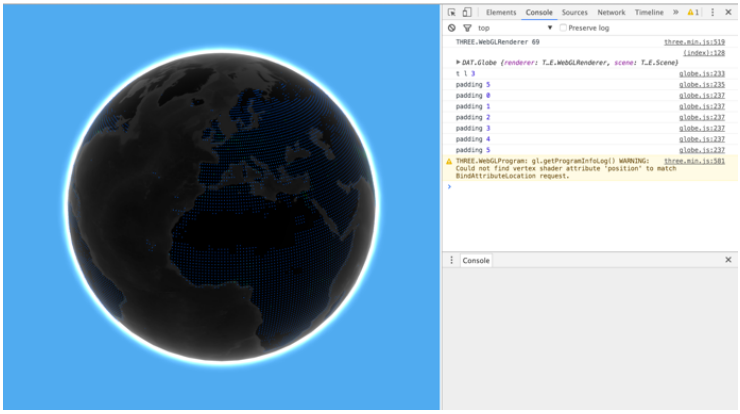


Figure 2: Globe Prototype with Glow

My first week of development was highly ambitious and I attempted to scrape all the first degree followers for a famous individual, Taylor Swift, who happens to have 75.1 Million followers. I found very quickly, that I would run into API limit requests if I continued to down this plan. In the end, I decided to gather random data about a user (through streams) and do a statistical analysis on an individual's twitter data to come up with some influence score for that individual.

Final Plan

My new idea was to have twitter streams tracking an individual person through their name and screen name. I was going to pull data from the stream, and analyze follower data extracted from the stream, which runs into less API limit calls. Using this data, I got a location (not longitude/latitude) and displayed where the tweet was sent from. I had to geocode the location (using the google geocoding API) before I could display the location.

Encountered Problems

API limits were what really limited me from creating the tool that I really wanted to make. I ran into many roadblocks and had to wait for the API keys to refresh. I ran into trouble with the Google Geocoding API, which only allows for 2500 requests per day. I also ran into trouble with the Twitter API where I tried to get Follower Lists, which would only return 180 users objects per call, but in chronological order. This call is limited to 15 times every 15 minutes. I tried to use another API call to get the Followers IDs, which would return 5000 user objects per call, with a limit of 15 times every 15 minutes. I would have to do yet another API request to get data from the user IDs. There is no way to collect the amount of data I needed fast enough from the Twitter API.

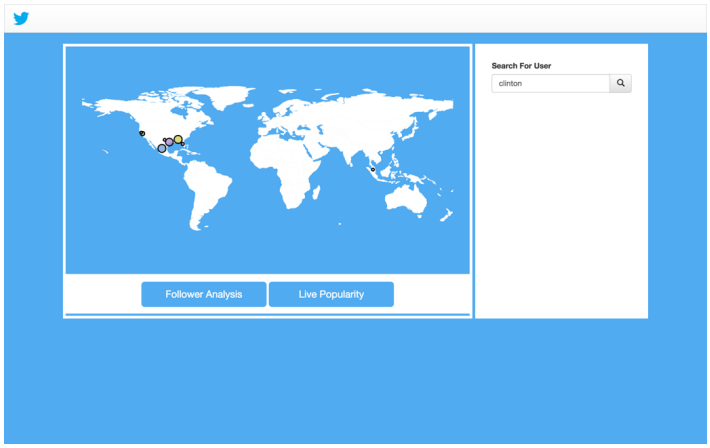
After considering just taking a sample from a user's follower's list, I ran into the trouble of collecting random data. The trouble being that the API call for follower's list returns a list in

the form of “pages” of chronologically sorted users and I could not get random pages because the id for the next page was some random, possibly hashed string. Without using random data, I couldn’t really do any meaningful analysis.

In the end I settled for the Twitter Streaming API call simply because the data coming in was random enough. Additionally, the streaming API gives users data in addition to live tweets. I spent time collecting user data and planned on doing analysis on this data.

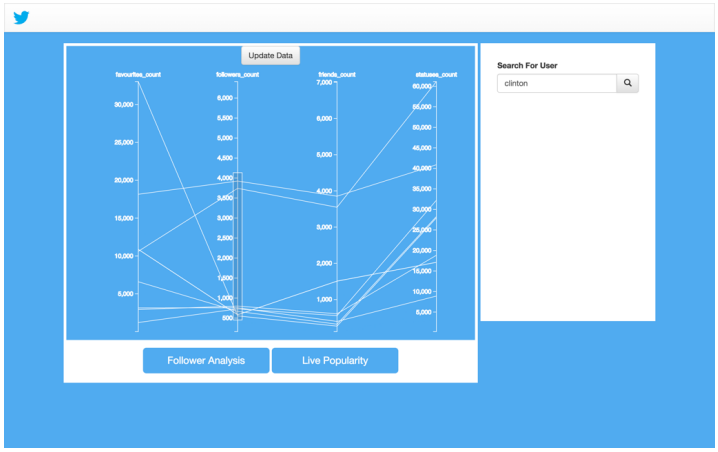
Final Plan

The core feature that I implemented is live popularity, which allows users to see live tweets and collect information about a person that could be used for analysis. The user data includes: name, screen name, account creation date, location, statues count, follower count, retweet count, friend count. I could analyze this information, and in the future come up with more interesting visualizations.



Another core feature is follower analysis, which allows users to interact and see trends in a person’s sample list of followers. Currently the information being compared are follower count, statuses count, favorite count, and friend count. One way this feature can be used for is to check which followers/tweeters might be fake given information such as statuses count and follower count (high statuses count with low follower/friend count).

Sensen Chen
Interactive Visualization Final Report



Other Features or Planned Features

One feature that I had planned to complete was popularity playback. Since I have times of when tweets are made I can play them back at in the order that they came in. This feature would be incredibly useful for analyzing popularity especially for politicians in the current political race.

Another feature that I had planned to complete was color coding on the parallel coordinates. I have researched metrics for figuring out fake users, cues such as low follower count and high statuses count, or high numbers of repeating tweets. I could easily analyze followers and categorize them as fake or not simply running the follower through a function to detect fake users. Color coding them would help users see different types of followers that they may have.

Some feedback I got from the class included increasing the length that the bubbles in live popularity stayed for and the colors for the bubbles. I wanted the bubbles to stay longer, but I realized really quickly that some people had several tweet coming in every millisecond while others had one tweet coming in every hour. Popularity really varies and another another feature I wanted to implement is flow size as a metric for bubble duration. From my demo, the colors did not mean anything, but some feedback on the class was to use color to indicate reaction of the tweet or device type of the tweet. I chose the latter because it is actually information that is passed through streams.

Bibliography

M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In Proceedings of the 4th International Conference on Weblogs and Social Media, 2010.

C. Wickramaarachchi, A. Kumbhare, M. Frincu, C. Chelmis and V. K. Prasanna, "Real-Time Analytics for Fast Evolving Social Graphs," *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on*, Shenzhen, 2015

Seok-Hee Hong, Weidong Huang, K. Misue and Wu Quan, "A framework for visual analytics of massive complex networks," *B ig Data and Smart Computing (BIGCOMP), 2014 International Conference on*, Bangkok, 2014