A Project Report

On

Hashtag Analysis of the Movie
The Kerala Story
And
Understanding the Network Structure

By

Naaga Varshini M
Research Scholar
Computer Science and Engineering
Sona College of Technology

Under the Supervision of

Dr. Yayati Gupta
Professor
Mahindra University, Hyderabad

NPTEL IIT MADRAS

SOCIAL NETWORKS
(4 Weeks Internship)

Hashtag Analysis of the Movie The Kerala Story and Understanding The Network Structure

Abstract

Social Media has become a place for everyone to share their opinions and spread information across the world. Hashtags in Social Media are a very important tool to diffuse any information to make it viral. Twitter is a Social Media with a large number of users to share their opinion on any news. We have used this tool to understand the network of people during the release of the movie The Kerala Story in India which faced a lot of issues that even before the movie's release ran into a controversy. It also led to a ban in some states of India. Some people supported the movie and some did not. In this paper, we collected the Twitter hashtag about this movie and how it spread across social media over time, and how the network of this hashtag and its co-occurring hashtags have been evolving since the movie's release which has led to an outbreak in the country among the people. In this project, we have detected the most retweeted Twitter account @KreatelyMedia and the network of that account and found they were the supporters of the movie and also the various global properties of the network.

Keywords – Social Media, Twitter, hashtags, information diffusion.

1. Introduction

As nowadays people spend most of their time on social media and they use these social media platforms to share their opinions on different bases. Twitter is one of the most popular social networks in which people share their opinions and views on different issues that are happening in the world. In this people share their opinions and anybody retweets the tweets and shares their individual opinion on that. This leads to the spread of any important issue happening in the world. So we decided to look up the hashtags of the movie The Kerala Story since it was a hot topic in our country India which also led to an outbreak in the country because of the portrayal of Islam. In this paper, we are going to look at the various properties of the movie network knowing how people supported or opposed the movie using the Twitter co-occurrence network and also by using the retweet network. This shows how the network changes as the days go by and how the information becomes viral on Twitter using hashtags. Firstly we can focus on some of the important concepts of the social network structure and then follow the methodology used and the results obtained from the structure. We have collected the dataset from the Network Tool Osome by filtering the dates of the movie release.

2. Literature Survey

Social Media is a place where people can speak about their own opinion. Studying and analyzing the social media network will help us to find the patterns that occur in the network by using graph theory. Researchers have done many properties of network density, modularity centrality analysis, and community detection[4].

Ladisliv Pilar proposed a Social Media Analysis based on hashtag research SMAHR was developed previously to examine and explore information diffusion through hashtags. They developed a framework to analyze the social media network using hashtags and provided a software module for working with hashtags [2].

Ying Xiong proposed research during the Metoo movement SMOs played a crucial role by analyzing the hashtag. A study was done on how social movement organizations use hashtags to involve in the Metoo movement, semantic network analysis was done on this issue[1].

Sifan Xu and Alvin Zhou proposed research by modeling the network using more than 1000000 tweets to examine the Twitter discourses and have done network analysis and investigated the homophily[3].

Twitter hashtag research on Saudi women who can drive cars has been done using Twitter hashtag using 2 million datasets using hashtags, retweet mentions, and co-mention networks. This helps in identifying the most influential people[4].

Ladisliv Pilaf proposed a study that examines the farmer's market. They collected the hashtags related to farmers' markets from Instagram for 1 day. This provides the customer's values and behavior in farmer marketing[5].

Proposed research on the hashtag of Indonesian COVID-19 tweets and performed Social network analysis. They collected the network dataset of 5000000 public tweets. They made some centrality measures, closeness, etc., and selected the hashtags with more scores[6].

3. Methodology

3.1 Data Extraction

Data is extracted from the Osome which is a Network Tool that contains the Twitter hashtag dataset. We have collected the data for the co-occurrence network specifically during the time of movie release which is from 2nd May 2023 to 8th May 2023. Then after finding the most influential hashtags, we made a list and collected the retweets and quotes network from the Osome Tool daily and made the analysis on those network structures. After performing this analysis we collected the retweet network dataset for these influential hashtags from 2nd May to 20th May to find the most retweeted account on Twitter.

3.2 Network Model

Here the network model represents how the network looks and nodes represent the hashtags and the edges represent the tweets in which the two hashtags co-occur[6].

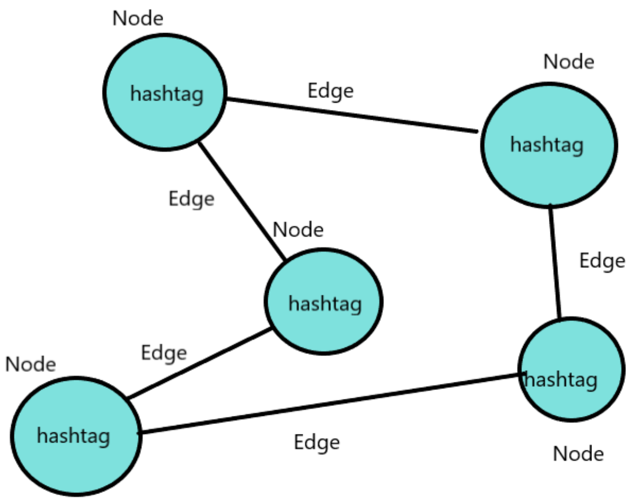


Fig 1: Network Model

In this paper we have used two types of network models one is the co-occurrence network for the hashtag #thekeralastory and the other one is the retweet and quotes network for the most influential hashtag found from this #thekeralastory hashtag.

3.3 Network Analysis Methods

3.3.1 Degree Centrality

Degree Centrality is used to measure the importance. The node's degree refers to the number of links that are connected to it i.e., the number of edges connected to the node i.

$$\mathit{Centrality}_{\text{degree}}(v) = d_V /(| N | -1)$$

where d_V is the Degree of the node v and N is the set of all nodes of the graph.

3.3.2 Closeness Centrality

Closeness Centrality is the measure of the average shortest distance from each node to each other node. It is calculated as the sum of the path length from a given node to all the other nodes.

$$\mathit{Centrality}_{\text{closeness}}(v) = (| R(v) | / | N | - 1) * (| R(v) | / |\Sigma_{u \varepsilon R(v)} d(v,u))$$

where R(v) is the set of all nodes v can reach.

3.3.3 Betweenness Centrality

Betweenness Centrality is the measure to which a node lies on paths between other nodes. Nodes with high betweenness will be having a higher influence and they control the information passing over the network.

$$\mathit{Centrality}_{\text{betweenness}}(v) = \Sigma_{s,t\ \varepsilon\ N} (\sigma_{s,t}(v) / \sigma_{s,t})$$

where $\sigma_{s,t}$ is the number of shortest paths between nodes s and t. $\sigma_{s,t}(v)$ is the number of shortest paths between nodes and t that pass through v.

3.3.4 Transitivity

Transitivity is also called the clustering coefficient of a network and it is the measure of the likeliness of the node to cluster together. When the transitivity is higher it indicates that the network contains the communities of nodes densely connected internally.

$$T = \frac{3 * \text{number of triangles in the network}}{\text{Number of connected triples of nodes in the network}}$$

3.3.5 Modularity

Modularity is the measure of the structure of the network that measures the strength of the division of a network also called communities. Networks with high modularity have dense connections between nodes in different modules.

$$Q= 1/2m\ \Sigma_{ij}(A_{ij} - \gamma\ k_i k_j /2m)\ \delta(c_i ,c_j)$$

Where m is the number of edges, A is the adjacency matrix of G, k_i is the degree if I, γ is the resolution parameter, and $\delta(c_i,c_j)$ is 1 if i and j are in the same community else 0.

4. Experiments And Results

We have used Python and the NetworkX library for the implementation of Social Network Analysis (SNA). For the hashtag #thekeralastory, we have made the following Centrality Analysis that includes degree centrality, closeness centrality, and betweenness centrality and found out the most influential node. Then after that using the top 5 most influential node i.e., hashtags we made the modularity and transitivity analysis on the retweet and quotes network for a large dataset. The co-occurrence network of #thekeralastory consists of 2952 nodes and 18390 edges and it is the dataset of 1 week from 2nd May 2023 to 8th May 2023. After finding the most influential hashtags from this we collected the retweet network from 2nd May to 20th May and found out the most retweeted Twitter account.

4.1 Data Profiling

We have collected the tweets based on the Kerala Story movie keywords. We selected the dates during the time of the movie release and after the movie release to find the most influential hashtags that play a very important role in spreading the movie on Twitter and also to find the most retweeted network using Social Network Analysis (SNA).

4.2 Co-Occurrence Network

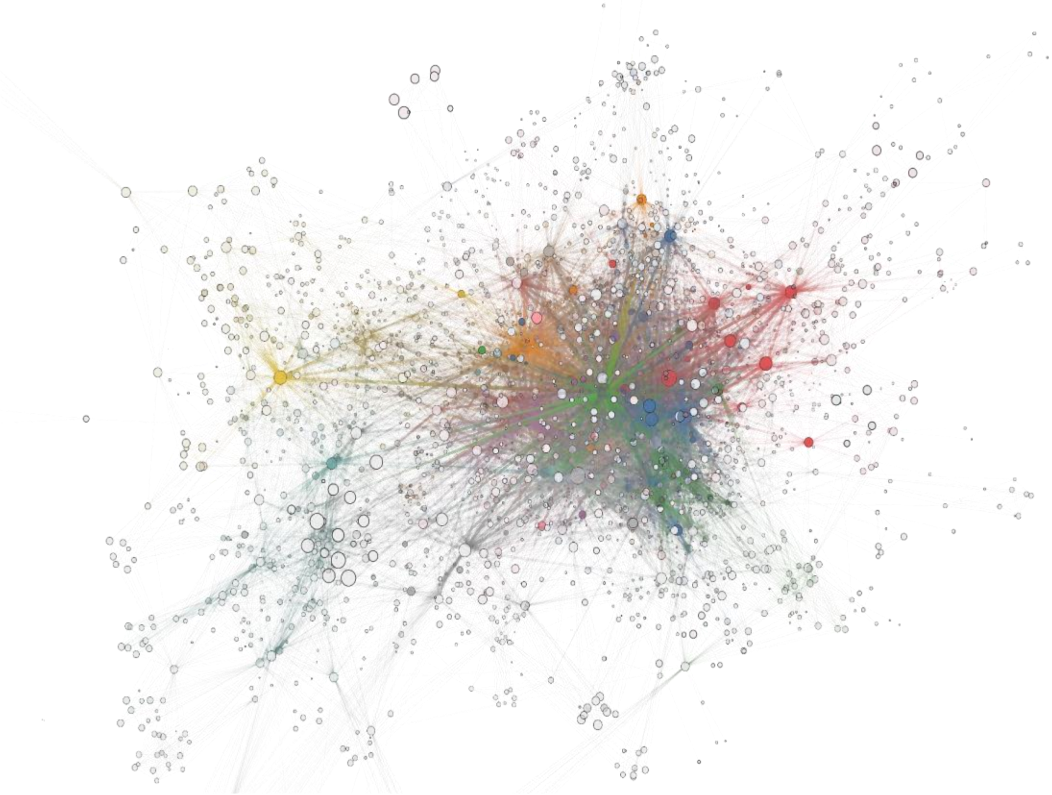


Fig 2. Network of the hashtag #thekeralastory

4.3 Centrality Measures

We have made the centrality measure of Twitter data about the movie The Kerala Story. This centrality measure includes degree centrality, closeness centrality, betweenness centrality, and degree distribution[6].

4.3.1 Degree Centrality

Degree Centrality is a measure of how many connections the node has in it.

Table 1. Degree Centrality

S. no	**Hashtags**	**Degree**
1.	#thekeralastoryreview	0.13859
2.	#keralastory	0.11962
3.	#adahsharma	0.11589
4.	#kerala	0.11555
5.	#india	0.07387

Based on these hashtags we can see that people use these hashtags often in uploading their tweets.

4.3.2 Closeness Centrality

Closeness Centrality is a measure of how much a hashtag is close to all the other remaining hashtags in the network.

Table 2. Closeness Centrality

S.no	**Hashtags**	**Closeness**
1.	#thekeralastoryreview	0.46715
2.	#adahsharma	0.46146
3.	#keralastory	0.45485
4.	#kerala	0.44370
5.	#india	0.43421

In table 2 we can observe that #adahsharma has been moved to second place because it is close to the remaining hashtags in the network.

4.3.3 Betweenness Centrality

Betweenness Centrality is used to measure how often the nodes pass a node to reach a particular node in the network. This shows the hashtag as a bridge in connecting the interactions in the network.

Table 3. Betweenness Centrality

S.no	Hashtags	Betweenness
1.	#thekeralastoryreview	0.10635
2.	#adahsharma	0.09141
3.	#keralastory	0.07750
4.	#kerala	0.06439
5.	#thekashmirfiles	0.05077

4.4 Degree Distribution

In degree distribution we can see which node has the highest degree and the plot on the degree distribution is displayed in Fig 3.

Table 4. Degree Distribution

S.no	Hashtags	Degree
1.	#thekeralastoryreview	409
2.	#keralastory	353
3.	#adahsharma	342
4.	#kerala	341
5.	#india	218

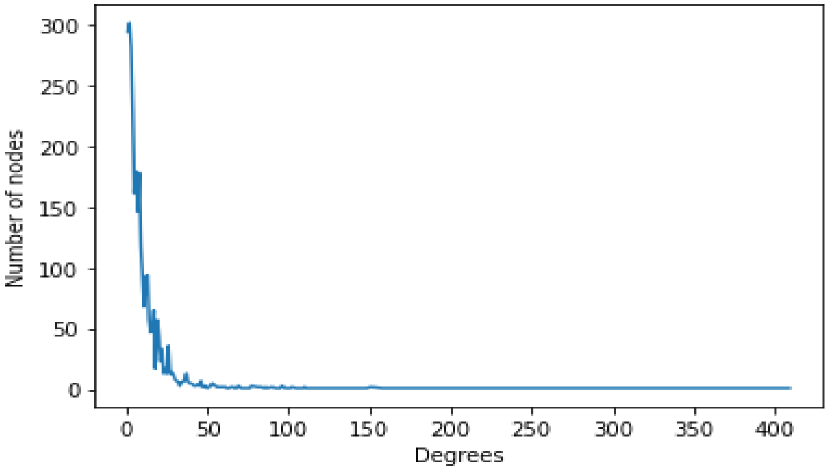


Fig 3. Degree Distribution of nodes

4.5 ReTweet Network

From the above observations, we found that these five hashtags #thekeralastoryreview, #keralastory, #adahsharma, #kerala, and #india are the most influential [4]. Next using these hashtags we moved to the retweet network to proceed with the further analysis. Here in this retweet network, we have collected data on a single-day basis from 2 -May-2023 to 20 May 2023 to analyze several other properties that include the number of nodes, Number of edges, transitivity, modularity, number of communities, and density of graph. After analyzing all these properties daily we have plotted their variation over time. We have plotted all these properties against the date to know on which day the property was in peak. After this, by combining all the single-day datasets we got a large network with 39844 nodes and 64672 edges and found the most retweeted Twitter account.

4.5.1 Number Of Nodes

The below fig.4 shows the variation in the change of nodes between 2 May 2023 to 20 May 2023. Here we can see that the number of nodes peaked on 5 May 2023.

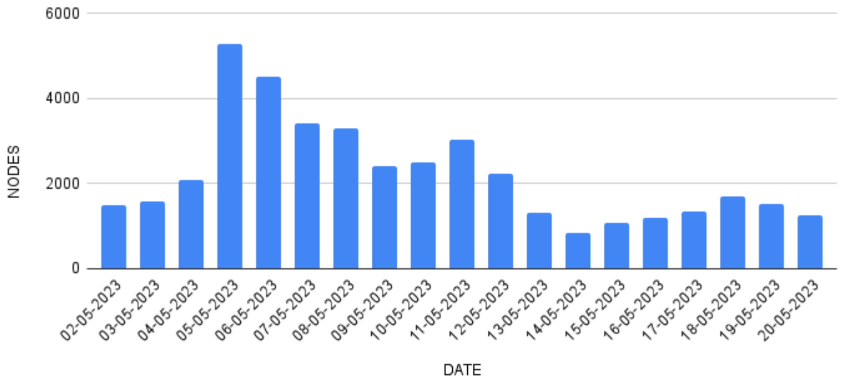


Fig 4. Variation in the number of nodes

The number of nodes on 5-May-2023 is 5289 this is because it was the day on which the movie was released.

4.5.2 Number of Edges

Next, we will look at the changes in the number of edges from 2nd May 2023 to 20th May 2023. Fig 5. Shows the variation in the number of edges

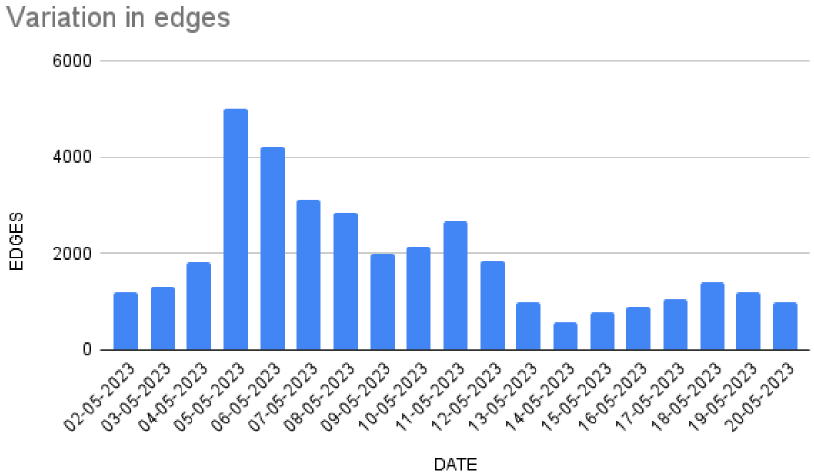


Fig 5. Variation in the number of edges

Here on May 5, 2023, it has the highest edges it has 5031 edges and it peaked on the chart

4.5.3 Transitivity

Transitivity means the clustering coefficient of the network. Figure 6 below shows the variation in transitivity as the days go on.

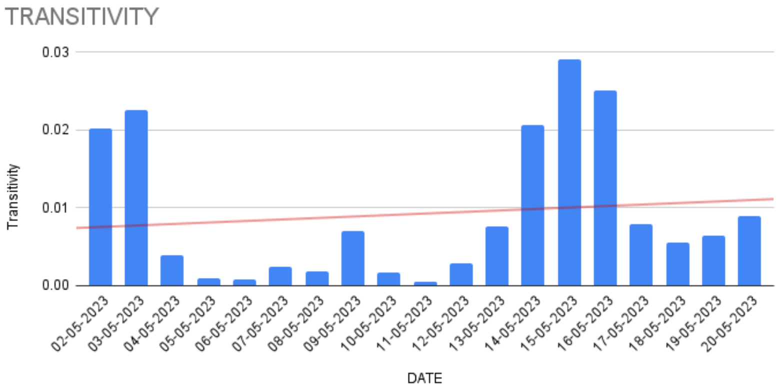


Fig 6. Variation in transitivity

Here the transitivity of the network peaked on 15th May 2023 the value is 0.0290936 which is the highest transitivity.

4.5.4. Density

The density of a network is defined as the fraction of edges present over all possible edges. Figure 7 shows the variation in the density of the network over time.

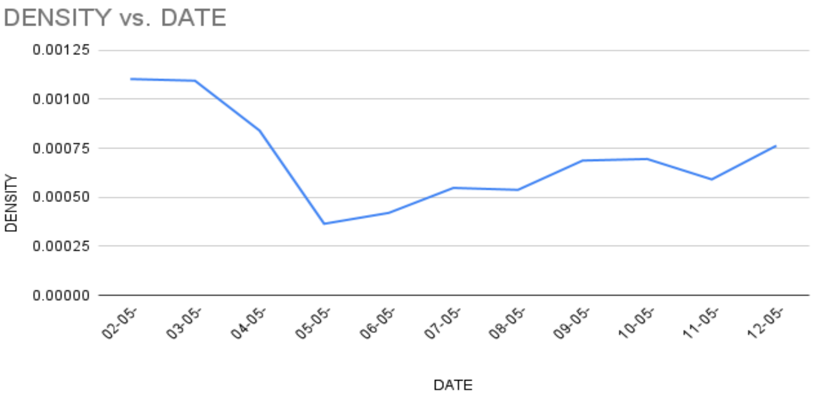


Fig7. Variation in Density

4.5.5 Modularity

The Modularity of the network means the measure of the strength of the division of the network. Figure 8 shows that the modularity has been decreasing and increasing over time.

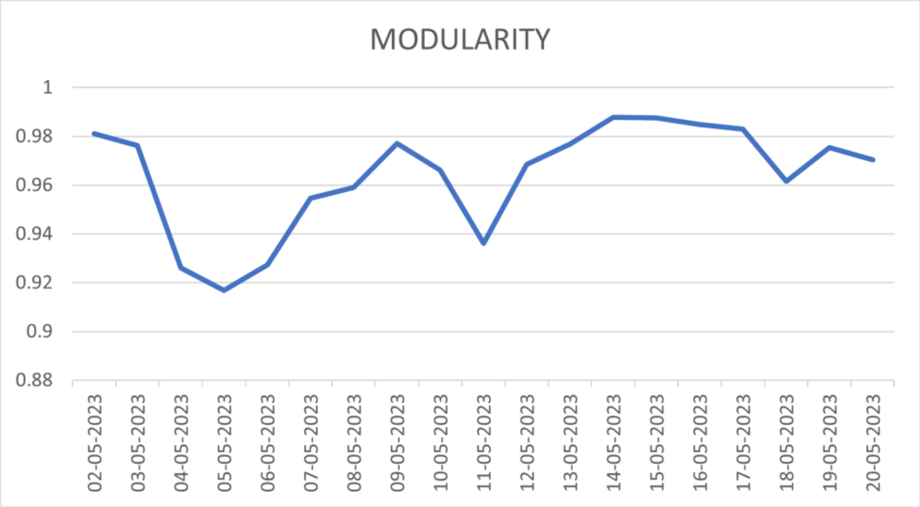


Fig 8. Variation in Modularity

4.5.6 Number of Communities

Communities in the network define the subset of nodes in the graph such that connections in the network are denser than the connections in the rest of the network.

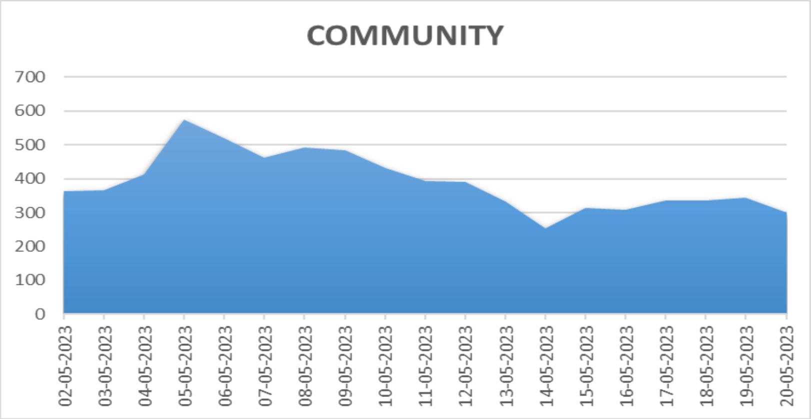


Fig 9. Number of Communities

4.6 Most Retweeted network detection

Now we have moved to the next phase to detect the strong network that is the most retweeted Twitter account's network. We downloaded the retweet network from 2nd May 2023 to 20th May 2023. The dataset contained 39,844 nodes and 34,372 edges with 5172 connected components.

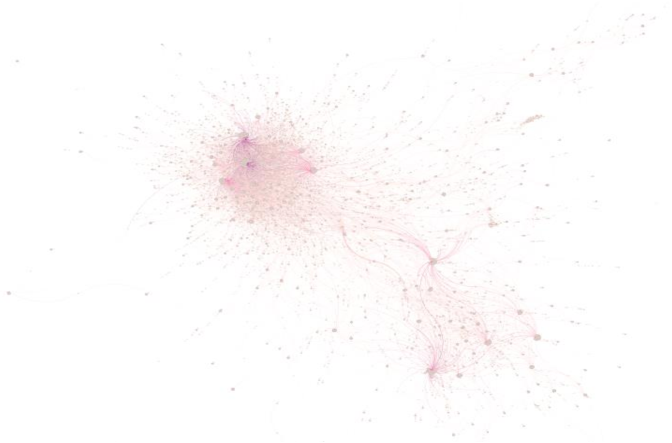
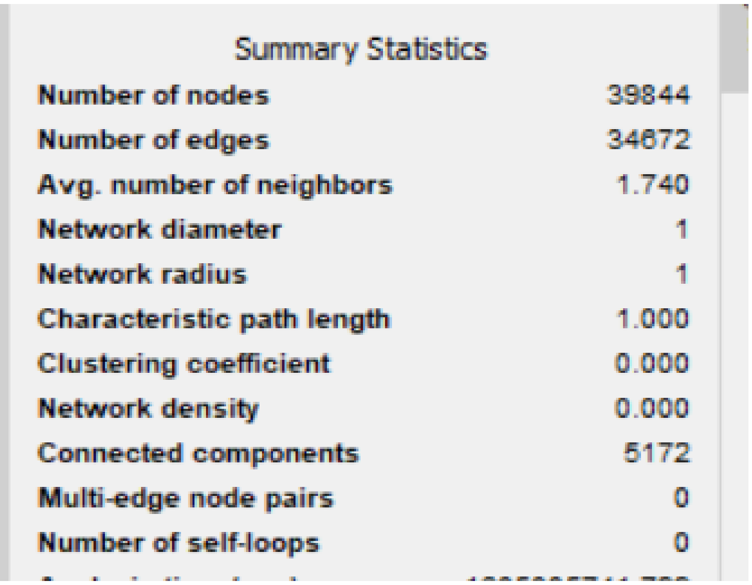


Fig 10. Retweet Network



Summary Statistics	
Number of nodes	39844
Number of edges	34672
Avg. number of neighbors	1.740
Network diameter	1
Network radius	1
Characteristic path length	1.000
Clustering coefficient	0.000
Network density	0.000
Connected components	5172
Multi-edge node pairs	0
Number of self-loops	0

Fig 11. Summary Statistics of the large dataset

Figure 11 shows the properties of the network. From this, we have analyzed the network with 1283 nodes and 1282 edges and this network has the ma number of nodes and edges as the clusters of the retweet network.

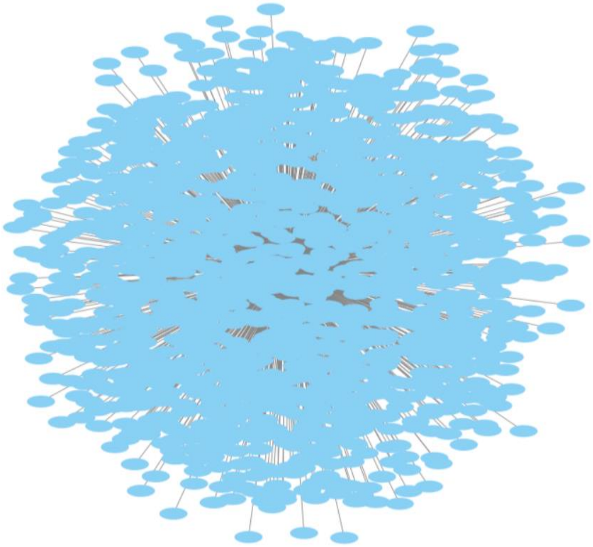
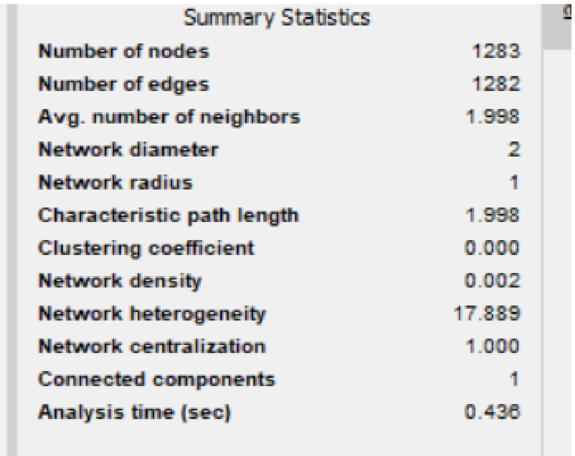


Fig 12. Most retweeted network

Next, we are going to see the summary of the above-mentioned Fig. 12 network.



Summary Statistics

Number of nodes	1283
Number of edges	1282
Avg. number of neighbors	1.998
Network diameter	2
Network radius	1
Characteristic path length	1.998
Clustering coefficient	0.000
Network density	0.002
Network heterogeneity	17.889
Network centralization	1.000
Connected components	1
Analysis time (sec)	0.436

Fig 13. Summary Statistics

We found the most tweeted account using this large dataset, named @KreatelyMedia. The below figure represents the network of that account.

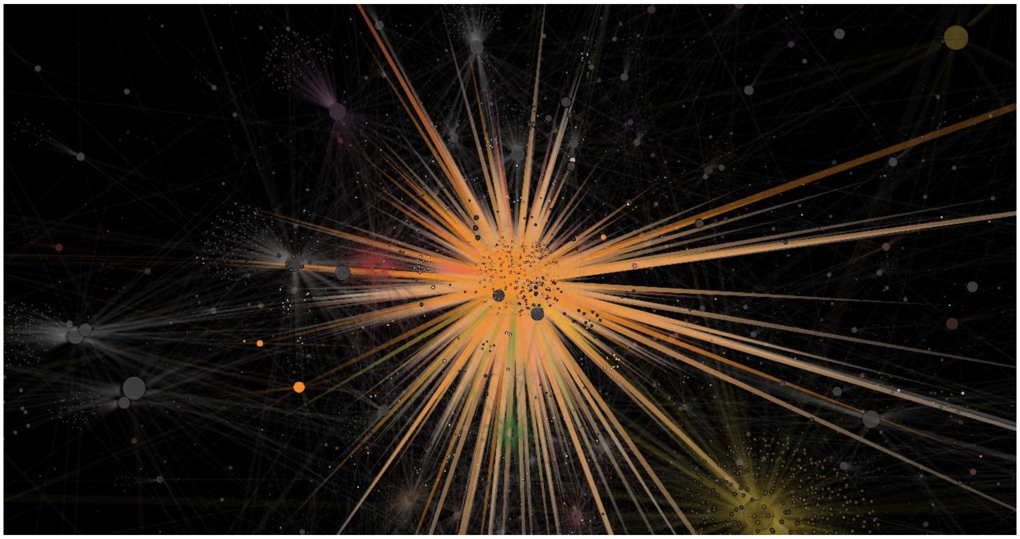


Fig 14. KreatelyMedia Twitter Account’s Network

This account was retweeted 1282 times and it is the most retweeted account. This was the result obtained from the analysis. This account has been supporting the movie The Kerala Story since the movie has portrayed only the real incident and we should understand that the movie doesn't portray the entire Islam as a wrong one it has only focussed on the particular ISIS terrorists and showed the real incident as a movie. The sovereignty of scriptures of all religions must come to an end if we want to have a united integrated modern India.

5. Conclusion

This research has analyzed the two types of networks which include co-occurrence networks and retweet networks. We have analyzed the co-occurrence network for the hashtag #thekeralastory and from this, we found out the top 5 most influential hashtags, are #thekeralastoryreview, #keralastory, #adahsharma, #kerala, #india. Then using the retweet network of these hashtags which consists of about 39844 nodes and 34672 edges and this network had 5172 connected components by analyzing all the clusters of this network we found a cluster with the highest number of nodes, edges i.e., with 1283 nodes and 1282 edges, on further analyzes on this network we discovered the Twitter account @KreatelyMedia which was retweeted 1282 times and it is highest in number when compared to all the other clusters in the network and we found that the account has supported and promoted the movie. In future analyses, we can also consider the top 10 clusters of the retweet network and discover the opinions of those clusters about the movie.

6. References

[1] Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement Ying Xiong, Moonhee Cho, Brandon Boatwright.

[2] Framework for Social Media Analysis Based on Hashtag Research Ladisliv pilar, lucie kvasnickova stanislavska, Roman Kvasnicka Petr Bouda, Jana Pitrova.

[3] Hashtag homophily in the Twitter network: Examining a controversial cause-related marketing campaign Sifan Xu, Alvin Zhou.

[4] Using Social Network Analysis to Understand Public Discussions: The Case Study of #SaudiWomenCanDrive on Twitter Zubaida Jastania1, Rabeeh Ayaz Abbasi Kawther Saeedi Mohammad Ahtisham Aslam

[5] Customer experience with farmers' markets: what hashtags can reveal done by Ladislav Pilaf, Tereza Balcarova, Stanislav Rojik, Ivana Ticha, Jana Polakova.

[6] Hashtag Analysis of Indonesian COVID-19 Tweet Using Social Network Analysis Muhammad Habibi Adri Priadana Muhammad Rifqi Ma'arif.

[7] W. Tan, M. B. Blake, I. Saleh, and S. Dustdar, ―Social-network-sourced big data analytics,‖ IEEE Internet Comput., vol. 17, no. 5, pp. 62–69, 2013

[8] A. Said, T. D. Bowman, R. A. Abbasi, N. R. Aljohani, S.-U. Hassan, and R. Nawaz, Mining network-level properties of Twitter altmetrics data,‖ Scientometrics, Apr. 2019.

[9] C. Casanueva, Á. Gallego, and M.-R. García-Sánchez, ―Social network analysis in tourism,‖ Curr. Issues Tour., vol. 19, no. 12, pp. 1190–1209, Oct. 2016.

[10] S. Rahimi, A. Abdollahpouri, and P. Moradi, ―A multi-objective particle swarm optimization algorithm for community detection in complex networks,‖ Swarm Evol. Comput., vol. 39, pp. 297–309, 2018.

[11] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, ―Defining and identifying communities in networks,‖ Proc. Natl. Acad. Sci., vol. 101, no. 9, pp. 2658–2663, 2004.