Hashtag Network Analysis

RITURAJ KULSHRESTH*, B18CSE046
SAURABH BUREWAR†, B18CSE050

Abstract: In this project, we have analyzed the social network formed by the users when they use a particular hashtag in a tweet on Twitter or post on Koo to understand how the network translates in the real world, and the social impact of a particular hashtag in the real world. This will be used to understand the network structure that is formed in the most popular hashtags. These hashtags are related to society and help us understand the online interaction that is necessary to make a hashtag/social cause successful online, using various social network analysis metrics.

Keywords: *Social Network Analysis, Twitter, Koo, Network Science*

1 INTRODUCTION

In this project, we have analyzed the social network formed by the users when they use a particular hashtag in a tweet on Twitter or post on Koo, to understand how the network translates in the real world and the social impact of a particular hashtag in the real world. This will be used to understand the network structure that is formed in the most popular hashtags. These hashtags are related to understanding human behavior is necessary to understand the behavior of people or objects that behave in a manner similar to society or that is bound by some social rules. Analysis of present social networks allows us to predict how an individual in a social network or a group of people belonging to a social network will react to a situation. It also helps us understand what a particular reaction among the members of the social network may lead to in the real world. Social network analysis is the process of investigating social structures through the use of networks and graph theory. In a social network, Nodes are used to represent the entities in the network and hold some kind of property or value that translates into a real-life value. The edges represent the connection between the nodes and may hold some property as well which may translate to real life. The edges denote how the nodes or the attributes in the real world are connected in the social network formed. This can mean interactions, relations, activity, or some other feature between the attributes. Some examples of uses of Social network analysis are social connections, virtual routing networks, physical electricity networks, roads network, biology relations networks, and many other relationships.

Twitter (initially called Twttr) is an American micro-blogging and social networking service on which users post and interact with messages known as "tweets".Twitter is one of the most popular social networks today and is a popular choice among data science and social network analysts to understand the behavior of society based on the tweets and posts put on the service. In this project, the analysis of the posts on Twitter will help us in getting a good understanding of the real-world information network that is present. Koo is also a budding micro-blogging and social networking service in India that has seen a recent surge in popularity and its social network analysis will be helpful in getting an insight into the information network present in India. It will help us understand the dynamics of how the network varies in and out of India on Twitter.

2 PROBLEM STATEMENT

- Analyse the Social Network formed by the users in Twitter while using hashtags and mentions and get the necessary details of the network formed to perform a social analysis of the equivalent real world scenario.
- Scrape the Koo website to get the post, follower, following data, similar to what is accessible using Twitter's API.

*Rituraj Kulshresth - Working with twitter API. Performed analysis of the networks and report 50%
†Saurabh Burewar - Working with twitter API. Scraping of Koo webapp and report 50%

3 RELATED WORK

Social networks provide a platform and an environment where people can share their opinions and instigate discussions that can lead to a social movement. Therefore, studying and analyzing social networks can help in learning the information flow across users [3][5][9]. Many similar studies have been done where social networks were analyzed to understand their effect of these networks on the spread of an ideology or a movement. One such study is [4], where public discussions were analyzed related to "Saudi women can drive" using certain keywords including the *#SaudiWomenCanDrive* to understand the spread of the idea in the community. [6] mined tweets to demonstrate the use of social media in understanding public thoughts on medical topics like drug detection systems. [2] used data mining techniques, embeddedness, betweenness, etc. and designed an affiliation graph to identify popular communities and understand social behaviour. Another research in [1] explored the use of Twitter in saving lives by analyzing blood donation requests. [7] analyzed the effect of Twitter users and their opinions on the Brexit referendum and the possibility of outcome of the referendum being influenced by these discussions. [8] studied the changes in teaching and learning through the study of the Next Generation Science Standards chat hashtag (#NGSSchat).

4 OUR CONTRIBUTION

Most of the studies done in Social network analysis are done on a specific hashtag in order to target a specific movement. One good example of this is the study published in [4]. We decided to build networks of different hashtags pertaining to successful and unsuccessful social movements to try and see the difference between them.

Since Koo is a new platform in the micro-blogging and social media market, there aren't any studies or datasets for the platform yet. We built our own dataset by scraping the Koo webapp and built our networks on those datasets to perform a study similar to what we are doing with the Twitter dataset.

5 EXPERIMENTS AND RESULTS

5.1 Data Collection and Preprocessing

In order to work on the analysis of the social network formed from the tweets on Twitter and Koo, we had to obtain the tweets for the various hashtags.

5.1.1 Twitter. We chose to collect the data regarding the tweets and the user who first wrote them. We used a Standard package developer account to get access to tweets which only gives the tweet data for the last two weeks. To access the twitter API easily, we used a python library called Tweepy. We used Tweepy to get the data of the tweets such as username, description, location, following, followers, total tweets, retweet count, text, has text, t_id. This data was used to understand the information format provided by Twitter. Following this, we selected the tweets belonging to a particular hashtag and collected 1000 to 2500 tweets. However, Twitter treats retweets as individual tweets while responding to a request through the API, hence, all the tweets that were collected had to be checked whether they were retweets or not and if they are already recorded. If the tweet id that we get is new it is stored for further processing. Twitter has a limit of 900 'Tweet ID' requests per 15 minutes which makes the process considerably slow and tedious. Twitter also has a rate limit of 75 access requests for the tweet details hence we did not collect the details initially. Once the individual, non-repetitive tweets were collected we then ran another program to get the tweet details and check for the retweets. the process was broken into two parts since the limit of 75 accesses per 15 min would make the collection of tweet ids very slow. Hence, we only collected the ids which have a higher access rate limit. Also, the rate limit for checking the details based on the tweet id and collecting the retweeter info was the same 75 tweets access per 15 min so it made sense to club these processes together into one and complete the data collection.

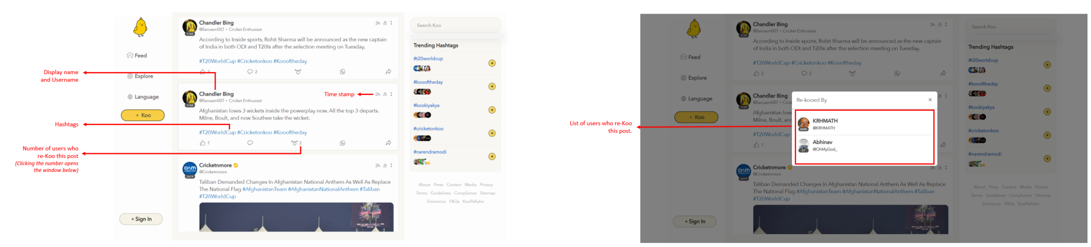
Next, we create a master file with data of the tweet and the individual who posted the tweet originally by exporting the pandas data frame storing the data. The data collected are username, following, followers, total tweets, retweet count, retweeter list. The retweeter list is used to create a network using networkx and exported as an edge list. This process is extremely slow since the data access rate provided by Twitter API is 75 requests per 15 minutes, therefore, it takes hours to collect the data of a tweet at a time. Also, multiple exceptions need to be handled since the connection between Tweepy and the API breaks after a minute of inactivity, hence we need to reconnect it again and again. Since the API request rate is so slow for Twitter we chose not to make the hashtag-hashtag network.

Table 1. Data from the Twitter API

Username	following	followers	totaltweets	retweetcount	retweeters
johnpringdns	3841	22135	48913	13	['Zebra$_t$*alks′,′ Russell_E__′,′ MDPAccessGroup′,′ ScrimshawsFire′,′ acagoldsmith′,′ purrrmeowpurrr′*, ...]
Autist$_W$*riter*	622	688	4119	0	[]
autismage	238	6857	11805	8	['Russell_E__', 'missiemill', 'BoycottSpect10k', 'johnpringdns', 'niwrowahanol', 'JaneCas65', 'ElizabethWorsl1']
FadeN2Yuu	917	324	19774	3	['Noodnood966', 'TheEndSarsBot', 'Gnrl$_S$*trike*$_B$*ot′*]
...	...	...	...	...	...

5.1.2 Koo. Koo is a fairly new social media platform and hence doesn't have public APIs that can be accessed to get easy access to data. So, we scraped the Koo webapp feed to get the data.

The Koo web app feed can be seen in two ways - A hashtag-specific feed where all posts are about a hashtag or a combined feed with all posts. We chose to scrape the hashtag-specific feed to get data targeted to a specific hashtag. The feed page itself has a simple layout with infinite scrolling which loads more posts as we scroll. Each post has post and user information like username, post content, hashtags, etc., and also has a "rekoo" option (similar to retweet) at the bottom, which on clicking also shows the users who have "re-kooed" the post.



(a) (b)

Fig. 1. (a) shows the Koo webapp feed and the information that can be scraped. (b) shows the modal of rekoo users list that pops up when the rekoo number (highlighted in (a)) is clicked.

The HTML of the feed page is simple with a container div containing all the posts. Each post is a div tag with a unique id which further contains all information about the post. Our program scrapes important information from each post like username, content, likes, hashtags, etc. using ids and class names of the corresponding div tags and if the post has been rekooed, it scrapes the usernames of rekooers as well. This is done by executing a click on the rekoo option which opens a modal (Fig. 1 (b)) and then reading the usernames of all users that appear in the list in the modal. The program also executes scrolling to load more posts and saves the ids of the posts that have been scraped to avoid duplicate posts in the feed. It also executes scrolling to load more posts once the existing ones are scrapped.

We used the Selenium library for web-scraping since it makes scrolling and opening and closing of the rekoo list possible using JavaScript executions.

Table 2. Data scraped from the koo feed

Username	date	content	likes	hashtags	rekoo-users
@pemakhandubjp	04-11-2021	India achieving...	6	['#VaccineCentury', '#CovidVaccination']	['@sethircno']
@Hindusthansamachar	04-11-2021	Prime minister...	1	['#vaccination', '#VaccineCentury']	NaN
@outbreak$_{i}ndia$	02-11-2021	Day 290:01-Nov, 7PM India ...	1	['#VaccineCentury', '#vaccination', '#vaccinated']	NaN
@DrTamilisaiGuv	29-10-2021	Several Covid warrio ...	22	['#Coimbatore,', '#VaccineCentury']	['@harish4M0HK', '@umeshtomarmp', '@Thilak$_{V}R'$]
...	...	...	...	...	...

NOTE: The data collected from the twitter API was taken from the week starting from 26/10/2021 to 1/11/2021 for all the hashtags. The data from Koo corresponds to 23/10/2021 to 06/11/2021 for *#t20worldcup* and all the available posts for other hashtags (since the activity is not as high).

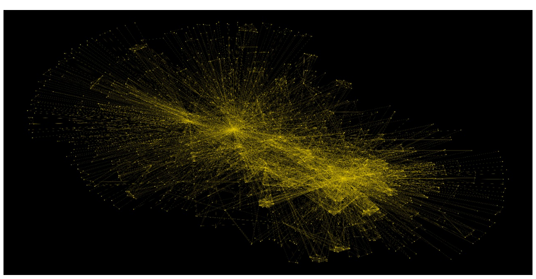
5.2 Networks

To analyse the data we extracted from the two social platforms, we built two types of networks from them.

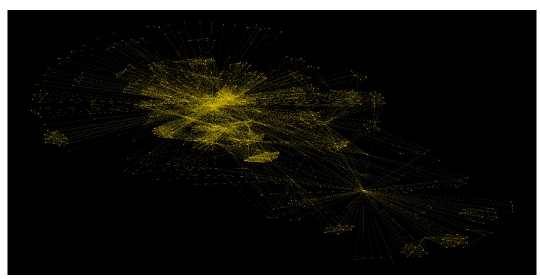
- Retweet network: Every node is a user and an edge is formed when a user retweets/rekoo-es another user.
- Hashtag-Hashtag network: Every node is a hashtag and when two hashtags appear in the same tweet/koo, there is an edge between them.

For twitter, getting hashtags would significantly increase the time to get the data due to API limitations, which limited us to retweet network only. For Koo, we were able to build both types of networks.

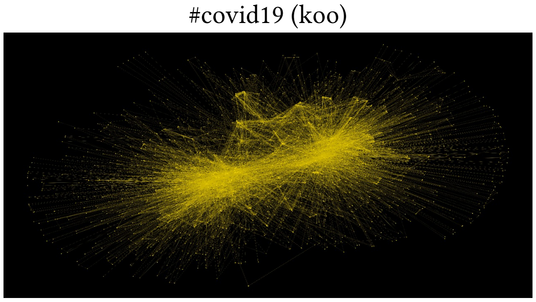
Below, in Fig.2 and Fig.3, are the images of all the networks built in this project. High resolution images of the networks can be found in the code repository here.



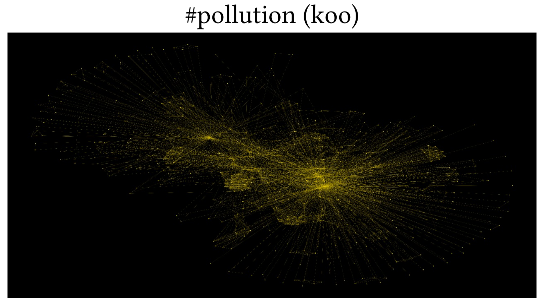
#covid19 (koo)



#pollution (koo)

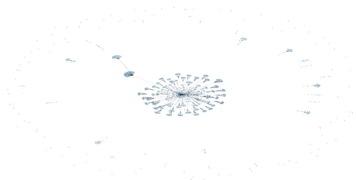


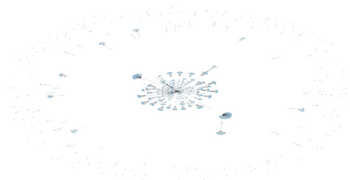
#t20worldcup (koo)

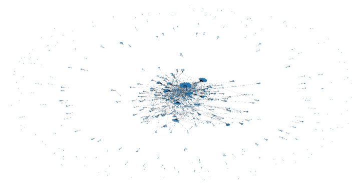


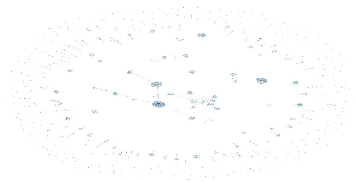
#VaccineCentury (koo)

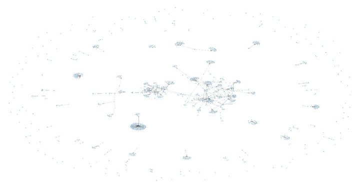
Fig. 2. Hashtag-hashtag networks for koo hashtags


#BlackLivesMatter

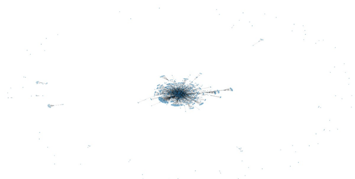

#BlackLivesMatter_update


#Diwali

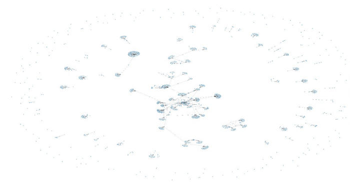

#LGBTQ


#ufo

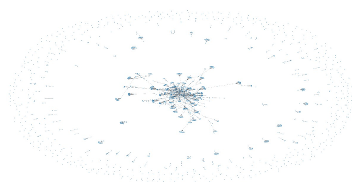

#whitelivesmatter

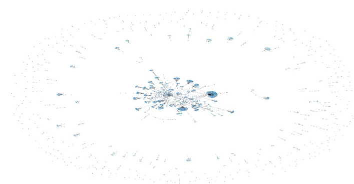

#BoycottBollywood

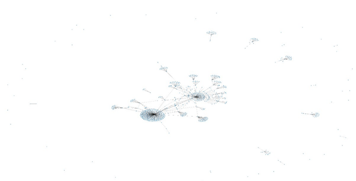

#T20WorldCup

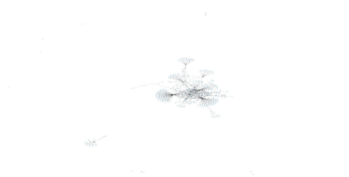

#metoo

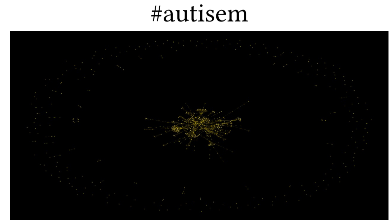

#crackerban

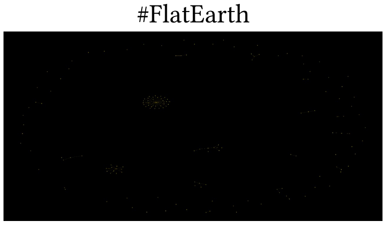

#viratkohli

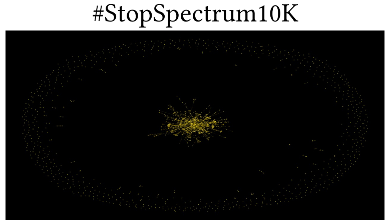

#autisem


#FlatEarth


#StopSpectrum10K


#covid19 (koo)


#pollution (koo)


#t20worldcup (koo)

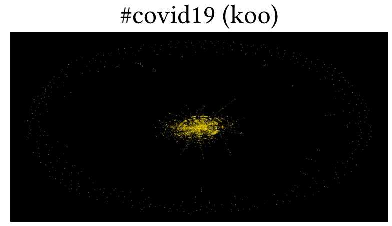

#VaccineCentury (koo)

Fig. 3. Networks of all the hashtags.

5.3 Experimental Setup

In order to understand the network we first read the edge list of the user retweeter that we have saved previously. We used various metrics to understand the network. The metrics used are as follows:

(1) **Node Count:** It is the number of nodes present in the network.
(2) **Edge Count:** It is the number of edges present in the network.
(3) **Average Degree:** The degree of a node is the number of edges connected to a node (undirected). The Average degree for all nodes is the Average degree.
(4) **Number Triangles:** It is used to detect communities and measure the cohesiveness of those communities. It can also be used to determine the stability of a graph and is often used as part of the computation of network indices, such as clustering coefficients.
(5) **Number Components:** The number of components is an important topological invariant of a graph. In topological graph theory, it can be interpreted as the zeroth Betti number of the graph. It is used in characterizing the graph with perfect matching and graph toughness.
(6) **Size Largest Component:** We have used the node count of the components as a measure of the size of the component. the largest component is defined by the largest connected sub-network of the original network that contains a fraction of the entire graph's vertices. It signifies the reach of the information that a person can transmit.
(7) **Clustering Coefficient:** It measures the proportion of your friends that are also friends with each other (i.e., what amount of mutual trust people have for each other).
(8) **Density:** The density of a graph is a measure of how many ties between actors exist compared to how many ties between actors are possible.
(9) **Assortativity:** Assortativity or assortative mixing is a preference for a network's nodes to attach to others that are similar in some way. For instance, in social networks, nodes tend to be connected with other nodes with similar degree values. This tendency is referred to as assortative mixing or assortativity. On the other hand, technological and biological networks typically show disassortative mixing, or disassortativity, as high degree nodes tend to attach to low degree nodes.
(10) **Degree Centrality:** The degree centrality for a node is simply its degree. This degree is scaled to a number between 0 and 1. the node with the highest degree in the network will have a degree centrality of 1, and every other node's centrality will be a fraction of its degree compared with that most popular node.
(11) **Betweenness Centrality:** Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. The vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness.
(12) **Closeness Centrality:** The farness/peripherality of a node is defined as the sum of its distances to all other nodes. The closeness is defined as the inverse of the farness. The more central a node is, the lower its total distance to all other nodes. Closeness can be regarded as a measure of how long it will take to spread information from v to all other nodes sequentially.
(13) **Eigenvector Centrality:** Eigenvector centrality is a centrality index that calculates the centrality of an actor based not only on their connections but also based on the centrality of that actor's connections. Eigenvector centrality measures a node's importance while giving consideration to the importance of its neighbors. For example, a node with 300 relatively unpopular friends on Facebook would have lower eigenvector centrality than someone with 300 very popular friends

Based on the results found in these metrics it becomes easy to understand the network structure for the user-retweeter network in Twitter and Koo. The same measurement metric is also used for the hashtag-hashtag network analysis for Koo.

5.4 Results

Table 3. Results for the various metrics used in user-retweet network.

Hashtag	Node Count	Edge Count	Average Degree	Triangles	Components	Largest Component	Clustering Coefficient	Density
#BlackLivesMatter	1434	1531	2.135286	5	31	1248	0.001578	0.001490
#BlackLivesMatter_update	1315	1356	2.062357	9	62	1045	0.006687	0.001570
#Diwali	8365	9217	2.203706	8	113	7464	0.000508	0.000263
#LGBTQ	1367	1240	1.814192	3	172	248	0.001864	0.001328
#ufo	1026	1074	2.093567	13	62	507	0.006901	0.002043
#whitelivesmatter	101	89	1.762376	0	14	45	0.000000	0.017624
#BoycottBollywood	1293	2844	4.399072	852	16	1223	0.058238	0.003405
#T20WorldCup	3028	3649	2.410172	1	38	2696	0.000771	0.000796
#metoo	1325	1358	2.049811	16	83	526	0.010970	0.001548
#crackerban	1428	1766	2.473389	46	63	1136	0.009687	0.001733
#viratkholi	1750	1874	2.141714	8	146	1071	0.003136	0.001225
#autism	1805	1790	1.983380	14	132	1225	0.006217	0.001099
#FlatEarth	741	791	2.134953	13	15	637	0.011272	0.002885
#StopSpectrum10K	359	568	3.164345	220	5	335	0.097917	0.008839
#covid19(koo)	659	777	2.358118	1	42	567	0.002040	0.003584
#pollution(koo)	98	80	1.632653	0	21	32	0.000000	0.016831
#t20worldcup(koo)	983	1277	2.598169	6	68	848	0.003150	0.002646
#VaccineCentury(koo)	1533	2351	3.067189	50	33	1459	0.012988	0.002002

Table 4. Results for the various metrics used in user-retweet network.

Hashtag	Assortativity	Degree Centrality avg	Betweenness Centrality avg	Closeness Centrality avg	Eigenvector Centrality avg
#BlackLivesMatter	-0.1515	0.0535	0.1990	0.3333	0.2214
#BlackLivesMatter_update	-0.2067	0.0600	0.1848	0.3044	0.2395
#Diwali	-0.1559	0.0566	0.1711	0.2178	-
#LGBTQ	-0.2466	0.0447	0.0133	0.0614	0.2016
#ufo	-0.2603	0.0466	0.0862	0.1218	0.2009
#whitelivesmatter	-0.4331	0.1420	0.0441	0.2631	0.2281
#BoycottBollywood	-0.1819	0.1056	0.1287	0.3993	0.2299
#T20WorldCup	-0.2855	0.0686	0.1357	0.2827	0.1968
#metoo	-0.4438	0.0352	0.0498	0.1070	0.2095
#crackerban	-0.4841	0.0276	0.0716	0.2059	0.2040
#viratkholi	-0.5579	0.0180	0.0418	0.1386	0.2475
#autism	-0.1821	0.0509	0.1382	0.2339	0.2067
#FlatEarth	-0.3360	0.1249	0.2284	0.3893	0.2009
#StopSpectrum10K	-0.4141	0.1749	0.2032	0.4400	0.2886
#covid19(koo)	-0.3280	0.0815	0.1438	0.2843	0.2686
#pollution(koo)	-0.3587	0.1093	0.0238	0.1938	0.2430
#t20worldcup(koo)	-0.2286	0.0464	0.1117	0.2583	0.2619
#VaccineCentury(koo)	-0.3189	0.0926	0.1355	0.3282	0.2597

Table 5. Results for the various metrics used in hashtag-hashtag network.

Hashtag	Node Count	Edge Count avg	Average Degree	Number of Triangles	Clustering Coefficient	Density	Assortativity
#pollution(koo)	539	4788	17.766	29247	0.907	0.033	-0.112
#covid19(koo)	1281	8067	12.594	33482	0.841	0.009	-0.156
#t20worldcup(koo)	1938	15967	16.477	85015	0.833	0.008	-0.159
#VaccineCentury(koo)	570	3136	11.003	11200	0.834	0.0193	-0.193

Since most of the networks consist of a very large component and other small and single node components, we decided to do the analysis for a large components of the networks to get a clearer numbers without the individual nodes disrupting the results. Table 6 shows the results for the largest component.

Table 6. Results for the various metrics used in user-retweet network for largest connected subgraph.

Hashtag	Node Count	Edge Count	Average Degree	Triangles	Clustering Coefficient	Density	Assortativity
#BlackLivesMatter	1248	1373	2.200	5	0.0018	0.0017	-0.1635
#BlackLivesMatter_update	1045	1142	2.1856	9	0.0084	0.00209	-0.2410
#Diwali	7464	8424	2.257	8	0.00056	0.00030	-0.1709
#LGBTQ	248	253	2.0403	1	0.0047	0.0082	-0.5979
#ufo	507	604	2.3826	13	0.0139	0.0047	-0.4928
#whitelivesmatter	45	44	1.955	0	0	0.045	-0.957
#BoycottBollywood	1223	2786	4.556	852	0.0615	0.0037	-0.1870
#T20WorldCup	2696	3351	2.4859	0	0	0.00092	-0.3101
#metoo	526	608	2.3117	11	0.01709	0.0044	-0.6551
#crackerban	1136	1532	2.697	46	0.0121	0.0023	-0.5441
#viratkholi	1071	1297	2.4220	5	0.0018	0.00226	-0.6824
#autism	1225	1325	2.1632	12	0.00681	0.0017	-0.2405
#FlatEarth	637	698	2.191	13	0.0131	0.00344	-0.383
#StopSpectrum10K	335	548	3.2716	220	0.1049	0.0097	-0.4297
#covid19(koo)	567	719	2.5361	1	0.00237	0.00448	-0.3687
#pollution(koo)	32	31	1.9375	0	0	0.0625	-1
#t20worldcup(koo)	848	1189	2.8042	6	0.00365	0.0033	-0.2667
#VaccineCentury(koo)	1459	2304	3.1583	50	0.01364	0.0021	-0.3284

Table 7. Results for the various metrics used in user-retweeter network for largest connected subgraph.

Hashtag	Degree Centrality avg	Betweenness Centrality avg	Closeness Centrality avg	Eigenvector Centrality avg
#BlackLivesMatter	0.0614	0.2628	0.3829	0.2214
#BlackLivesMatter_update	0.0754	0.2928	0.3830	0.2395
#Diwali	0.0634	0.2149	0.2441	-
#LGBTQ	0.1895	0.4092	0.3395	0.2016
#ufo	0.0613	0.3540	0.2467	0.2914
#whitelivesmatter	0.2182	0.2089	0.5979	0.2281
#BoycottBollywood	0.1116	0.1439	0.4222	0.2299
#T20WorldCup	0.0771	0.1712	0.3175	0.1968
#metoo	0.0731	0.3171	0.2698	0.2877
#crackerban	0.0347	0.1133	0.2588	0.2040
#viratkholi	0.0271	0.1118	0.2266	0.2475
#autism	0.0750	0.3003	0.3447	0.2067
#FlatEarth	0.1450	0.3093	0.4530	0.2009
#StopSpectrum10K	0.1874	0.2335	0.4716	0.2886
#covid19(koo)	0.0947	0.1945	0.3305	0.2686
#pollution(koo)	0.2258	0.2000	0.6066	0.2430
#t20worldcup(koo)	0.0538	0.1502	0.2995	0.2619
#VaccineCentury(koo)	0.0973	0.1496	0.3449	0.2597

6 OBSERVATIONS

The Community of the hashtags *#BoycottBollywood* and *#StopSpectrum10K* are very cohesive i.e. many users have retweeted each other's tweets in a cycle., this may suggest a very close community among the users and some real-life relations such as family or friends.

In most of the retweet networks, the number of nodes in the largest component is almost equal to the number of nodes in the network this suggests that these communities are somehow connected by their users. However, the communities of *#ufo*, *# LGBTQ*, *#whitelivesmatter*, *#metoo*, *#pollution(Koo)* do not have well-connected sub-communities since the largest components are less than 50 percent of the size of the network, suggesting many smaller groups within the community.

All these graphs are very low-density graphs, i.e. these are loose-knit networks and are more sparsely connected. Their actors are reachable to one another via a very restricted set of pathways.

The clustering coefficients of *#BoycottBollywood*, *#metoo*, *#StopSpectrum10K* are high with regard to the number of nodes. This suggests that there is a strong desire in the community to forward the information among themselves (transitivity).

The assortativity of all the networks is negative meaning that the correlation between the degree property of the nodes is low i.e. nodes of dissimilar degrees are connected suggesting that in most networks the retweet of some of the major users is more than that of others. This is more prominent in the hashtags *#metoo* and *#viratkohli* suggesting that some celebrity/prominent accounts have the most retweets of the network.

An observation from the highest betweenness centrality measure of each hashtag suggests that in the network for *#BlackLivesMatter*, *#LGBTQ*, *#autism*, the nodes with the highest betweenness centrality coefficient have values higher than 0.7 suggesting that these people have the most influence and are the key players in the network without whom the communication fails. However, the highest betweenness coefficient for hashtags *#pollution(Koo)*, *#whitelivesmatter* is very close to 1 suggesting that the network is completely dependent on the key player in the network.

From the betweenness centrality coefficient, the size of the largest component, and other metrics it can be concluded that the networks for hashtags *#whitelivesmatter*, *#pollution(Koo)* are extremely unsuccessful.

Other observations from all the metrics could be : *#BoycottBollywood* is the most successful hashtag with its high density and node count suggesting that many people support the hashtag with their posts. The clustering coefficient is high suggesting a strong social community. The high triangle count suggests that the hashtag is currently very active and popular among smaller groups of people. The high associativity suggests that the community is not dependent on many high activity accounts and individuals also contribute to the network. The centrality measure of the highest activity accounts is almost equal to each other, suggesting many key players despite the low associativity. Despite that, the centrality measure is not too high suggesting a strong community structure.

Thus it can be concluded that for a good social movement using hashtags the structure of the retweet/rekoo network should be similar to that of *#BoycottBollywood*.

6.1 Hashtag-Hashtag networks

The hashtag-hashtag networks built here are highly connected graphs suggesting that each hashtag is used with many other hashtags in the network. No other useful observation could be taken out from the hashtag-hashtag network since all the networks are similar. This suggests that people use hashtags without any proper pattern or method.

6.2 Full network vs Largest components

For many hashtags like *#BlackLivesMatter*, *#BoycottBollywood*, *#T20WorldCup*, *#StopSpectrum10K*, *#covid19(Koo)*, and *#VaccineCentury(Koo)*, the graph is highly connected making the size of the largest component very close to the total number of nodes. For such tags, we do not see much difference in the Average degree and Assortativity of the largest connected component from the full graph, since the largest component is very similar to the full network. Other hashtags like the *#LGBTQ*, *#ufo*, *#metoo*, *#crackerban*, *#viratkohli*, and *#autism* show a significant change in the Assortativity since the smaller components and individual nodes which formed a significant part of the full network are now removed. Some networks like *#whitelivesmatter* and *#pollution(Koo)* are very small and hence the largest component which is much smaller is almost fully connected, giving very low Assortativity. Due to this, all nodes in these components are connected to a node leaving no neighbor and hence giving a clustering coefficient of zero.

7 CONCLUSION AND FUTURE WORK

From the above analysis, we can conclude that the best network structure was observed in *#BoycottBollywood*. Thus it can be concluded that a hashtag will have more chance of being as successful as the *#BoycottBollywood* if it follows a structure similar to *#BoycottBollywood*. In the future, we can test out the networks for Twitter and check if they are similar to that of Koo. Also, a comparative study of Koo and Twitter based on hashtags can be done. We can also collect more data over a longer period and work with more computational power to find structures among larger graphs which may translate even better in the real world. The above analysis can be improved by introducing more customized metrics like assortativity based on the following of the user, etc.

REFERENCES

[1] Rabeeh Abbasi, Onaiza Maqbool, Mubashar Mushtaq, Naif Aljohani, Ali Daud, Jalal Alowibdi, and Basit Shahzad. 2017. Saving Lives Using Social Media: Analysis of the Role of Twitter for Personal Blood Donation Requests and Dissemination. *Telematics and Informatics* 35 (02 2017). https://doi.org/10.1016/j.tele.2017.01.010

[2] Hameed Al-Qaheri and Soumya Banerjee. 2013. Measuring Homophily in Social Network: Identification of Flow of Inspiring Influence under New Vistas of Evolutionary Dynamics. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Extended Papers from Science and Information Conference 2013* 3, 4 (2013). https://doi.org/10.14569/SpecialIssue.2013.030403

[3] Mark Boukes. 2019. Social network sites and acquiring current affairs knowledge: The impact of Twitter and Facebook usage on learning about the news. *Journal of Information Technology & Politics* 16, 1 (2019), 36–51. https://doi.org/10.1080/19331681.2019.1572568 arXiv:https://doi.org/10.1080/19331681.2019.1572568

[4] Zubaida Jastania, Mohammad Ahtisham Aslam, Rabeeh Ayaz Abbasi, and Kawther Saeedi. 2020. Using Social Network Analysis to Understand Public Discussions: The Case Study of SaudiWomenCanDrive on Twitter. *International Journal of Advanced Computer Science and Applications* 11, 2 (2020). https://doi.org/10.14569/IJACSA.2020.0110230

[5] David Knoke and Song Yang. 2008. *Social Network Analysis*. Vol. 154. https://doi.org/10.1007/978-1-4614-6170-8_362

[6] Daisuke Kuroshima and Tina Tian. 2019. Detecting Public Sentiment of Medicine by Mining Twitter Data. *International Journal of Advanced Computer Science and Applications* 10, 10 (2019). https://doi.org/10.14569/IJACSA.2019.0101001

[7] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. 2004. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America* 101 (04 2004), 2658–63. https://doi.org/10.1073/pnas.0400054101

[8] Joshua Rosenberg. 2019. Exploring the Next Generation Science Standards Chat (NGSSchat) Professional Network on Twitter Through Social Network Analysis. https://doi.org/10.31219/osf.io/uwza6

[9] Lea Vega and Andres Mendez-Vazquez. 2019. Detecting of topic-specific leaders in social networks. *Procedia Computer Science* 151 (01 2019), 1188–1193. https://doi.org/10.1016/j.procs.2019.04.170

OTHER INTERNET RESOURCES

- Oreilly.com - Social Network Analysis
- Geeks for Geeks - Tweepy reference
- Tweepy documentation
- Stack overflow - debugging

A CODE

Codes of the proposed methodology and data collection are provided here.