Centrality Measures to Classify Event and Non-event Synchrony
Mini Project Report

Submitted by:

Prashant Nigam- 16UCC069

Himanshu Daga- 16UCC037

Mangamuri Leela Surya Teja- 16UCC051

Under Supervision of:

Dr. Sakthi Balan Muthiah

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

The LNM INSTITUTE OF INFORMATION TECHNOLOGY

February 2019

Contents

Chapter 1

Introduction

1.1 Overview of the Project

Nowadays, Social media has become an integral part of our life. Studying and analyzing trends in social media has become very important. Our work mainly focuses on differentiating event and non-event and synchrony using centrality measures. Synchrony is defined as a social phenomenon where number of people performing certain acts first increases and then decreases.

In this project, we used Twitter as a platform to analyze the behavior of people in social media. During an event, no of tweets about the hashtag related to the event increases hugely and then decreases over a period. Due to this behavior, we label the event as synchrony. During non-event, there is a rise in the number of tweets but not related to an event. It is a generic hashtag where people share their experiences, emotions, feelings etc. So, this can be termed as non-event synchrony. Ex #mondaymotivation is used by the people to motivate themselves on Monday.

1.2 Brief description of NetworkX

NetworkX is a Python package for the creation, manipulation, study of dynamic structures and functions of complex networks.

NetworkX provides:

- tools for the study of the structure and dynamics of social, biological, and infrastructure networks;
- a standard programming interface and graph implementation that is suitable for many applications;
- a rapid development environment for collaborative, multidisciplinary projects;
- an interface to existing numerical algorithms and code written in C, C++, and FORTRAN.
- the ability to painlessly work with large nonstandard data sets.

With NetworkX we can load and store networks in standard and nonstandard data formats, generate many types of random and classic networks, analyse network structure, build network models, design new network algorithms, draw networks, and much more.

Documentation of NetworkX

Chapter 2

Social Network Analysis

2.1 Synchrony

Social synchrony is a social phenomenon in which many people respond or perform an activity together over a short duration of time. The number of individuals performing that activity first increases, remains above a certain threshold for some time, and then decreases to normal. It can be observed in our day to day life in the form of synchronised applause after some performance, choir singing etc.

It is also observed in social media platforms too. Let’s take the example of Twitter. When a hashtag is trending, i.e., a topic becomes popular lots of users tweet and retweet their views over that topic for some duration of time. On analysing the number of tweets and retweets with respect to time, it is found that generally the number of tweets first increases and then decreases in this short time span when that topic is trending. This phenomenon is termed as Online Social synchrony. This can be used to detect the presence of shared concern by the collective activity of users.

2.1.1 Event Synchrony

Whenever there is any popular event, it catches the attention of many Twitter users. They tweet or retweet about the event using a hashtag related to it to share their emotions, experiences and feelings. This forms an intertwined network of tweets and retweets. It causes the upsurge in the number of tweets related to that event. After the completion of that event number of tweets related to that event decreases forming online social synchrony. We can label this synchrony as event when retweet ratio is high. It means there is a greater number of people involved in this activity who retweeted to a tweet rather than creating their own tweets

2.1.1 Non-Event Synchrony

Sometimes some hashtags become popular due to various reasons other than the occurrence of an event. The surge of tweets over that hashtag is known as Non-Event Synchrony. It may happen that many people are using the same generic hashtag to share different types of information leading to an increase in the number of tweets with no actual significance. Or it might be a collective activity with many people talking about the same thing under the banner of a generic hashtag like #India or #Sunday. There would be rise and fall in an ample number of tweets over a period forming an Online Social Synchrony. We label this as a non-event when the retweet ratio is very low. It means there is no connection between the people involving in this activity. Number of people retweeted or replied to the tweet is very less in this activity.

2.2 Network Analysis

Firstly, let's see what a network is. A network is a cluster of closely related components. May it be people, cities, or any other related stuff. Their relationship with each other is represented diagrammatically in the form of graphs which are called Networks.

Few terminologies used in this domain:

Node: Every individual component is represented using a dot in the network diagram which is referred to as a node. There will be connections between the nodes (vertices for convenience) and they form the edges.

Degree of a node: The number of other nodes that are connected to a node constitutes the degree of the node.

There are again two types of degrees, namely:

Indegree: In a directed graph, the number of edges from other nodes that are directed towards a node forms the indegree of the node.
Outdegree: In a directed graph, the number of edges from a node towards other nodes constitutes the outdegree of the node.

2.3 Centrality

In graph theory and network analysis, indicators of centrality identify the most important vertices within a graph. Applications include identifying the most influential person(s) in a social network, key infrastructure nodes in the Internet or urban networks, and super-spreaders of disease.

2.3.1 Degree Centrality

The degree is a simple centrality measure that counts how many neighbors a node has. If the network is directed, we have two versions of the measure: in-degree is the number of incoming links or the number of predecessor nodes; out-degree is the number of outgoing links, or the number of successor nodes. Typically, we are interested in in-degree, since in-links are given by other nodes in the network, while out-links are determined by the node itself.

A node is important if it has many neighbors, or, in the directed case, if there are many other nodes that link to it, or if it links to many other nodes.

2.3.2 Closeness Centrality

By this method of measuring the importance of a node in a given network, we take the average distance of a node from all the other nodes. Whichever node has a lesser average value is the node which is nearer to every node and thus it is the node with the higher centrality.

Closeness centrality measures the mean distance from a vertex to other vertices.

2.3.3 Betweenness Centrality

In this, all the shortest paths between the nodes are taken and observed which node is visited most of the times while traversing the shortest paths and such a most visited node is the one with the highest centrality or that is the node with greater importance.

Betweenness centrality measures the extent to which a vertex lies on paths between other vertices. Vertices with high betweenness may have considerable influence within a network by their control over information passing between others.

2.3.4 Eigenvector Centrality

A natural extension of degree centrality is eigenvector centrality. In-degree centrality awards one centrality point for every link a node receives. But not all vertices are equivalent: some are more relevant than others, and, reasonably, endorsements from important nodes count more.

A node is important if it is linked to by other important nodes.

2.3.5 Katz Centrality

A major problem with eigenvector centrality arises when it considers directed graphs. Centrality is only passed on when we have (outgoing) edges, and in special cases such as when a node is in a directed acyclic graph, centrality becomes zero, even though the node can have many edges connected to it. In this case, the problem can be rectified by adding a bias term to the centrality value. The bias term is added to the centrality values for all nodes no matter how they are situated in the network (i.e., irrespective of the network topology).

A node is important if it is linked from other important nodes or if it is highly linked

2.3.6 Local Clustering Coefficient

The local clustering coefficient of a vertex (node) in a graph quantifies how close its neighbors are to being a clique (complete graph). A graph G= (V, E) formally consists of a set of vertices V and set of edges E between them.

The neighborhood of aa vertex is defined as follows:

$$N_i = \{v_j : e_{ij} \in E \vee e_{ji} \in E\}.$$

We define k_i as the number of vertices, $|N_i|$, in the neighborhood, N_i, of a vertex.
The local clustering coefficient for a vertex is then given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. The **local clustering coefficient for directed graphs** is given as

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

An undirected graph has the property that the edges are considered identical. Thus, the **local clustering coefficient for undirected graphs** can be defined as

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

2.3.7 Global Clustering Coefficient

The global clustering coefficient is based on triplets of nodes. A triplet consists of three connected nodes. A triangle, therefore, includes three closed triplets, one centered on each of the nodes (n.b. this means the three triplets in a triangle come from overlapping selections of nodes). The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed).

Chapter 3

Getting Started

3.1 Gathering Knowledge

We started this project by gathering knowledge in the domain of Social Media Analysis and Social Synchrony through following papers[1],[2],[3],[4],[5]. Then we decided to apply centrality measures on some synchronies to differentiate them among event and non-event. But for this purpose, we need knowledge about centrality measures and data to do the analysis. We learnt more about centrality measures from Social Media Mining book [13] and some other internet sources which we have described above in chapter 2.

3.2 Collecting Data

For further analysis, we took 10 synchronies that were detected as either event or non-event in [3].Then by using Twitter API, we collected all the tweets related to the hashtags of each syncrony for its duration. This provided us dataset of tweets with 16 features. We did pre-processing and cleaning of this data to finally get a dataset of 8 columns for each of the 10 chosen hashtags.

3.3 Implementation

Now, we used the NetworkX python library for exploratory data analysis of this cleaned dataset as it is good for generating visuals for graphs. This gave us a good understanding about the connectedness of users in particular synchrony.
Let's discuss more about the graphs

3.3.1 Graphs

Each synchrony dataset is represented in the form of a directed graph and each node in the graph indicates twitter_id of the user. Node exists in the graph if the user representing that node makes a tweet about the hashtag related to that synchrony or retweets about that hashtag. Assuming if there is an edge directed from node(A) to node(B), it indicates that tweet of node(A) was retweeted or replied by node(B).

Analysis of these graphical representations is done in chapter 4.1.The observations from these graphs and obtained results after applying centrality measures on these hashtags have been captured in a tabulated form in Table 3 in chapter 4.2.

Chapter 4

Analysis of Project

We have used following 10 synchronies that were detected in the Social Synchrony paper [3]. Out of 10, 6 hashtags were detected as event synchronies and other 4 were detected as non-event synchronies by the author. These hashtags are depicted in Table 1 and Table 2 below.

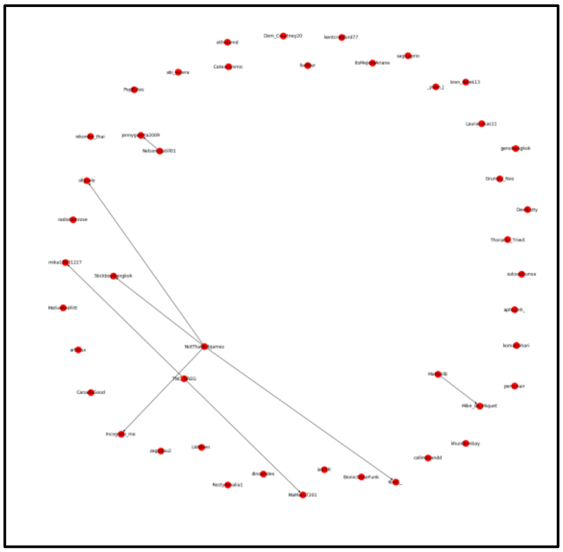
Detected event synchronies	
#hashtag	**Duration**
#bts	3 days
#favpinoynewbieinigo	2 days
#iheartawards	3 days
#kca	5 days
#savesyrianchildren	4 days
#spiritualleadersaintrampalji	3 days

Table 1: The above table represents the duration of event synchronies.

Detected non-event synchronies	
#hashtag	**Duration**
#bangkok	2 days
#india	2 days
#love	2 days
#singapore	5 days

Table 2: The above table represents the duration of non-event synchronies.

4.1 Graphical Analysis: Images of Networks Formed by Tweets and Retweets in each of these synchronies



Synchrony Type: **Non-Event**.

Total number of tweets made **51**.

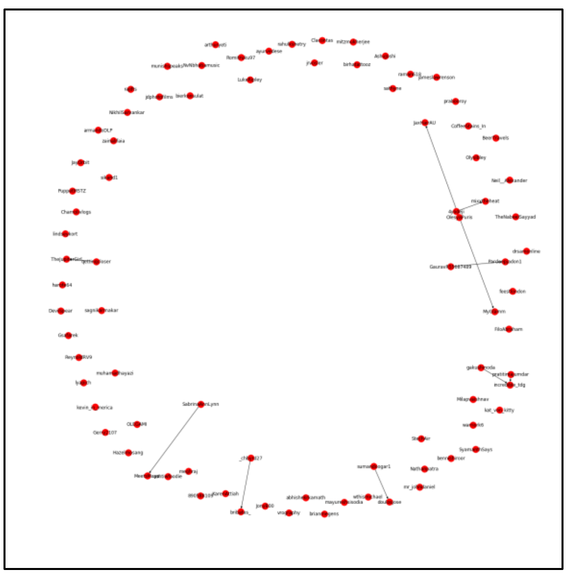
Total number of users involved were **45**

Total number of edges (retweets) in the graph are **8**

Average degree of this graph is **0.1778**

Duration of this synchrony is 2 **Days**. `

Figure 1: The above graph represents the event synchrony related to **#bangkok**



Synchrony Type: **Non-Event**.

Total number of tweets made **87**

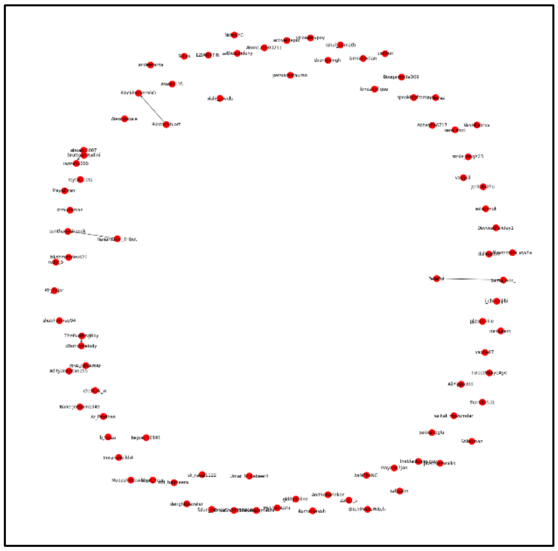
Total number of users involved were **82**

Total number of edges (retweets) in the graph are **11**

Average degree of this graph is **0.1341**

Duration of this synchrony is 2 **Days**.

Figure 2: The above graph represents the event synchrony related to **#india**



Synchrony Type: **Non-Event**.

Total number of tweets made **103**

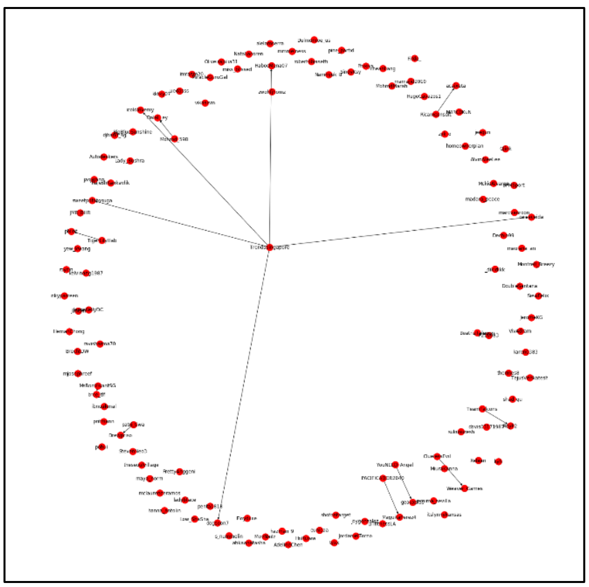
Total number of users involved were **86**

Total number of edges (retweets) in the graph are **5**

Average degree of this graph is **0.0581**

Duration of this synchrony is 2 **Days**.

Figure 3: The above graph represents the event synchrony related to **#love**



Synchrony Type: **Non-Event**.

Total number of tweets made **168**

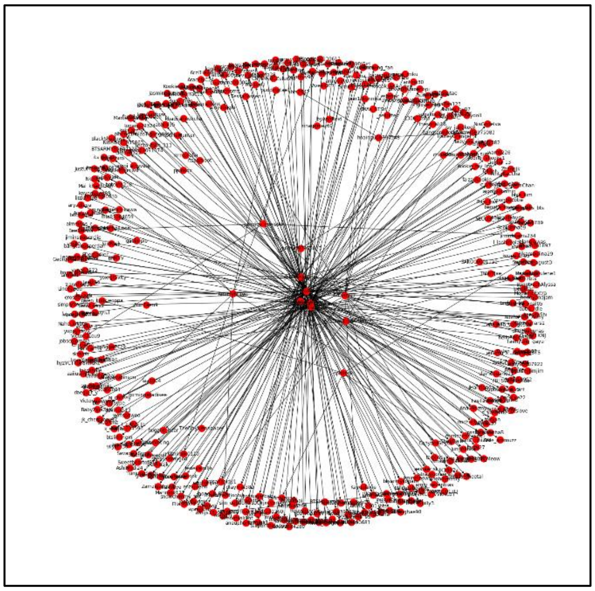
Total number of users involved were **119**

Total number of edges (retweets) in the graph are **16**

Average degree of this graph is **0.1345**

Duration of this synchrony is 5 **Days**.

Figure 4: The above graph represents the event synchrony related to **#singapore**



Synchrony Type: **Event**.

Total number of tweets made **581**

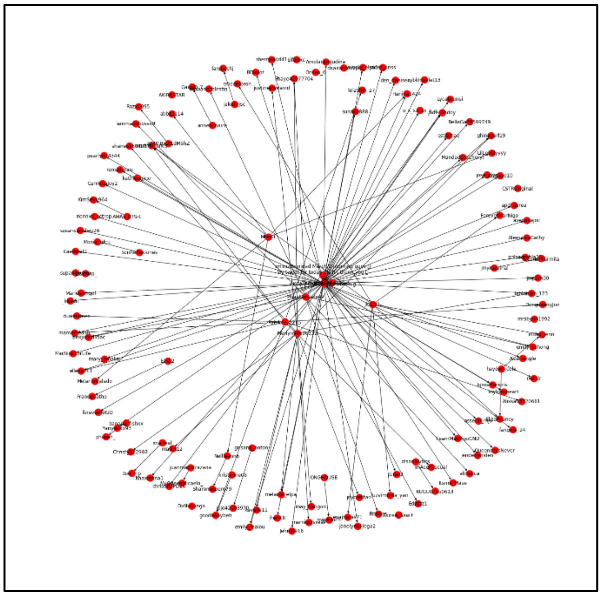
Total number of users involved were **395**

Total number of edges (retweets) in the graph are **345**

Average degree of this graph is **0.8772**

Duration of this synchrony is 3 **Days**.

Figure 5: The above graph represents the event synchrony related to **#bts**



Synchrony Type: **Event**.

Total number of tweets made **588**

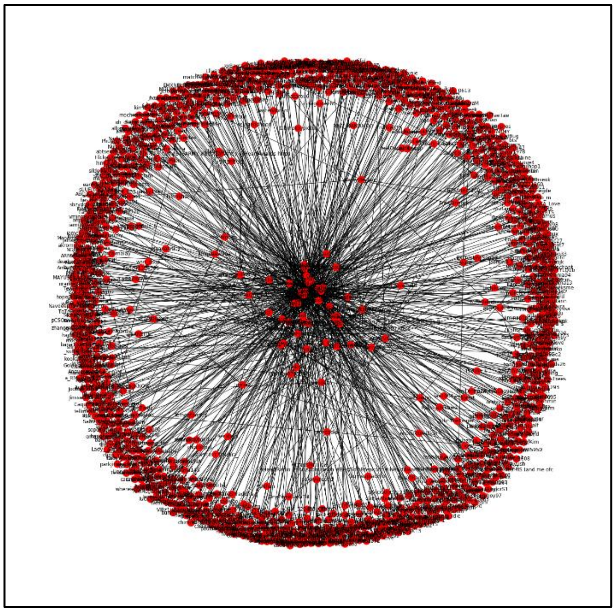
Total number of users involved were **137**

Total number of edges (retweets) in the graph are **124**

Average degree of this graph is **0.9051**

Duration of this synchrony is 2 **Days**.

Figure 6: The above graph represents the event synchrony related to **#favpinoKOynewbieinigo**



Synchrony Type: **Event**.

Total number of tweets made **1722**

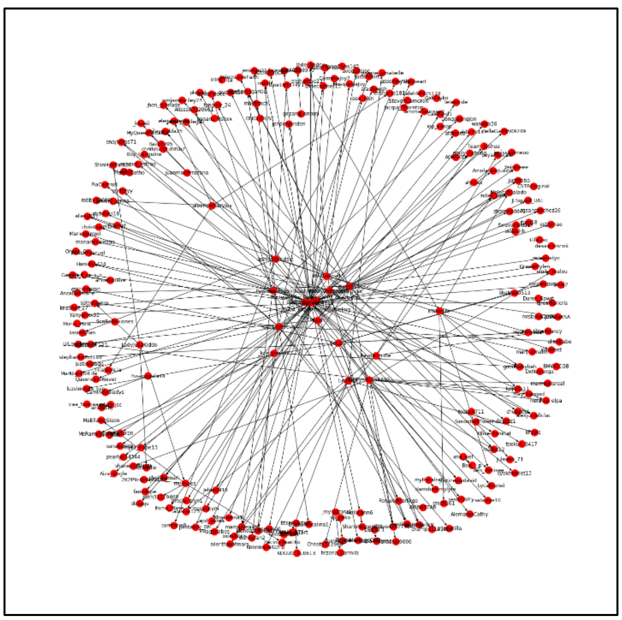
Total number of users involved were **1114**

Total number of edges (retweets) in the graph are **962**

Average degree of this graph is **0.8394**

Duration of this synchrony is 3 **Days**.

Figure 7: The above graph represents the event synchrony related to **# iheartawards**



Synchrony Type: **Event**.

Total number of tweets made **708**

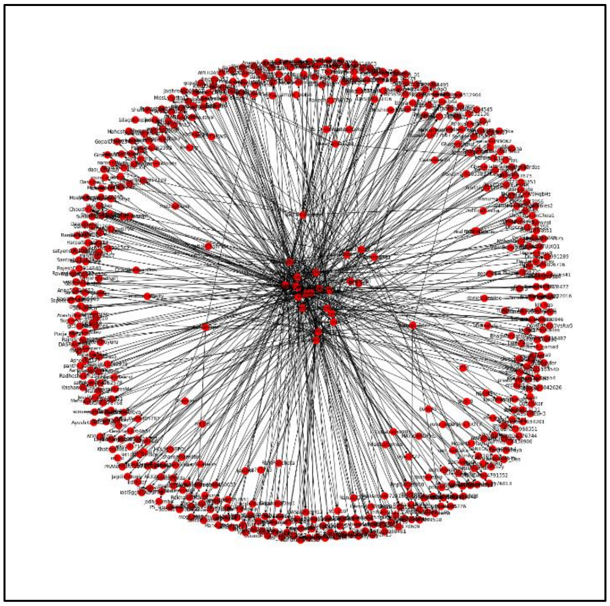
Total number of users involved were **241**

Total number of edges (retweets) in the graph are **218**

Average degree of this graph is **0.9046**

Duration of this synchrony is 5 **Days**.

Figure 8: The above graph represents the event synchrony related to **#kca**



Synchrony Type: **Event**.

Total number of tweets made **800**

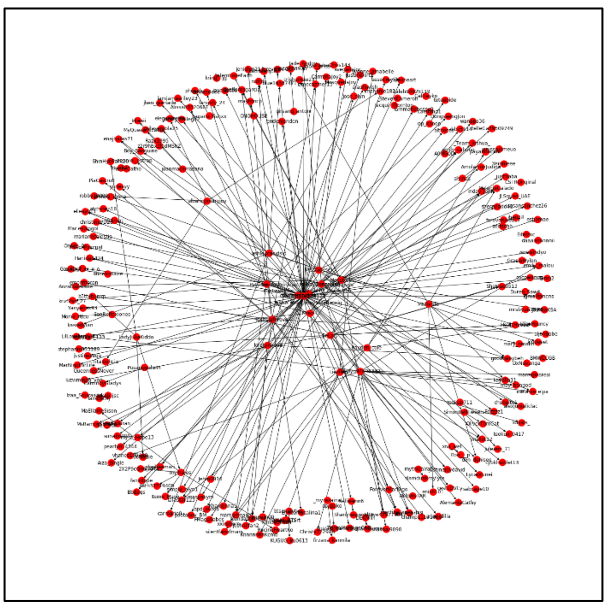
Total number of users involved were **515**

Total number of edges (retweets) in the graph are **558**

Average degree of this graph is **1.0835**

Duration of this synchrony is 5 **Days**.

Figure 9: The above graph represents the event synchrony related to **#spiritualleadersaintrampalji**



Synchrony Type: **Event**.

Total number of tweets made **1151**

Total number of users involved were **626**

Total number of edges (retweets) in the graph are **582**

Average degree of this graph is **1.0756**

Duration of this synchrony is 4 **Days**.

Figure 10: The above graph represents the event synchrony related to # **savesyrianchildren**

4.1.1 Conclusion of Graphical Analysis

- By comparing the network graph structures of events i.e. Figure 1,2,3,4 and non-events i.e. Figure 5,6,7,8,9,10 we can see that the graph of events is much denser than non-events in general.

- Average degree of non-events i.e. Figure 1,2,3,4 are very smaller when compared to the event i.e. Figure 5,6,7,8,9,10

- Number of people involved in the activity of events i.e. Figure 5,6,7,8,9,10 is more than that of non-events i.e. Fig 1,2,3,4

- **Figure 9,10** are the only two graph networks where the average degree is greater than 1.Both of them are the events. These are **#spiritualleadersaintrampaljil** and **#savesyrianchildren** respectively which are highly emotion based because of which more than **500** users are involved in these networks which are large when compared to other graph networks in our report. Both Figure 9,10 also form the most centralized networks from which we can say that there are some people or groups who are trying to influence or control the whole network.

- **Figure 3** is the graph network of **#love** which is a non-event and has the least average degree of all the other networks with value **0.0581**.Number of people involved in this synchrony are 82 which is small. Being the most sparse network of all, it indicates that there is no connection between the users involved in this network. No of retweets are exiguous which is 5 in number.

4.2 Centrality Analysis

- By observing Table 3 we can see degree centrality measure of an event is almost 10 times greater than that of non-event. This implies that number of people who retweeted and replied to the particular user about an event are more when compared to non-event of generic hashtags like #India or #love.

- In Table 3 Numbers of closeness, betweenness,kartz, eigenvector centrality, local and global clustering coefficients all tend to almost null values both for event and non-event synchrony. So, these methods are insignificant to draw any conclusion about the event and non-event synchrony.

- In the Table 1 and Table 2 Duration of all synchronies including both event and non-event varies from 2 to 5 days. The maximum length of all the synchronies is of 5 days.

#hashtag	Centrality Measures					Clustering Coefficient	
	Degree	Betweenness	Closeness	Eigenvector	Kartz	Local	Global
#bts*	1.7468	0	0	0.176	0.0775	0	0
#favpinoynewbieinigo*	1.8088	0	0	0.10	0.12	0	0
#iheartawards*	1.6589	0	0	0.17	0.49	0	0
#kca*	1.8059	0	0	0.1	0.1	0	0
#savesyrianchildren*	2.15409	0.6e-07	0.001	0.11	0.119	0.008	0.002
#spiritualleadersaintrampalji*	2.16	0	0	0.14	0.12	9.1e-05	3.8e-04
#bangkok	0.355	0	0	0.282	0.195	0	0
#india	0.243	0	0	0.100	0.172	0	0
#love	0.094	0	0	0.223	0.134	0	0
#singapore	0.208	0	0	0.133	0.133	0	0

*** indicates event synchronies**

Table 3: The above table compares centrality measures of all the hashtags.

Chapter 5

Conclusions and Future work

In this work, we have made graph network for each synchrony using network package NetworkX and implemented Centrality measures on the each of the synchrony. From the results obtained from centrality measures, we conclude that event synchrony and non-event synchrony can be differentiated based on degree centrality measure. Degree centrality measure of event synchrony is very high when compared to non-event synchrony. There is almost ten times the difference between degree centrality measure of event synchrony and non-event synchrony

There is a theoretically similar phenomenon of the sudden upsurge in people's response known as Online firestorm. We analyzed this from the following papers: [8].[9],[10],[11] and summarized it in following presentation: "Online Firestorm: A tempest in a teapot"
We believe that comparing Online Firestorm and Social Synchrony with Centrality Measures will yield more useful information which will lead to more robust solutions for detection algorithms in both the domains. We tried to do this but couldn't follow due to unavailability of data for predicted Online Firestorms as this is relatively young and open research domain. So, this might serve as a soul for many new research problems like-

- Detection of Online Firestorm
- Who initializes Online Firestorm?

Chapter 6

References

1. Unankard, S., Li, X., & Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. *World Wide Web*, *18*(5), 1393-1417.

2. Panagiotou, N., Katakis, I., & Gunopulos, D. (2016). Detecting events in online social networks: Definitions, trends and challenges. In Solving Large Scale Learning Tasks. Challenges and Algorithms (pp. 42-84). Springer, Cham.

3. Sivaraman, N. K., Muthiah, S. B., Agarwal, P., & Todwal, L. (2017, June). On Social Synchrony in Online Social Networks. In Proceedings of the 2017 ACM on Web Science Conference (pp. 417-418). ACM.

4. Dong, X., Mavroeidis, D., Calabrese, F., & Frossard, P. (2015). Multiscale event detection in social media. Data Mining and Knowledge Discovery, 29(5), 1374-1405.

5. Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. Social network analysis and mining, 3(4), 1277-1291.

6.Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. Journal of the American Statistical Association, 113(522), 607-614.

7. Suter, L., Birkland, T., & Larter, R. (2009). Disaster research and social network analysis: Examples of the scientific understanding of human dynamics at the National Science Foundation. Population Research and Policy Review, 28(1), 1-10.

8. Drasch, B., Huber, J., Panz, S., & Probst, F. (2015). Detecting Online Firestorms in Social Media.

9. Moretti, E. (2017). Report of the Editor: Journal of Economic Perspectives. American Economic Review, 107(5), 716-18.

10. Lappeman, J., Patel, M., & Appalraju, R. (2018). Firestorm Response: Managing Brand Reputation during an nWOM Firestorm by Responding to Online Complaints Individually or as a Cluster. Communicatio, 44(2), 67-87.

11. Goswami, S., Köbler, F., Leimeister, J. M., & Krcmar, H. (2010, June). Using online social networking to enhance social connectedness and social support for the elderly. Association for Information Systems.

12. Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. PLoS one, 11(6), e0155923

13 Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: an introduction. Cambridge University Press

Acknowledgement

We are overwhelmed in all humbleness and gratefulness to acknowledge our deep gratitude to all those who have helped us put our ideas to perfection and have assigned tasks, well above the level of simplicity and into something concrete and unique. We, whole heartedly thank **Dr. Sakthi Balan** for having faith in us, selecting us to be a part of his valuable project and for constantly motivating us to do better. During the process, we got a chance to see the stronger side of our technical and non-technical aspects and also strengthen our concepts of social media analysis including Online Firestorms, Social Synchrony and Centrality measures of a network.