PROJECT 1: Social Network Analysis

Final Report

Prepared BY:

Tinh Cao| Uday Ramesh

Issue Date:

Feb 14, 2022

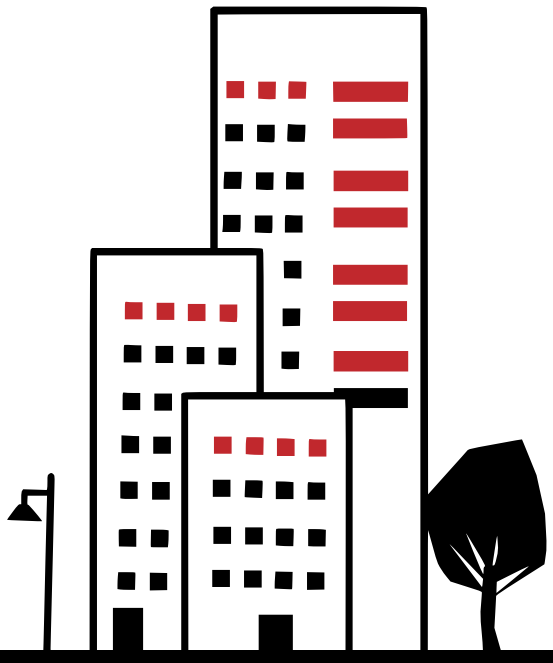


Table of Contents

1. Introduction

1.1. Project Description

This project explores some fundamental concepts in the network analysis bases of widely used social media platforms. Such networks are more relevant than ever when we consider the fast-paced moving of news in the contemporary period. With such beliefs, we choose the topic of public sentiments about vaccination against Covid-19, the pandemic that wreaks global havoc, and the mixed emotions associated with its vaccination. To capture the degree of distributed news, we believe that the diffusion network can best illustrate the significant portion of the analysis. Each node is the entity that can propagate information, and the edge is the direction to which information travels. During the initial phase, some current media platforms such as Facebook, Instagram, Twitter, and Tiktok were reviewed to decide under criteria like developer API's friendliness, user right's and liability of obtaining public dataset, and the associated public libraries and third-parties tools available which would be best to perform. Due to the exploratory nature of the project on data mining, we will develop our code to generate the dataset used later for the public datasets, create a graph structure to visualize those data, and apply some network measurements at the end. The Github repository for this project can be accessed through https://github.com/tccao/TwitterCrawlerAnalysis.

1.2. Review of Project Objectives

Stated Project Objectives	Status	Comment
Decide on the project platform.	*Met*	*Twitter was chosen among other platforms.*
Data Collection	*Met*	*Using Twitter Public API-v2 library to crawl data.*
Data Visualization	*Met*	*Through file "Twitter Analysis.ipynb"*
Network Measures Calculation	*Met*	*Through file "degree-of-distributions.ipynb"*

1.3. Project Schedule

Milestone	Scheduled Completion Date	Actual Completion Date
Platform finalized	01/14/2022	01/18/2022
Initial credential Setup	01/19/2022	01/22/2022
Data's crawler setup	01/25/2022	01/27/2022
Graphing Phase	02/04/2022	02/14/2022
Network Analysis	02/08/2022	02/14/2022

2. DATA COLLECTION

2.1. Platform Review

Initially, we picked Facebook as our target network due to its enormous user-base and popularity among a wide range of ages. Obtaining the developer's account for Facebook took us some time due to its strictly-enforced regulations on accessing their user information after the infamous scandal's Cambridge Analytica. Even after being approved for a developer account, the use of Facebook official APIs was limited down to a predesigned role that Facebook made available, like enterprise, apps, marketing, etc. Most of the interactions with the API need to come through an app, or a page, that the developers currently own, which did not apply to us and defeated the purpose of generating a public dataset that fitted into our exploration. Further reviews of Facebook'permission and legality document revealed that obtaining public user data like public user posts/page posts were prohibited, so there was no way to use their official API to generate the dataset. Similar issues also arose from others media platforms such as Tiktok and Instagram, which we planned as the alternative options.

This brought us to the unofficial method of generating the dataset from the website without the API, the Web crawler, and the web scraping method. A web crawler, or crawler, would systematically browse our target website and copy all the web contents down, or contents indexing, to search the website data more efficiently. There are many concerns about using such tools. One of them is the heavy loads they put on web server systems. We spent some time reviewing the basics of building a custom web crawler from scratch, but it turned out that we did not have enough time to develop a fully-functional scraper. Another barrier from using a web crawler is listed in the Robot.txt files from many major social media websites that officially prevent using such tools to crawl data. As such, we reconsidered our approach once again and ended up with the Twitter platform due to its excellent API library and supportive permission control.

2.2. Data Crawler

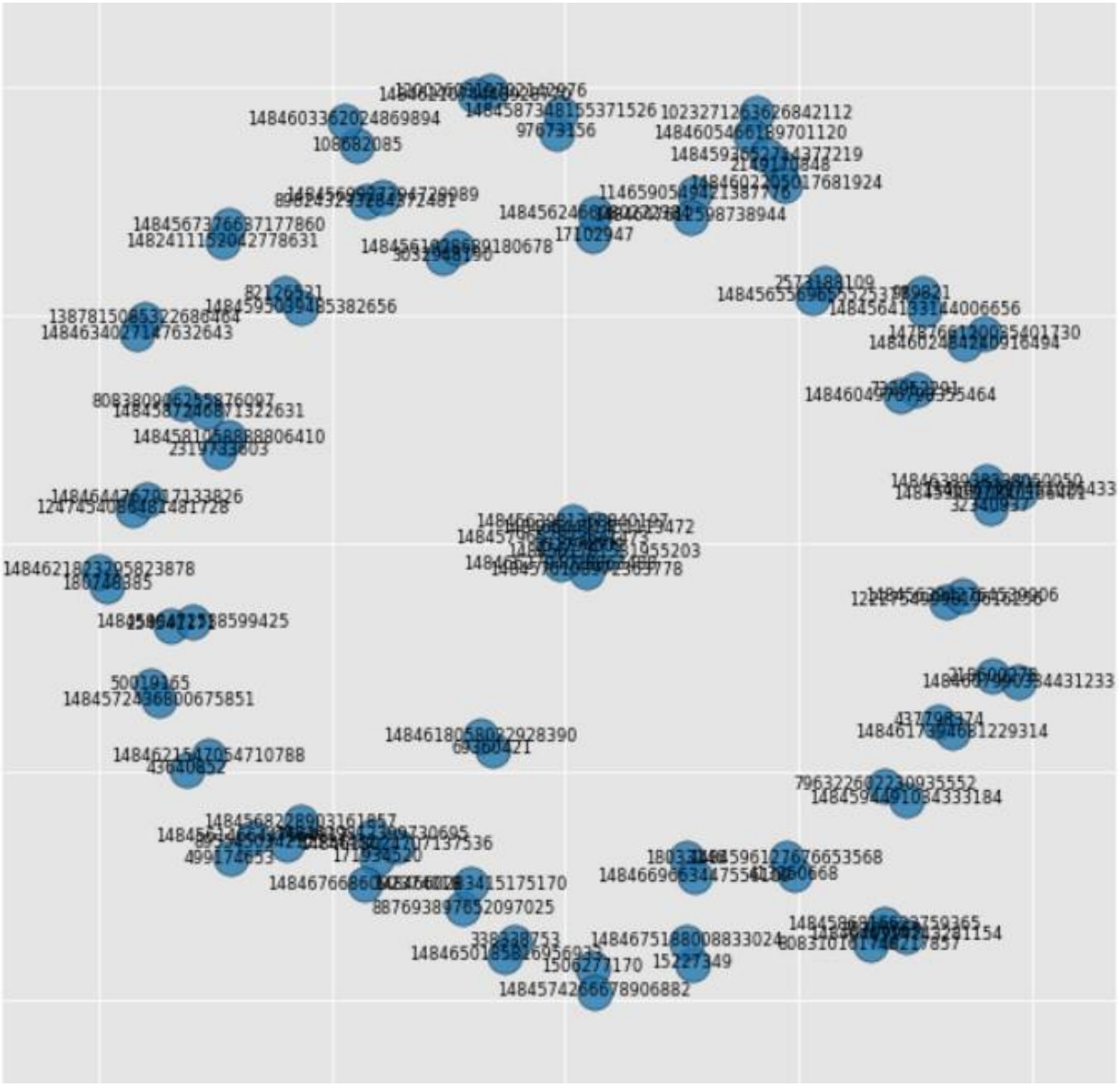
After choosing Twitter as our social media platform, we set up the necessary account and asked for Elevated permission, which allowed more API calls to Twitter's server. Getting the account set up took a few days since Twitter's team needed to review our permission request. After everything is done, we start working on getting ourselves familiar with the terms and jargon of the Twitter library.

The big picture of the crawler was to make a recent search query call containing hashtags of covid-19 and vaccination to Twitter's server and get all the original tweets that use those tags. Using what was returned, we started the second round of tweets look-up API calls to search for any retweets associated with those original tweets, representing the directed edges in the graph later on. The third round of user look-up API calls looked up the user information that owned the retweets, which meant the node that received information from our original author. Every tweet would involve 2 API calls, one to get the edge information and the node information. However, after making the third API call, we had reached the maximum call limits in the 15 minutes timeframe for the recent search query, leaving us with a network of 120 nodes. All of the data were written into three different comma-delimited files for ease of graph import later. The final dataset for the network would be extracted and recreated from these three raw files. The method

would have more room to improve, and the performance was not matched with third parties' publicly available. Still, it was adequate for our exploratory purpose.

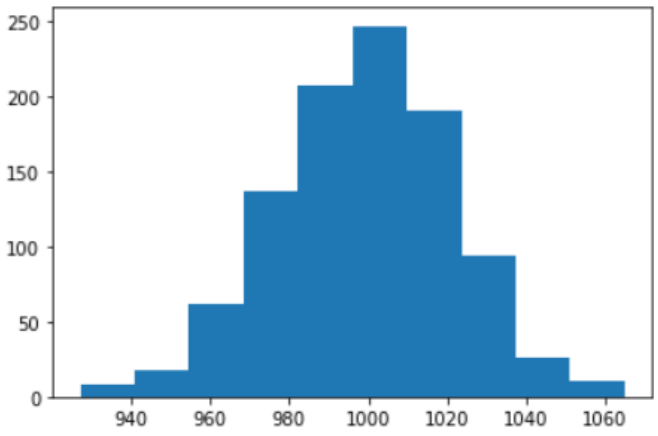
3. GRAPH AND NETWORK MEASUREMENTS

Here is the graph for file *retweet_user_data*.



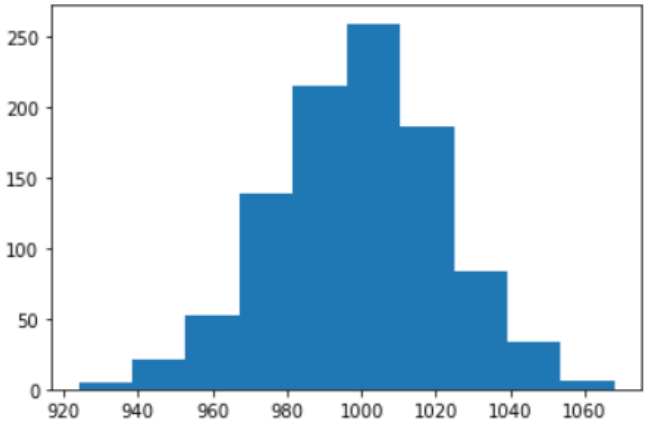
Here, we can also see the degree distribution of each node and its histogram.

	nconst	degree
0	14254409	1
1	Raymond	1
2	17573273	1
3	Arthur Caplan	1
4	1257284551	1



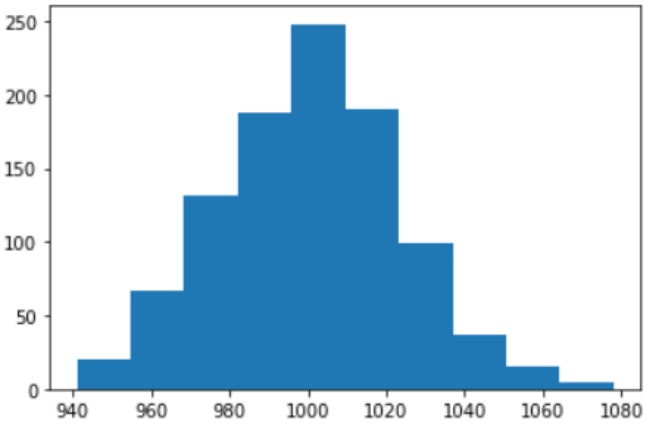
The next would be for *original_tweets_data:*

	nconst	degree
0	252794509	6
1	171934520	3
2	2149170848	2
3	1484676686092374018	1
4	15227349	1



Finally, this is the *retweet_tweets_data*:

	nconst	degree
0	1257284551	1
1	1477621785889849347	1
2	1463396344555446276	1
3	1485036636836564992	1
4	25059481	1



4. TEAM EFFORT

4.1. Tinh Cao

I was in charge of the first half of the project, including making platforms review, completing data collection step, setting up appropriate Github's repository for collaboration, task split, and compiling the write-up report.

4.2. Uday Ramesh

I was in charge of visualizing data into a graph network and performing associated network measurements.

5. References

Edward, A. (2022, January 6). *Collecting tweets from Twitter API v2 using Python 3 | Towards Data Science*. Medium. https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a

Lloy Pearson, B. (2021, May 28). *How to Use the Python Requests Module With REST APIs*. Nylas. https://www.nylas.com/blog/use-python-requests-module-rest-apis/

Russell, M. A., & Klassen, M. (2019). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More* (3rd ed.). O'Reilly Media.

Twitter API Documentation. (n.d.). Docs | Twitter Developer Platform. https://developer.twitter.com/en/docs/twitter-api