Network Analysis of Twitter Data: West Bengal Elections 2021

Samidha Verma
MS(Research) Student
IIT Delhi
csy207575@cse.iitd.ac.in

ABSTRACT

General election to the Legislative Assembly of West Bengal for 292 constituencies of the 294 constituencies in West Bengal were held between 27 March to 29 April 2021 in 8 phases[1]. Voting for 2 remaining constituencies was delayed and are scheduled to be held on 16 May 2021. During elections, social media platforms like twitter and facebook are abuzz with opinions and discussions taking place all over India. Thus, twitter provides rich content that can be used for analysis of the general opinion of the population, as well as study the relative popularity of various political parties. This paper covers the results of network analysis study done on follow networks and retweet networks of some popular twitter users that are associated with various political parties contesting in the elections in West Bengal. Further, this paper highlights the correlation between the result of the election and the observations made while conducting this study. The code has been made public[1] for further analysis.

1. INTRODUCTION

Social media is an integral part of people's lives today. Many people communicate their feelings, opinions, thoughts, etc. on a regular basis. When some specific event like elections take place, social media becomes a platform for promotion campaigns of political parties, as well as a media where general people discuss about their political opinions and inclinations. Twitter as a micro-blogging website that is very popular today, receives huge volume of tweets from many users. These tweets and the network of the users who post these tweets becomes a goldmine of data from which many interesting insights can be drawn.

Keeping the above points in mind, in this study we have mined twitter timelines of some key users associated with political parties contesting in the general elections to the Legislative Assembly of West Bengal, 2021. Through this study we will try to find out the popularity of various political parties via the general opinion of users on twitter using hashtags. We would also take a look at the twitter networks in terms of followers, friends and retweeters of the key users, and perform network analysis on them.

[1] `https://github.com/Samidha09/WestBengalElections2021`

Section 2 of the report covers the methods used for data acquisition, follow and retweet networks' construction. Further, the details about the experiments conducted are also covered in this section. In section 3, the results of the experiments and the analysis of the results have been covered. The limitations and possible future work is discussed in section 5 of this work.

2. METHODS

Python programming language has been used for the programming assignment. We have particularly used the python library **tweepy**[2] for accessing the twitter API with ease.

2.1 Data Collection

First, 14 key twitter users affiliated with different political parties were shortlisted. Their twitter handles are '@BJP4Bengal', '@DilipGhoshBJP', '@SuvenduWB', '@AITCofficial', '@MamataOfficial', '@cpimspeak', '@SitaramYechury', '@INCWestBengal', '@aimim_national', '@derekobrienmp', '@DidiKeBolo', '@MahuaMoitra', '@tathagata2', and '@Sujan_Speak'.

2.1.1 Hashtags

The 14 twitter users mentioned above formed our base set. We extracted the top-500 tweets from the timeline of each of the user. In total 7000 tweets were collected. If a tweet was retweeted, we stored it separately in order to use it to build the retweet network. We further, stored data regarding the hashtags mentioned in the tweets, in order to find out the most popular hashtags associated with the West Bengal elections, 2021 on twitter.

2.1.2 Follow Network

Using the twitter API, we extract approximately 40 friends and 40 followers of each of the 14 users shortlisted for the base set. Now we have three sets, the base set of users, the set of their followers, and the set of the users they follow. Let's call these three sets B, C, and A respectively. Through the data mined as mentioned above, edges from set B to set A, and from set C to set B were found. Edges from set B to C, A to B, C to A, and A to C still remained. These relations were found by the method described below.

We further mined followers and friends(or followees) of set A and set C. Let's call them $A_{followers}$, $A_{friends}$, $C_{followers}$, $C_{friends}$.

- Edges from set B to set C: If a user b of set B, was a follower of some user c of set C, then an edge was drawn with source as b and target as c. Such a relation would be found in set $C_{followers}$.
- Edges from set A to set B: If a user a of set A, was a follower of some user b of set B, then an edge was drawn with source as a and target as b. Such a relation would be found in set $A_{friends}$.
- Edges from set C to set A: If a user c from set C is a follower of a user a from set A, then either such a relation would be found in $C_{friends}$ or $A_{followers}$. If such a relation exists, then an edge was drawn with source as c and target as a.
- Edges from set A to set C: If a user a from set A is a follower of a user c from set C, then either such a relation would be found in $C_{followers}$ or $A_{friends}$. If such a relation exists, then an edge was drawn with source as a and target as c.

We mined a total of 29587 edges. Out of all these edges, when following the constraints mentioned above for insertion into a networkx[3] DiGraph, we attained a graph with 979 nodes and 1564 edges.

2.1.3 Retweet Network

Out of the 7000 tweets mined as mentioned in Section 2.1.1, 1911 tweets were such that they had been retweeted by the 14 users in the base set. each tweet out of these 1911, would lead to formation of a single edge in the retweet graph, such that the users in base set would be the source node, and the person whom they retweeted would be the target node. Further, we were left with 5089 tweets which were either originally tweeted or quoted by the users in the base set. We chose 500 most retweeted tweets among these 5089 tweets, and for each of them we collected the twitter screen names of 10 users who retweeted them. Thus, with these, we were able to get retweeters of the users in the base set.

Now, we will have three sets. The base set of users, the users whom the base set users retweeted, and the users who retweeted the tweets of base set users. Let's call these three sets as B, A and C respectively. Through the data mined as mentioned above, edges from set B to set A, and from set C to set B were found. Edges from set B to C, A to B, C to A, and A to C still remained. These relations were found by the method described below.

We further mined users who retweeted tweets posted by users in set A and users whom set A users retweeted. Let's call these sets $A_{retweeters}$ and $A_{retweeted}$ respectively. Similarly, we mined users who retweeted tweets posted by users in set C, and users whom set C users retweeted. Let's call these sets $C_{retweeters}$ and $C_{retweeted}$ respectively.

- Edges from set B to set C: If a user b of set B, was present in $C_{retweeters}$ and hence retweeted a tweet posted by some user c of set C, then an edge was drawn with source as b and target as c.
- Edges from set A to set B: If a user a of set A, retweeted a tweet by some user b of set B, then an edge was drawn with source as a and target as b. Such a relation would be found in the set $A_{retweeted}$.
- Edges from set C to set A: If a tweet by a user a from set A is retweeted by a user c from set C, then either such a relation would be found in $C_{retweeted}$ or $A_{retweeters}$. If such a relation exists, then an edge was drawn with source as c and target as a.
- Edges from set A to set C: If a tweet by a user a from set A is retweets a user c from set C, then either such a relation would be found in $C_{retweeters}$ or $A_{retweeted}$. If such a relation exists, then an edge was drawn with source as a and target as c.

After adding all these edges to a networkx DiGraph[3], following the above constraints, we attained a graph with 2568 nodes and 4402 edges.

Note: In the case of a non-null intersection between set C and set A, i.e., when a user in set B follows some other user and is followed back by that user, it may be possible, that multiple edges in the same direction exist between two users, but the networkx DiGraph[3] handles such a case, and treats those edges as one. This applies to both follow and retweet networks.

2.2 Experiments

2.2.1 Hashtags

In this experiment, we have conducted a study to discover the most popular hashtags on twitter associated with West Bengal elections, 2021. This has been done, in order to get an idea of how closely the popular opinion of twitter data, mirrors the actual truth in terms of popularity of certain political parties contesting in the election as well as the results of the election.

2.2.2 Follow Network and Retweet Network

In this experiment, we have done the network analysis of the follow and retweet networks of the key users associated with political parties contesting in the West Bengal elections, 2021 as described before. We find out community structure in these networks using Louvain's community detection algorithm[4] and Girvan Newman method[5]. We also study the in and out degree distributions of both the networks. Top-10 central users in each network are reported on the basis of degree, betweenness, closeness and eigen vector centrality measures. Network properties like number of strongly and weakly connected components, average clustering coefficient, degree assortativity are also looked into. Further we have also reported the top-10 users in terms of VoteRank[7] and PageRank[6].

3. RESULTS & ANALYSIS

3.1 Hashtags

- The top-10 hashtags in tweets were 'MegaPublicRally', 'TMC200Paar', 'LokhhoSonarBangla', 'Vote4AsolPoriborton', 'VoteForTMC', 'Nandigram', 'ModiMadeDisaster', 'COVID19', 'EbarBJPEbarSonarBangla'. The frequency distribution of hashtags followed a power law distribution. There were less hashtags with greater frequency as compared to hashtags with lower frequency. The frequency distribution of tweets has been shown

in figure 1. We have removed hashtags with frequency less than 3, in order for ease of plotting graph.

- We can observe that hashtags supporting two major contestents, BJP and AITC are the most popular ones.
- Another observation is that hashtags supporting AITMC are more popular, which also is correct, considering the fact that AITMC won the West Bengal elections, 2021.
- The second wave of Covid19 pandemic led to increase in unrest among the Indian public against BJP government. This factor was also observed in this study. At first, before the disaster associated with Covid19 second wave spread, the hashtags supporting BJP were most popular. However, as evident from a particular hashtag 'ModiMadeDisaster', as well as the issues associated with the failure of healthcare systems in the face of second wave of Covid19 that the general public is facing, along with the centre and state governments blaming each other, or denying the issues altogether, has lead to widespread unpopularity of BJP government. Thus, by observation of the social media hashtags, as well as the ground truth of election results of AITMC winning 213 seats and sweeping the election results in its favour, it may be possible that Covid19 second wave is a major reason for BJP losing the election in West Bengal.

3.2 Follow Network

3.2.1 *Number of strongly and weakly connected components*

Number of strongly and weakly connected components are **780** and **1** respectively. The large number of strongly connected components can be attributed to the fact that there were a large number of users with zero in-degree and out-degree due to the manner in which the network was constructed. Further, several small strongly connected components may have occurred, due to the fact that people with similar opinions or as in this case similar political opinions are likely to follow each other. The fact that there is only one weakly connected component also highlights the fact, that some neutral users exist that follow, opposing parties as well.

3.2.2 *In-degree and Out-degree distributions*

The in-degree and out-degree distributions are shown in figure 2 and 3. We have removed degree 0, and those degrees that had frequency less than 2, in order to plot a more interpretable graph. However, as expected in real world graphs, power law distribution was observed in both the degree distributions. Nodes with low degree are high in number and high degree nodes are low in number.

3.2.3 *Central Nodes*

Top-10 central nodes computed based on various centrality measures have been reported in **table 1**. The centrality measure appropriate for a particular network depends upon which aspect of the network topology it captures. Some key observations are:

- Degree and betweeness centrality measures are able to capture the leaders, or the relevant users from the base set. This is because degree centrality captures the local popularilty of a particular vertex in terms of the number of connections. Clearly, popular political party or leader would have higher degree centrality.
- Closeness centrality represents the average distance, or average shortest path, to all other vertices in the network. The idea is that a central vertex will be closer on average than other, less central vertices. Closeness centrality is not able to capture the leaders in the base set of users, rather captures users like 'rahtrapatibhvn' as it is likely to be neutral and connected to most political parties and hence closer on average to all nodes in the network.
- Betweenness centrality captures the ability of a vertex to control the flow of communication and it indicates how many times a vertex is located on the shortest path between two other vertices. It can differentiate

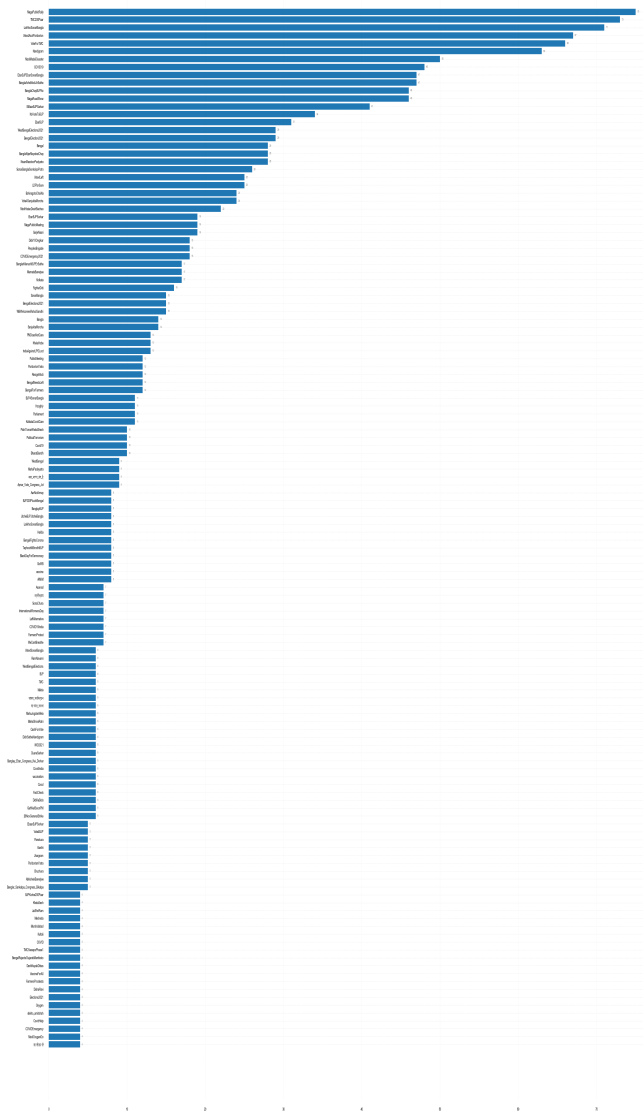

Figure 1: Frequency Distribution of Hashtags: West Bengal Elections, 2021

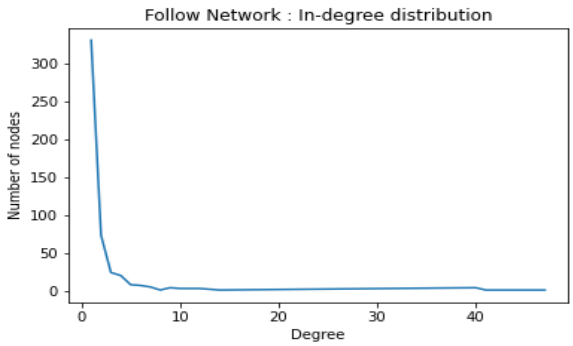

Figure 2: In-degree distribution of Follow Network

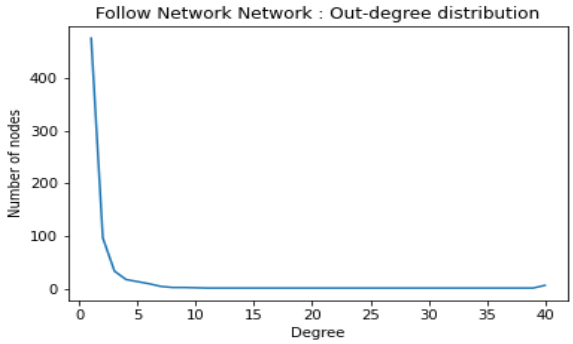

Figure 3: Out-degree distribution of Follow Network

popular leaders and political parties probably because individuals who lie on communication paths can exert a greater influence on that communication.

- Eigen vector centrality is not able to distinguish between the political parties or leaders and their followers as can be seen from **table 1**. This may be because a political party or leader is likely to be followed by general public as also observed, whose centrality would actually be low. Since eigen vector centrality of a node depends upon the centrality of its neighbours, therefore, the base sets users or political leaders/parties are likely to have lower eigen vector centrality.
- PageRank[6] and VoteRank[7] are able to distinguish base set users based upon their popularity as well. However, as seen in the results VoteRank is more accurate, since it computes a ranking of the nodes based on a voting scheme. With VoteRank, all nodes vote for each of its in-neighbours and the node with the highest votes is elected iteratively. The voting ability of out-neighbors of elected nodes is decreased in subsequent turns. The concept of VoteRank is more aligned with the election process, and hence, ranks AITCofficial at the top, which was actually the winning political party in the elections, followed by BJP4Bengal which was the second most voted for political party having won 77 seats in the West Bengal Legislative Assembly.

3.2.4 Average Clustering Coefficient

The average clustering coefficient of the follow network is **0.0588**. The low value of clustering coefficient is intuitive since, if the definition of users A, B and C from section 2.1.2 are taken, then although edges are likely to exist between set C and set B, and from set B to set A, however, it is less likely that there will be an edge between set A and set C, because most users in set C would be general users, while most users in set A would be popular political leaders or parties, and are less likely to follow the general users. Thus, most of the triplets are likely to be open than close.

3.2.5 Degree Assortativity

Assortativity is a preference for a network's nodes to attach to others that are similar to it in some way. Since the follow network's nodes have no attributes, the degree assortativity actually measures whether high degree nodes are likely to connect to other high degree nodes or not. The degree assortativity of the follow network is **-0.597**. This is actually intuitive since, low degree nodes, that is the general users, are likely to follow high degree nodes, which are the political leaders or parties. Even though high degree nodes are likely to follow other high degree nodes in this scenario, still due to the power law distribution, that is larger number of nodes with low degree, the network will be **disassortative** as observed by the negative value of degree assortativity.

3.2.6 Community Structure

Community detection techniques are useful for social media algorithms to discover people with common interests and keep them tightly connected. It would be useful to find out communities in the follow network. This may also help in classifying users as well on the basis of their political inclination as left wing or right wing supporters, and for other use cases as well. We have used the community API of networkx[3] for community detection.

Louvain's Algorithm Louvain[4] community detection algorithm was proposed as a fast community unfolding method for large networks. This approach is based on modularity, which tries to maximize the difference between the actual number of edges in a community and the expected number of edges in the community. However optimizing modularity in a network is NP-hard, therefore have to use heuristics. Louvain algorithm is divided into iteratively repeating two phases:

- Local moving of nodes
- Aggregation of the network

The limitation of the networkx[3] louvain's community detection algorithm is that it doesn't work for directed graphs, so we had to convert are directed graph, to an undirected one to get the results. As a result of this although the correct communities have been detected, but a better partition could have been found by using the actual graph. A total of 12 communities were detected, and the modularity of the partition was **0.799**. The communities[2] can be seen in Figure 4.

[2]Communities in follow network with labels: `https://github.com/Samidha09/WestBengalElections2021/blob/main/follow_network_communities.png`

Table 1: Top-10 Central Node(Users) in Follow Network

Rank	Degree	Betweeness	Closeness	Eigen Vector	VoteRank	PageRank
1	AITCofficial	tathagata2	rashtrapatibhvn	hiran_chatterji	AITCofficial	ClassStressed
2	tathagata2	AITCofficial	INCWestBengal	krishmenon11	BJP4Bengal	tathagata2
3	BJP4Bengal	INCWestBengal	jdhankhar1	KhanSaumitra	SitaramYechury	AITCofficial
4	DilipGhoshBJP	Sujan_Speak	mayukhrghosh	RajibBanerjeeWB	INCWestBengal	MamataOfficial
5	INCWestBengal	MamataOfficial	rajnathsingh	NisithPramanik	derekobrienmp	DilipGhoshBJP
6	derekobrienmp	cpimspeak	PMOIndia	iamharsh1000	tathagata2	INCWestBengal
7	SitaramYechury	gagandeepseven	tathagata2	amit0360	cpimspeak	SuvenduWB
8	Sujan_Speak	UtpalAdhikary14	Gautamdebmic	Andri54527441	aimim_national	Sujan_Speak
9	cpimspeak	ZartabKhan141	Indrani39664132	RRRMovieeee	Sujan_Speak	derekobrienmp
10	aimim_national	SuvenduWB	AITCofficial	Saiyada10803839	DilipGhoshBJP	BJP4Bengal

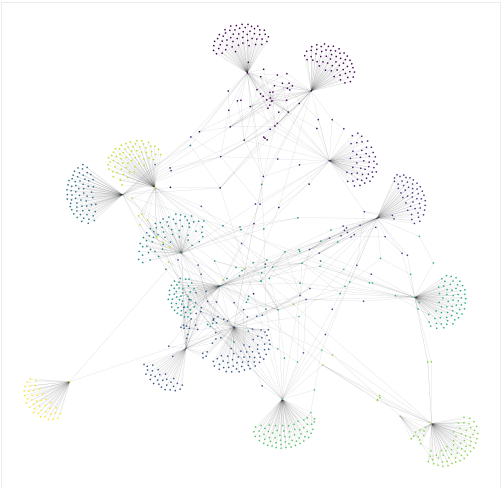

Figure 4: Community Structure: Follow Network
Communities can be distinguished by a colour of nodes

Girvan Newman Algorithm In the Girvan-Newman[5] algorithm, the communities in a graph are discovered by iteratively removing the edges of the graph, based on the edge betweenness centrality value. The edge with the highest edge betweenness is removed first.

This method worked with the original directed graph, and hence by observation was better than louvain's algorithm at detecting communities. There was no direct way of calculating modularity which was a weakness of this method. However, by observation the best partition was when 6 communities were detected with respect to classifying the base set of users in a single community based on the political parties they are a part of. The exact communities can be seen using the FollowNetwork.ipynb file uploaded in the repository.

3.3 Retweet Network

The retweet and the follow graphs are similar in nature with retweeted users likely to be affiliated with a political party, or being political leaders, and the retweeters likely to be general users. Therefore the explanations for results in the follow network may also apply in case of retweet network.

Note: Download the file and zoom in to see the labels clearly.

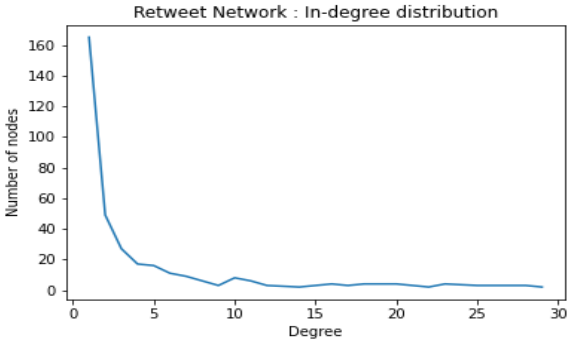

Figure 5: In-degree distribution of Retweet Network

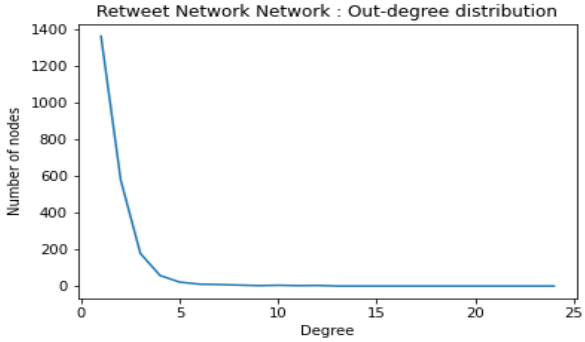

Figure 6: Out-degree distribution of Retweet Network

3.3.1 *Number of strongly and weakly connected components*

Number of strongly and weakly connected components are **2528** and **1** respectively.

3.3.2 *In-degree and Out-degree distributions*

The in-degree and out-degree distributions are shown in figure 5 and 6. Just like the case of follow network(Section 3.2.2) zero degree, and degrees with frequency less than 2 have been removed. Similarly, a power law distribution was observed in both the degree distributions.

3.3.3 *Central Nodes*

Top-10 central nodes computed based on various centrality measures have been reported in **table 2**. Some observations are:

- Degree, closeness and betweenness centrality measures and VoteRank are more accurate in terms of the ranking of political leaders and contesting political parties.
- Closeness, eigen vector centrality and PangeRank give many media houses higher ranks, and social media influencers like Kunal Kamra with left wing ideology, a higher rank. These align with the left wing ideology of the AITC political party as well. Since most retweeted tweets were those of members of AITC political party, like MahuaMoitra and MamataOfficial, and they might retweet these media houses, therefore, the media houses might have higher eigen vector and PageRank centrality. Further, since many users follow media houses and social influencers in general, it is intuitive that their closeness centrality is also high.

3.3.4 Average Clustering Coefficient

The average clustering coefficient of the retweet network is **0.0642**. The reason for this may be the same as mentioned in section 3.2.4.

3.3.5 Degree Assortativity

Degree assortativity of the retweet network is **-0.296** and hence the network is **disassortative**. The reason for this may be the same as mentioned in section 3.2.5.

3.3.6 Community Structure

The method used for detecting communities in the retweet network was the same as mentioned in Section 3.2.6 and the limitations mentioned there also apply for the retweet network.

Louvain's Algorithm A total of **13** communities were detected, and the modularity of the partition was **0.591**. The communities can be seen in Figure 7. It can be observed that basically there is a community associated with each user of the base set. For a detailed look at the community structure, the reader may refer to the community structure plot with labels uploaded in the github repository[3]. The largest community was associated with the user 'MahuaMoitra' whose tweets were retweeted the most.

Girvan Newman Algorithm As mentioned in section 3.2.6, Girvan Newman method of networkx was able to detect communities using the original graph and gave a good division of communities. However, since the retweet network was larger than the follow network it took a significant amount of time to run the algorithm and it would not scale for larger graphs. The exact communities can be seen using the RetweetNetwork.ipynb file uploaded in the repository.

4. LIMITATION AND FUTURE WORK

Due to rate limit constraints, sufficient amount of data could not be collected for better understanding of the network.

[3]Communities in retweet network with labels: `https://github.com/Samidha09/WestBengalElections2021/blob/main/retweet_network_communities.png`
Note: Download the file and zoom in to see the labels clearly.

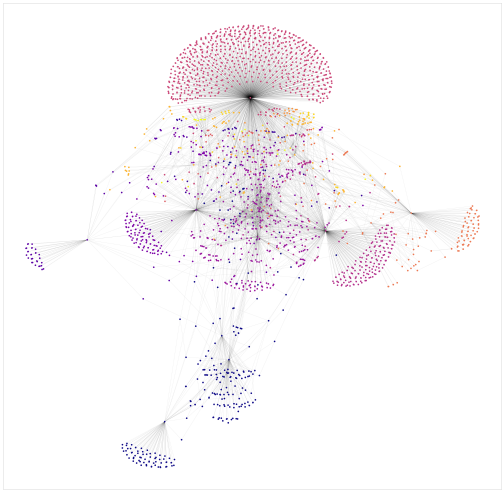

Figure 7: Community Structure: Retweet Network
Communities can be distinguished by a colour of nodes

The base set of users could have been larger, with key members and political parties contesting in all constituencies. Also, since the retweeted tweets were large in number, only the 500 most retweeted tweets could be used for expanding the network due to rate limits. This introduced a bias in the retweet network. Therefore for future work, the analysis could be conducted with larger volume of data to get more intuitive results. Further analysis of tweet content can also be done to find out the sentiment of the users towards each contesting political party, and through analysis of twitters we may also find out the reasons behind the results of the election. Another, option could be to do a temporal study of how the networks and tweets evolved with key events occurring in West Bengal like Covid19 second wave, cyclone Amphan, citizenship, immigration and refugee related issues.

5. CONCLUSION

The study of hashtags and the network analysis of follow and retweet networks of twitter accounts of key political parties and leaders actually mirrored the ground reality of the election results. One of the most popular hashtags like 'TMC200Paar' also gave a premonition to the fact that AITMC won 212 seats in the Legislative Assembly of West Bengal. Further, the two most popular parties in this election were AITMC and BJP, which was also evident from the experiments. Other political parties like INC, CPI(M), AIMIM, etc were not as popular as the other two, and actually won 0 seats. This also was evident from the hashtags, and follow, retweet networks. Network analysis results showed power law degree distributions which actually occur in social media networks and in real world graphs. Further this had an impact leading to low average clustering coefficient and disassortative networks. Finally, a different centrality measure called VoteRank which aligns more with the election

Table 2: Top-10 Central Node(Users) in Retweet Network

Rank	Degree	Betweeness	Closeness	Eigen Vector	VoteRank	PageRank
1	MahuaMoitra	derekobrienmp	MahuaMoitra	aimim_national	tathagata2	MahuaMoitra
2	MamataOfficial	AITCofficial	MamataOfficial	ANI	derekobrienmp	MamataOfficial
3	derekobrienmp	DidiKeBolo	ANI	Zee_Hindustan	DidiKeBolo	kunalkamra88
4	AITCofficial	MahuaMoitra	ndtv	aajtak	aimim_national	derekobrienmp
5	SitaramYechury	MamataOfficial	kunalkamra88	CNNnews18	Sujan_Speak	ndtv
6	DidiKeBolo	DilipGhoshBJP	BrutIndia	AimimZuber	DilipGhoshBJP	ANI
7	tathagata2	BJP4Bengal	VTankha	asadowaisi	INCWestBengal	BrutIndia
8	SuvenduWB	SuvenduWB	frontline_india	syedasimwaqar	cpimspeak	ThePrintIndia
9	DilipGhoshBJP	aimim_national	ThePrintIndia	MIRZARahmathBa6	SuvenduWB	VTankha
10	aimim_national	abhishekaitc	HomeBengal	MIMBahadurpura	Subhasi83803995	frontline_india

methodology in India was proposed and it actually predicted central users or key political parties/leaders well in both follow and retweet networks.

6. ACKNOWLEDGEMENTS

I would like to thank Professor Abhijnan Chakroborty, IIT Delhi for giving us the opportunity to learn and dig deeper into Social Computing with this programming assignment and also for the constant guidance he gave whenever any doubts arose. I would also like to thank Twitter for providing access to its data, by giving me permission to use a developer account.

References

[1] https://en.wikipedia.org/wiki/2021_West_Bengal_Legislative_Assembly_election#cite_note-3.

[2] https://www.tweepy.org/.

[3] https://networkx.org/.

[4] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

[5] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826. DOI: 10.1073/pnas.122653799. URL: http://dx.doi.org/10.1073/pnas.122653799.

[6] Lawrence Page et al. "The PageRank Citation Ranking: Bringing Order to the Web". In: (1999).

[7] Jian-Xiong Zhang et al. "Identifying a set of influential spreaders in complex networks". In: *CoRR* abs/1602.00070 (2016). arXiv: 1602.00070. URL: http://arxiv.org/abs/1602.00070.