Analysis of tweets hashtag connectedness in tweets related to Covid-19 during the pre-pandemic stages of the Covid-19 pandemic

Matt Stirling   
*University of Oulu*Oulu, Finland  
mstirlin18@student.oulu.fi

*Abstract*—This is the abstract of my paper

Keywords—social network analysis, Twitter, COVID-19, structural balance

# Introduction

Text ...

# Problem Description

The problem description as I see it was to analyse a graph that was to be constructed from a subset of the TweetsCOV19 dataset [1]. The dataset would be limited to tweets between December 2019 and February 2020, and the graph would be constructed by representing individual tweets as nodes and adding edges between any two nodes where the tweets that they represent share at least one hashtag. The graph would be saved in a manner that allows for it to be loaded into a graphing library (in my case an edge list) and the adjacency matrix of the graph would be created.

Then, the following analysis would be done on the network:

* Idenfity the three largest connected components and determine their average path lengths.
* Plot the degree centrality of the graph and investigate whether it is power law distributed.
* Plot the evolution of *cumulative* average positive and negative sentiment scores of the tweets over the selected timespan and comment on the descrepancy between the two graphs.
* Analyse the evolution of transitivity of the network by looking at the cumulative graph up to each point in time.
* Analyse the evolution of structural balance of the network by looking at the propotion of balanced and unbalanced triangles.
* Investigate the occurrence of unbalanced triangles as a virus propogation case.

After this analysis, I am supposed to comment on the limitations of the overal reasoning pipeline and suggest literature to support the findings.

# Dataset Description

The dataset from which the network was constructed was Part 1 of the TweetsCOV19 dataset, which is a semantically annocated corpus of tweets about the COVID-19 pandemic. It contains over 8 million tweets between October 2019 and April 2020 which were selected from the TweetsKB dataset, a much larger corpus of nearly 3 billion tweets spanning over 9 years. The basis for selection from this larger dataset was if the tweet text contained keywords related to the pandemic; a seed list of 268 keywords was used. The text of the tweets themselves were not made available in the TweetsCOV19 dataset and the user IDs were also encrypted for anonymity.

Each tweet instance contains 12 features related to the tweet.

1. Tweet Instance Features

| # | Feature | Description |
| --- | --- | --- |
| 1 | Twitter ID | Used internally by Twitter to uniquly idenfity twitter objects, inlcuding tweets, users, direct messages, etc [2]. |
| 2 | Username | Encrypted |
| 3 | Timestamp | in the format “EEE MMM dd HH:mm:ss Z yyyy” |
| 4 | #Followers | Number of followers the user has |
| 5 | #Friends | Number of friends the user has |
| 6 | #Retweets | Number of times the twees was retweeted by others |
| 7 | #Favourites | Number of times the tweet was favourited by others |
| 8 | Entities |  |
| 9 | Sentiment | Both a positive and negative sentiment score was calculated for each tweet (is is unclear how this was done) and these values are listed as a string represented by a whitespace char |
| 10 | Mentions | Usernames (which were unencrypted in this case) that were mentioned in the text of the tweet |
| 11 | Hashtags | Hashtags used in the text of the tweet |
| 12 | URLs | URLs used in the text of the tweet |

Some features – such as seniment or hashtags – need to be parsed as strings to extract the desired data they contain. The dataset is available for downlad in both Notation3 and TSV (Tab Separated Value) formats.

# General Methodology

It is worth starting by considering the nature of the graph we will create via the described method. Each hashtag that is associated with more than one tweet will produce a complete graph where the number of edges is:

where n is the number of tweets that use the given hashtag. Each node in this subgraph will therefore have a *minimum* degree centrality of nC2 and the minimum number of complete subgraphs of the whole network will equal the number of hashtags associated with more than one tweet.

As a result of the construction method, the network as a whole will be very dense with lots of edges, which can potentially make some types of network analysis challenging. However, some aspects of the network could be discovered or estimated by instead investigating an abstracted version of the netwok, where we represent the hashtags as nodes and add edges between nodes wherever a tweet shares both hashtags. For example, we can find the diameter of the full network by finding the diameter of this abstracted graph and adding one (finding the diameter was not part of the problem description, however). We could also find connected components of the whole graph, estimate the average shortest path, or detect communities (also not part of the problem description).

# Detailed Methogoloy

I have conducted all the analysis in python. I mainly used the

* **Pandas** for loading data into data frames
* **Dask** for loading data into data frames that are partitioned and loaded in a *lazy* manner to allow to handling of data that exceeds available RAM

|  |
| --- |
| // Helper functions  **function** find(x, parents)  **if** parents[x] ≠ x  set parents[x] to find(parents[x])  **return** parents[x]  **function** union(a, b, parents, ranks)  set ap to find(a)  set bp to find(b)  **if** ranks[ap] < ranks[bp]  set parents[ab] to bp  **else if** ranks[ap] > ranks[bp]  set parents[bp] to ap  **else**  set parents[ab] to bp  increase ranks[bp] by 1  // Initialize union find  set PARENTS to (0, 1, 2, …, n-1, n)  set RANKS to (0, 0, 0, …, 0, 0)  **for (**source, target) **in** EDGES  union(source, target, PARENTS, RANKS)  // produce list of nodes in each component  set COMPONENTS to empty container  set NODES to {x : (x,x) in EDGE LIST }  **for** node **in** nodes  p = find(node, PARENTS)  add node to COMPONENTS[p] |

1. Pseudocode of Union Find algorithm for finding list of nodes in each connected component given an edgelist of the graph.

# Results and Descussions

Text ...

A graph of a graph of a number of hashtags

Description automatically generated

1. Histogram of amount of hashtags per tweet, ignoring tweets that contain no hashtags (which are not added to the network, although they are included in sentiment analysis later on).

The total number of unique hashtags used by at least one tweet was 292,264. However, the number of hashtags that were used by more than one tweet was 81,931 (28%), the other 78% of all unique hashtags used will not contribute edges to the network.

A graph of a number of hashtags

Description automatically generated

1. Frequency of use of top 10 hashtags.

A graph of the number of coronavirus

Description automatically generated

1. Amount of edges in subgraph added to the network by the hastag. This is the minimum amount of edges added by that hashtag.

## Connected Components

Via my implementation of the Union Find algorithm, I was able to find 7,746 connected components in the graph with more than one node (tweets that did not connect to any other tweets were not included in the edge list and are therefore ignored). As we can see from the table below, the vast majority of nodes (94%) are in the largest of these components.

1. Amount of Nodes

|  |  |
| --- | --- |
| **nth largest component** | **Amount of nodes** |
| 1 | 385,468 |
| 2 | 95 |
| 3 | 80 |
| 4 | 78 |
| 5 | 71 |
| 6 | 68 |
| 7 | 67 |
| 8 | 59 |
| 9 | 57 |
| 10 | 52 |

The average size of these components is 53, and if we ignore the largest component, the average drops to 3.3. These isolated communities

A close-up of a network

Description automatically generated

1. 2nd largest connected component

A network of blue dots and lines

Description automatically generated

1. 3rd largest connected component

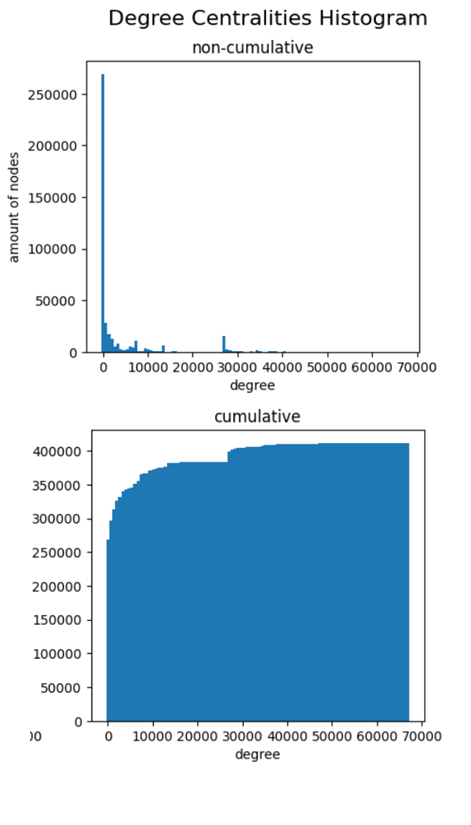
A network of blue dots and lines

Description automatically generated

1. 6th largest connected component

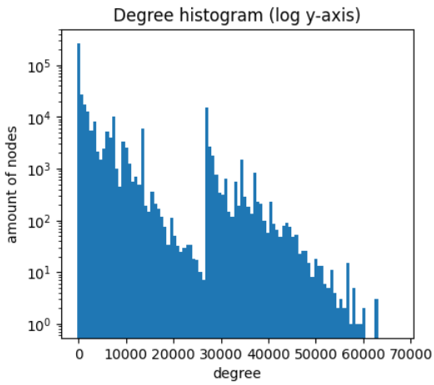
## Degree Centrality

The degree centrality



1. Histogram of degree centralities

Text



1. Log plot of degree centralities

Text

A graph with red dots

Description automatically generated

1. LogLog plot of degree centralities

# Conclusion and Prespectives

Text …

##### References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

<https://developer.twitter.com/en/docs/twitter-ids>