Community Detection and Analysis of Twitter Social Data

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Aman Sharma

(Roll No. 801332001)

Under the supervision of:

Dr. Rinkle Rani

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Community Detection and Analysis of Twitter Social Data*", in partial fulfilment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Rinkle Rani and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Aman Sharma.

Signature:

(Aman Sharma)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Rinkle 07/07/15

(Dr. Rinkle Rani)

Assistant Professor,

Computer Science & Engineering Department

Thapar University,Patiala

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science & Engineering Department

Thapar University

Patiala

(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

Acknowledgement

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success. First of all I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds. With the profound sense of gratitude and heartiest regard, I express my sincere feelings of indebtedness to my guide **Dr. Rinkle Rani**, Assistant Professor, Computer Science and Engineering Department, Thapar University for her positive attitude, excellent guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all her blessings. She has been a source of inspiration for me. I am grateful to **Dr. Deepak Garg**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar University for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academics Affairs in the University, for making provisions of infrastructure such as library facilities, computer labs equipped with internet facility, immensely useful for the learners to equip themselves with latest in the field.

Last but not the least I would like to express my heartfelt thanks to my parents and my friends who with their thought provoking views, veracity and whole hearted co-operation helped me in doing this thesis.

Aman Sharma

Aman Sharma

(801332001)

Abstract

"Human is a social animal" this line itself explains the importance of society in one's life. Society brings stability, a medium to express thoughts. Society leads to social interaction which eventually brings thoughtful minds. Humans have the intrinsic nature of analyzing and opinionating things and persons. This keen nature of human has emerged a new field of analysis that is social data analysis. Internet has merged the world today and as a result human social circles have expanded. There are various peculiar social networking sites available on internet, some of them are Facebook, Twitter, LinkedIn and many more. Each maintains accounts of billions of active users and huge amount of data is being produced as a result of interactions over such sites. Hence analyzing this data is a tedious task. But analysis of such online social communities and predicting their behaviour is of great importance for businesses and academics.

For our research purpose we have used Twitter as a key medium for social data. This thesis aims to develop a research based application using twitter and R-tool for social data analysis. Further comparison of Community detection algorithm(s) on CPU and GPU technology are performed. We have used Nvidia's CUDA toolkit which provides a possibility of increasing the computational efficiency of Community detection algorithms and metrics.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

"Human is a social animal" this line itself explains the importance of society in one's life. Society brings stability, a medium to express thoughts. Society leads to social interaction which eventually brings thoughtful minds. Humans have the intrinsic nature of analyzing and opinionating things and persons. This keen nature of human has emerged a new field of analysis that is social data analysis. Internet has merged the world today and as a result human social circles have expanded. There are peculiar social networking sites available on internet, also some of them are Facebook, Twitter, LinkedIn and many more. Each maintains accounts of billions of active users and huge amount of data is being produced as a result of interactions over such sites. Hence analyzing this data is a tedious task. But analysis of such online social communities and predicting their behaviour is of great importance for businesses and academics.

The concept of "social network analysis" (SNA) was coined by Radcliffe-Brown, the great anthropologist, as a structural concern. During 1930-1970 various anthropologist articulated the concept of "social structure" as a "web or fabric" of social life. But 1950, mark the era of formal definition and technical research in the field of SNA. Hence the inter connection of social networks and their interlocking aspects got a technical approach in SNA research. Various research writings emerged as a result which formed the ground basis for upcoming field of SNA. Ever since SNA has received huge attention, various techniques for network analysis evolved. In recent years a significant increase in SNA research has been sparkled by increase in various social networking sites like Facebook, Twitter and LinkedIn etc..

Online Social data is a fuel for the research in the field of SNA (Social Network Analysis). Information plays a vital role in modern world; it is influencing various complex business decisions, understanding customer's behaviour. This data comprises of great potential in making business decision, analyzing customer behaviour and sentiment analysis etc.. Online social interaction results in massive explosion of data in the form of tweets, posts, reviews etc. Social analysis is a new paradigm of data analysis where information is used as a derived source to identify various trends,

analyzing the mood variation and opinions that are popping up after a certain event. It gave the new dimension to social analysis, using which we can even predict the future trends and make recommendations based on prior analysis. We can identify the hidden association between the different online social groups. Use of graph mining in social analysis has emerged the new science of social analysis i.e. "Community Detection". In community detection the main focus is on determining the hidden association between the different conceptual social groups that has been formulated as a result of their interactions.

1.1. Social Media and Social Computing

Traditionally the medium of communication was mostly one to one or many to one. But now a days the medium of communication has been changed to many to many kind of.

Fig. 1 : Broadcast Media: One-to-Many

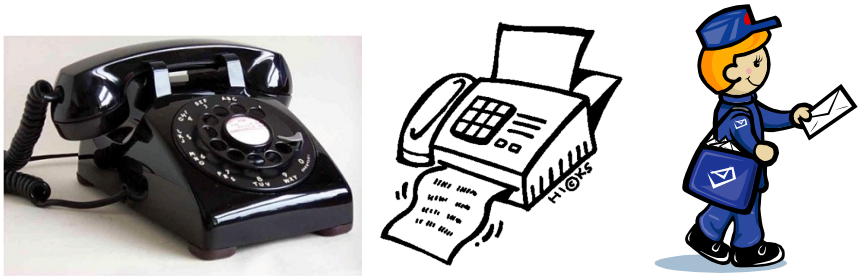


Fig. 2 : Communication Media: One-to-One

But now a day's various sources of online communication has widened the aspect of communication. Consumers are acting as producers simultaneously and the communication medium so emerged is many to many.

Fig. 3 : Communication Media: Many-to-Many

1.2 Features of Social Networks

All the social networks comprise of a structure consisting of actors/entities (e.g. people, companies etc.) and their relationships. How an actor respond, will be decided by his position and kind of connections it is holding with its neighbour actors. Nodes and edges form the network. Edges can be directed as well as undirected according to the communication structure of the problem intended to address.

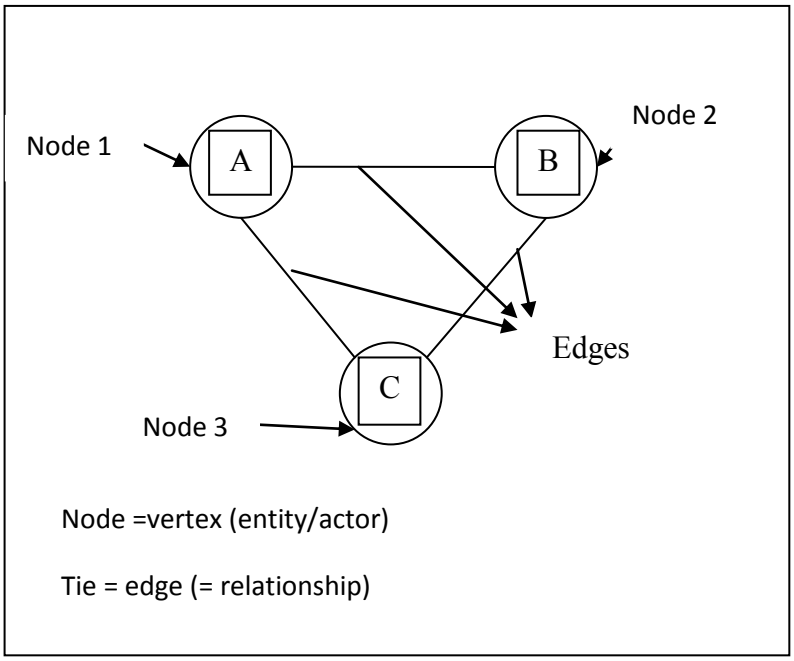


Fig. 4 : SNA Graph Structure

1.2.1 Relationship

It defines how the different nodes are connected to each other. A relationship can be defined from one node to another node through an arrow (A--->B). This arrow can be directed or undirected. Fig. 4. describes the simple undirected relationship between the nodes. Social network analysis maps and measures the relationship and flow of interaction between individuals, groups, organisations etc.. Thus various links in a network show relationships or interaction flow between the nodes.

1.2.2 Centrality

In order to understand networks and their participants, locations of actors have to be evaluated in the network. Centrality of a node is measured within the network to find the location of the node. These centrality measures give a detailed overview of various groups in a network. There are various kinds of degree of centrality, but three among them are most popular namely degree, betweenness, closeness centrality.

1.2.2.1 Degree Centrality

For measuring network activity for a node most of the social network researchers use the concept of degrees (It defines the number of direct connections a node has with neighbouring nodes). If a node has high degree centrality then it is considered an active participant in a network. Generally a node with high degree corresponds to connector or hub. High degree doesn't guarantee that the node is most connected.

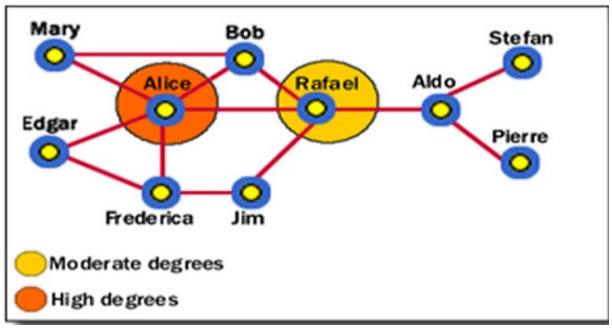


Fig. 5 : A SNA Network describing Nodes Degrees

1.2.2.2 Betweenness Centrality Betweenness centrality identifies a node's position within a network in terms of its ability to make connections to other pairs or groups in a network. A node with a high betweenness centrality generally represents a powerful

position in the network. It represents a single point of failure and have a power to influence the network.

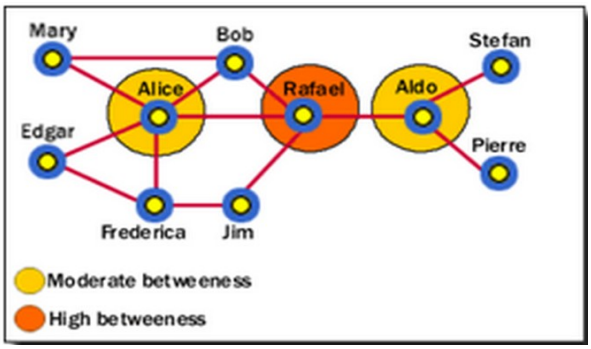


Fig. 6 : A SNA Network describing Nodes Betweeness

1.2.2.3 Closeness Centrality

Closeness centrality measures how quickly a node can access more entities in a network. An entity with a high closeness centrality generally has quick access to other nodes in a network. Such nodes describe a shortest path to other nodes. Such nodes can define about the status of the network.

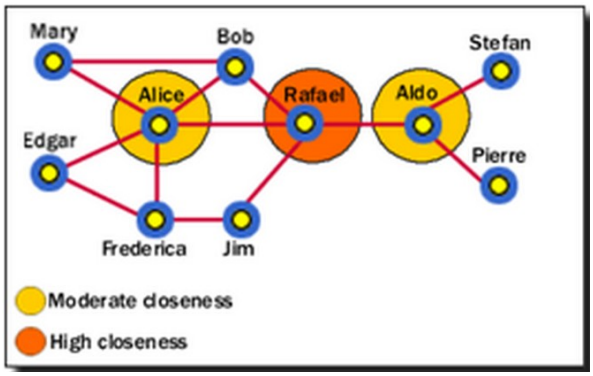


Fig. 7 : A SNA Network describing Nodes Closeness

1.2.3 Eigen value

It is a measure to identify how close a node is to other highly close nodes within a network. Eigen values define the most central nodes in a overall network. A high value indicates the node with high central influence in a network.

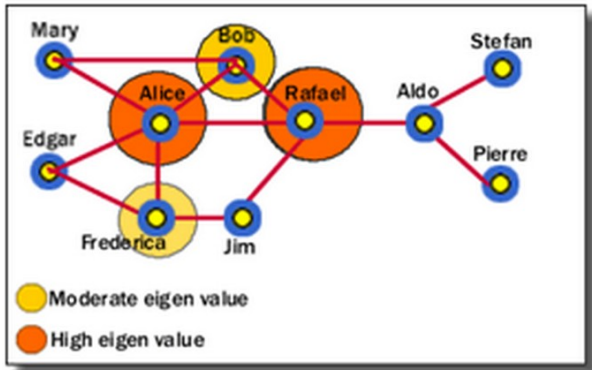


Fig. 8 : A SNA Network describing Nodes Eigen Value

1.3 Data Visualisation

A huge amount of data is being produced by social sites; this explosive growth of social media is one of the reasons that 90% of all the data in the world has been generated in the last two years alone. The reason why social data is getting so much attention is the social or personal context it carries along. Various platforms like Facebook and Twitter, stores our social interactions, photos, and log data in order to intervene our public interactions, and private communications. Such platforms are so well structured and convenient that, it is easy for us, our friends and our families to access them. We can review our social media timelines; can see the stories of our own lives in a very summative way. Visualization of our social media data has outbreak the craze for social data analysis and has simplified our experience with data analysis. It gives us insights into our own lives that we might never achieve on our own. Social graph visualizations, for example, help us make sense of the social dynamics that are playing out around us. The LinkedIn InMap can visualize our social network relationships, allowing us to see how we are related to our friends and colleagues and how closely they are related to each other. By creating a visual map of one's connections we can understand current networking efforts and how different connections are related to each another. We can also identify the scope of improvement by visualising map of professional connections. We can identify various complex relationships, patterns and social groups by visualization. We can infer about the overlapping between our social and professional network. Fig. 9 shows the map of LinkedIn network of a individual. Apart from these libraries we can also visualize our data using statistical tools like R-tool. It is a software language used for performing statistical analysis and visualization. It includes installation of R

environment. We can use CRAN-mirror for downloading and installing different R packages and libraries. In India it is maintained by IIT Madras.

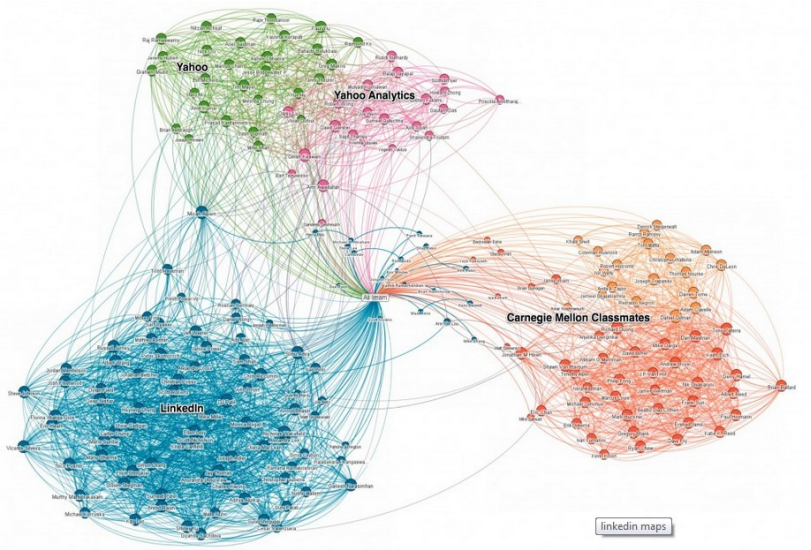


Fig. 9 : LinkedIn InMap Network

Various tools are available for data visualization. Some of the popular data visualisation tools are Gephi, NodeXl and R tool etc.. R tool provides a set of best possible libraries and packages for data visualization of extracted and processed data. But what if we want 3-D visualization of data using day to day useful things like lamp, smoothie jar etc., then there is a need for IoT solutions, as they connect these two ends.

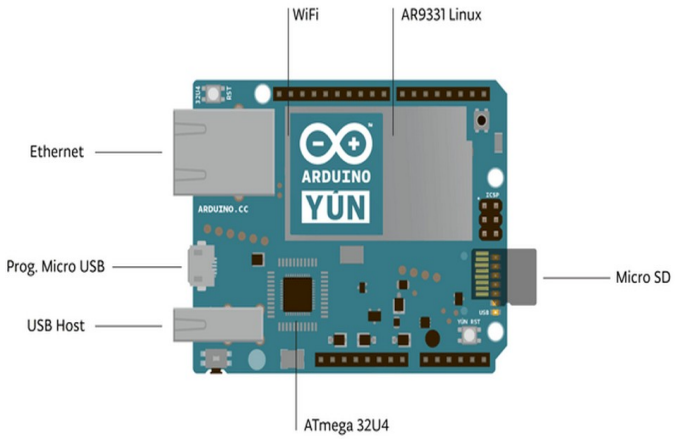


Fig. 10 : Arduino Board-YUN

IoT has given developers wings to fly their imagination into their programming skills. IoT has given a new concept of development, i.e. DNA (Device Network

Application). You can connect any device to network and can operate or sense it using application. Various interesting and innovative applications are emerging since the launch of Arduino. Hence a new approach to 3-D visualization is also driven with its advent.

TABLE 1 : Various Arduino Boards and Related Shields

S.No.	Arduino & Shields	Features
1.	ArduinoBT	Provides built-in Bluetooth module, allows wireless communication.
2.	ArduinoPro	Provides flexibility and low-cost
3.	ArduinoYún	Can be config.d to connect to your Wi-Fi network.
4.	ArduinoEthernet shield	Can be connected to internet with Ethernet library and can read/ write SD card with SD library.
5.	Arduino ISP	You can program AVR micro-controllers.
6.	Arduino GSM shield	You can connect your arduino with internet, send/receive SMS and make voice calls with GSM library.
7.	Arduino Wireless shield	Wireless communication using Zigbee can be made on Arduino.

Some of the applications [38, 39, 40, 41] that got attention from social network researchers and enrich the IoT market with smart devices and even help the social data analyst to dig the new visualisation paradigms:

1.3.1 Twitter Mood Light

There is well stated quote with this application "The World's Mood in a Box". If you are a news junkie, and keen about knowing everything happening around you then it is a perfect application for you. Suppose you slept at night and mean while something unusual or something trendy has happened overnight, and you want that if you wake-

up early in the morning then there should be something seeing which you can get this information. Then you have a solution for this "Twitter MOOD Light". A lamp will glow a unique colour based on the frequency of tweets related to particular mood. For example red for Anger, orange for happy, green for sadness. You need an Arduino Duemilanove, Wifly Shield. An Arduino can be connected directly to any wireless network via the WiFly module. It then repeatedly searches Twitter for sentiment tweets, maps the tweets for each emotion, analyzes the data, and fades the colour of an LED to reflect the current World Mood.

Fig. 11 : Mood Lamp for Twitter Tweets(Orange-happy)

1.3.2 Tasty Tweets

It is a data visualization experiment, which helps users to analyze twitter trends through their taste buds on a single press of a button. It collects tweets from twitter using Twitter API. Tweets having mentions of keywords of fruits like blueberry, pineapple, apple and carrot are collected and based on their frequency, a blend of smoothie are made which represents the trending graph for tweets. Flavour of smoothie changes as the trend changes. So every time you will get a unique flavour. Arduino is connected to various jars to extract the smoothies in equivalent portions to tweets extracted. This unique way to experience twitter trends and visualize them. Fig. 12. shows visual graph of smoothie being prepared so.

Fig. 12 : Tweets Smoothie

Other than these visualization applications we can also create interesting applications using arduino and Twitter. Some of these applications are:

1.3.3 Twitter Enabled Coffee Pot

It is one of the kinds of IoT solution that has revolutionized control of device remotely. It is a twitter enabled device which helps you to instruct your coffee maker to make a coffee anytime from anywhere. All this requires a Power switch tail, an arduino board, a computer with arduino IDE and python and of course Drip Coffee Pot.

1.3.4 Twitter Theft Control

This device will help you to monitor your things well and ensure their safety. If someone comes too close to your thing, then it will send an update to twitter. It includes an Arduino, distance sensor, buzzer and a serial cable. Now you can safeguard your things even if you are not nearby them.

1.4 Some of Recent Applications of Social Network Analysis

Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. SNA provides both a visual and a mathematical analysis of human relationships. Management consultants may use their project methodology with their business clients and call it Organizational Network Analysis

(ONA). Many datasets can be described in the form of graphs or networks where nodes in the graph represent entities and edges represent relationships between pairs of entities. A common property of these networks is their community structure, considered as clusters of densely connected groups of vertices, with only sparser connections between groups. The identification of such communities relies on some notion of clustering or density measure, which defines the communities that can be found. SNA has its foundation in a number of fields, including mathematics, spatial geometry, sociology, and anthropology, and focuses on the nature and characteristics of relationships. Citation analysis, Social media analysis, Friendship Networks, Web Graph and many more are the few application domains of SNA (community detection).

1.4.1 Ebola Outbreak(EVD)

Recently a disease named "Ebola" was a active topic of discussion in social and digital media. Even various social scientists focused their interest toward this burning topic. Many data scientists worked collaboratively to identify the source from where this disease outbreaks. They collected the data from social networking sites like twitter and further created a network of connected users based on tweets and retweets of posts. Data so collected is managed and stored using graph database like Neo4j which stores information in a native graph structure. Data is stored in the form of relationship between the entities. The main objectives of this research were:

- To identify users who are (or could be) influential and who are actually active to spread the messages within and outside particular community.
- To identify the hidden community structure that may be popping around or has evolved over the period of outbreak.
- To identify individuals who are prime participants at spreading potentially damaging misinformation about the problem/disease.

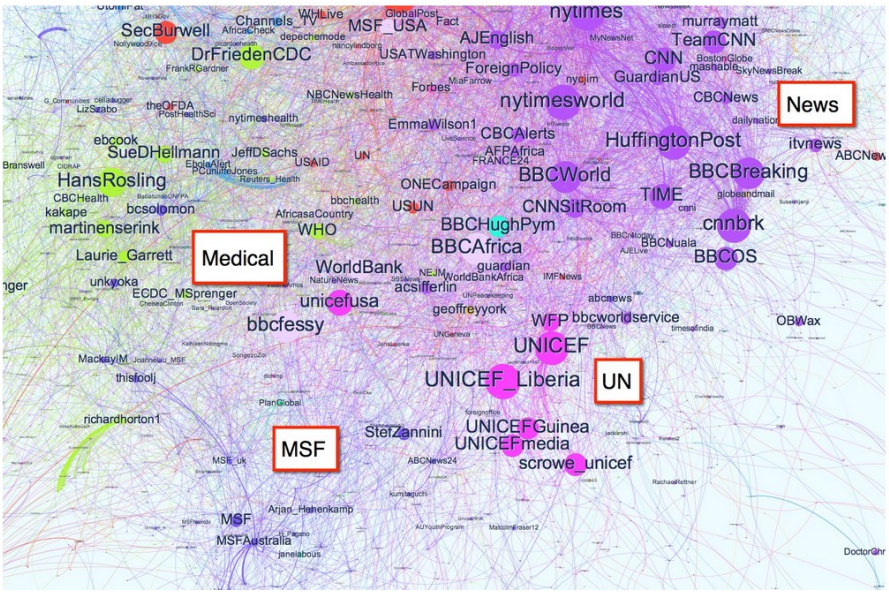


Fig. 13 : Detection and Separation of Several Distinct Community in Tweets Network

Whenever an infection or disease starts spreading, it seldom takes its time. If the disease is Ebola then any delay in its ailment can lead to real toll in human lives. The major drawback of such diseases is that it takes time for people to even fig. out that an outbreak has occurred. But thanks to our data scientists and powerful analytics tools that can help to identify the outbreak of such diseases much before it can show its devastating effects. An algorithm was designed to identify Ebola outbreak and it had spotted it nine days before the WHO (World Health Organization) announced the outbreak of Ebola in Africa.

How this algorithm works?

- It collects data from various data sources like news reports, social media, international health organizations, government websites, and even the personal blogs of health care workers operating in affected areas.
- A specialised web crawler is designed which traverse the RSS feeds and APIs. It then analyzes the text from accumulated data sources and identify the geography and status of disease severity.
- A feedback loops in algorithm helps it to improve further.

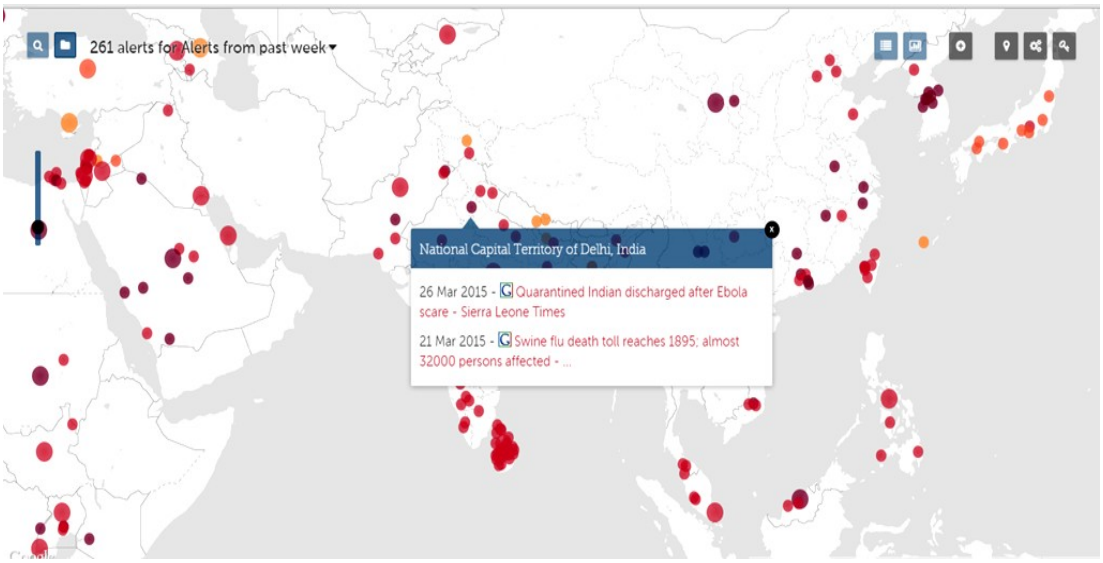


Fig. 14 : Snapshot of Healthmap of India-Delhi

1.5 Differnt R Packages used for Implementation

- twitteR- Interface to Twitter web API is provided by twitteR package developed in R. It provides access to various function like getTrends, getUser, search Twitter and many more for fetching data from twitter.
- stringr- stringr is a R package that has simplified working with strings by providing various built in string functions. They all are easy and handy to use. It provides consistent function and argument names which help in easy adoption and usage of these functions, NA's and zero length character are properly handled by each function appropriately, and there is appropriate matching between the input and output data structures corresponding to particular functions.
- RCurl- HTTP requests are composed with the help of this package, also provide functions to gather URL's, get & post forms and returns the result fetched by web servers.
- ROAuth- Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.
- tm- It provides a framework for text mining applications within R.
- RJSONIO- Conversion to and from Java script object notation (JSON) is provided by this package. R objects are being inserted into Javascript/ECMAScript/ActionScript code and provide R programmers the ability to read JSON content and their conversion to R objects. It is alternative to RJSON.

It was discarded because of slow conversion of R objects to JSON and was also inextensible.

- wordcloud- Pretty word clouds.You can plot a cloud of words shared across document and plot a cloud comparing the frequencies of words across documents.
- gridExtra- It provides high-level of Grid functions. It provides functions for Gridgraphics. Example you can draw an open rectangular borderdraw(borderGrob(type = 1, colour = "white", vp = NULL, ...) where vp is viewpoint.
- plyr- plyr is a set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each pieces and then put all the pieces back together.
- shiny- Shiny makes it easy to build interactive web applications using R. Automatic binding between inputs and outputs and extensive pre-built widgets make it possible to build beautiful, responsive, and powerful applications with minimal effort.
- ggplot2- An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources. It also implements a sophisticated multidimensional conditioning system and a consistent interface to map data to aesthetic attributes.
- reshape- Reshape lets you flexible restructure and aggregate data using just two functions: melt and cast

Chapter 2 Literature Review

Social media is one of the biggest sources of interaction among people through which they create their virtual communities and exchange information and ideas. This idea of social communities provides them a way to create, share and innovate intellectually and communally. With the increase in number of social networking platforms and increase in the user's participation on them has opened a new paradigm of social analysis for data scientists. For example, Facebook involves over a billion users interaction and Twitter produces tweet messages around 400 million per day. Moreover, smart phones and mobiles have provided added benefits by offering ubiquitous access to these platforms from everywhere and anytime [2]. Therefore, huge amount of data is being generated at tremendous rates make real time analysis of such a fast growing data a bottleneck. Since its launch in 2006, Twitter has gained tremendous users participation and has come up as a most popular micro blogging site [3]. It is simple interface provides easy facility to users for sending small messages known as 'tweets' into digital space which is hyper-real, normally called Twitter sphere [4-5]. More than 400 million tweets are sent worldwide each day [6]. Data is being used for detecting, monitoring and analysing various events (disasters and its impact, mood variations, communal opinion building).

2.1 Twitter

Twitter provides a full range of tools for developers. Twitter provides developers with handful of API's. There is REST API which provides programmer the ability to access, read and write Twitter data. The Streaming API provides continuous access to the data pertaining to particular search query. Stream queries run continuously till there is internet connection or until they are killed but REST queries dies eventually based on the parameters you are querying. Various streaming APIs provided by Twitter gives developers access to twitter's global stream data. Twitter offers several streaming endpoints [7]:

- Public streams: It provides access to data streams which are publically available on Twitter. Mostly they are used for following specific topics, individual and for data mining.

- User streams: They provide access to twitter data related to a particular user. You get access to tweets, followers and friends of a user you wish to.
- Site streams: It is a multi-version of user streams. Site streams are used for servers which connect to twitter as a proxy server for many users.

Table 2 : Parameters for the "POST statuses/filter" Endpoint

Parameter	Description
Follow	A comma separated list of user ID's, indicating the users to return statuses for in the stream.
Track	Keywords to track. Phrases of keywords are specified by a comma-separated list.
Locations	Specifies a set of bounding boxes to track.
Delimited	Specifies whether messages should be length-delimited.
Stall_warnings	Specifies whether stall warnings should be delivered.

2.1.1. Public Streams

For real time analysis of twitter data Public streams provide a good bench source of API. Public streams are divided in three endpoints:

- POST statuses/filter
- GET statuses/sample
- GET statuses/firehose

The "POST statuses/filter" returns status related to the various filter parameters. Multiple parameters can be provided for filtering the search queries in streaming mode as shown in the Table 2. At least one filter parameter (follow, track, locations) should be given to proceed with the search process. The resource URL for accessing this endpoint is the following:

https://stream.twitter.com/1.1/statuses/filter.json

Various parameters can be provided as: track=loo&user=3434. The "GET statuses/sample" endpoint returns small sample of all public statuses gathered

randomly. This endpoint is good for collecting information about the trending topics related to particular moment. The resource URL is:

https://stream.twitter.com/1.1/statuses/sample.json

Finally, the firehose endpoint returns all public statuses. However special access permission is required by this endpoint.

2.1.2. Data format

Default data format for extracting and presenting twitter data is JSON. The fig 1. is an example of a status update (a tweet) encoded in JSON. The information provided in JSON format contains tweet related data: tweet creation date, tweet text, tweet location, different hash tags, urls or user mentions, etc. Retweet information is also provided along with the original information. Information regarding creator of the post is also coupled along in the tweet information.


```
{
"created_at":"Sat Apr 13 16:15:44 +0000 2013",
"id":323107241500753920,
"id_str":"323107241500753920",
"text":"My car needs a shower asap, thank you dust storm for that :)",
"source":"\u003ca href=\"http://twitter.com/download/iphone\" rel=\"nofollow\"\
    u003eTwitter for iPhone\u003c/a\u003e",
"truncated":false,
"in_reply_to_status_id":null,
"in_reply_to_status_id_str":null,
"in_reply_to_user_id":null,
"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,
"user":{
  "id":112814547,
  "id_str":"112814547",
  "name":"Farah",
  "screen_name":"Farah327",
  "location":"www.changebyfas.com",
  "url":null,
```


Fig. 15 : A Tweet Encoded in JSON

2.1.3. Rate Limit

All public streaming tweets are fetched with the help of firehose. Normally any twitter user can only access around the 1% of the total twitter data. If filter parameters exceed the defined percentage of 1%, a rate limit notice is sent to user about the exceed percentage. Normally message containing information that how many tweets are not included is sent to user. Rate limits are difficult to increase, so for same reason we can use other providers of Twitter data also.

2.1.4 Other Providers of Twitter Data

Twitter data need constrained till rate limits are easily fulfilled by Twitter API's but above that we have to look for other sources. There are some cases in which we need much more than 1% of the whole stream. GNIP1 is one of the largest providers of social data and has tied up with Twitter since 2010. GNIP provides the full Twitter firehose. The only difference is in syntax other than that major parameters, attributes and information is same.

2.2 Community Detection

In the recent years a lot of interest has been seen in the area related to network science and as a result various graph based approaches have evolved. Other than social network properties like power law and small-world properties various real world networks are evolving and making a world a big cluster of real world entities. Community Detection (or graph clustering) algorithms aim to detect these communities/clusters of nodes, given the graph. A tremendous amount of work has been done in the field of community detection and various methods have been proposed over the years.

2.2.1 Overview of Graphs

Graphs are among the most interesting and useful objects in mathematics. Any situation or idea that can be described by objects with connections is a graph, and one of the most prominent examples of a real-world graph that one can come up with is a social network. A graph (or network) G (V, E) is a set of vertices (or nodes) V , and a set of edges (or links).

The number of nodes (or order) of the graph is the number of elements of set V and is denoted by |V |

The number of edges (or size) of the graph is the number of elements of set E and is denoted by |E|

2.2.1.1 Graph Parameters

Edge: A tuple (u,v) is called an edge where u,v ∈ V. The tuple signifies a connection between a real world entity u and another entity v. The edge set $E \subseteq V \times V$.

Generally edges in a graph are directional means if there is an edge from entity u to entity v then its direction will be mentioned in the graph and edge from u to v is not same as edge from v to u. However there are undirected graphs also means a edge from u to v is same as an edge from v to u.

An edge in a network or a graph has a real valued number or weight associated with it. Weights on a graph depend on the kind of network for which it is used. For e.g. a graph for a social network of twitter data has tweets and retweets as the edge weight along with them. Other than weighted graphs there are unweighted graphs too. They have generally Boolean values associated along with them either 0 or 1, depicting if there is an edge or not.

2.2.1.2 Graph Representation Techniques

A graph can be represented by a matrix consisting of binary values and such matrices are called adjacency matrix.

```
Matrixfill(**p )

{

If (edge between i and j)

A (i,j) =1 ;

Else

Return 0;

}
```

The count for the connection a vertex is having with other vertices is defined as the degree of the vertex. There is also a matrix named transition probability matrix X associated with a graph such that it is given by $X = A*B^{-1}$ where B^{-1} is the inverse of matrix B. Such a matrix represents the probability of a next state from present state.

P (i,j)=Probability that we will go from state i to j

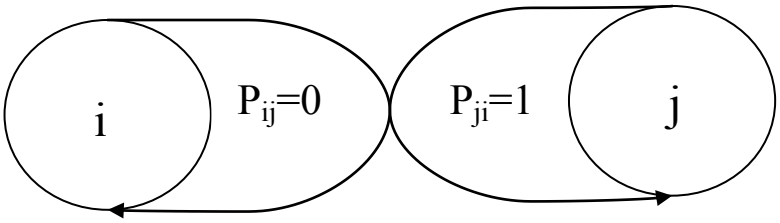


Fig. 16 : State Diagram

2.2.2 Creating Social Media Network

We can create social networks as a result of various online transations that occurs daily. We can record these transations to create a network out of them. Every transaction have entities involved with it; considering the example of Flickr tag assignment where we have entities like user, a photo and a tag, similarly a blog network involves entities like blog commenter, the blog article and the text. Hence a association is made between the entities of same kind and results in the network which has popped as a result of association among various entities. Any network formed represents the subset or snapshot of the social transactions that are going on. Hypergraphs are used for mathematical representation of complex networks. As hypergraphs can handle multi-way edges too. But one of the drawback of hypergraphs is that the majority of network analysis methods, and community detection in particular, are not applicable to hypergraphs or k-partite graphs. This is the reason that we focus on simplified network which describes the partial aspect of the big and complex evolving networks. So in order to have simplified networks we prefer to have one- or two-mode vertices and edges which are simple in nature, that is conneted to two vertices. Hence network formed is used for social analysis.

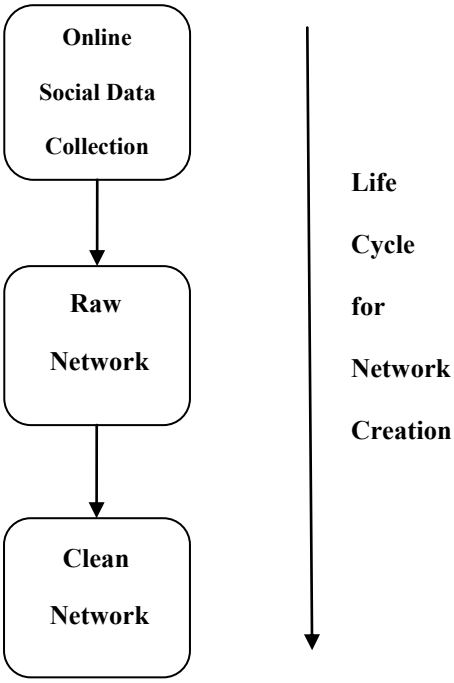


Fig. 17 : Pre-analysis Steps

Thus a huge amount of preprocessing is required before analysis on large complex networks. Most of the preprocessing task involves creation of simplified and clean networks.

2.2.3 Community

One of the most simplest defination of community is to say that a community is a subset of vertices that are completely connected to one another. Technically speaking, a community is a sub-graph which forms a clique. Sometimes an n-clique is also called a complete graph on n vertices, denoted K_n. Fig. 18. is an example of a 5-clique in a larger graph.

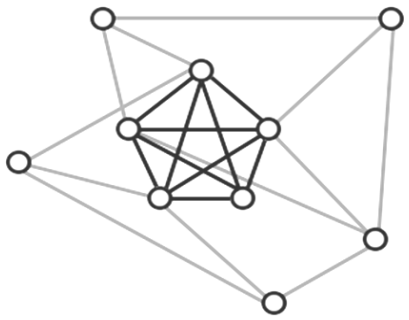


Fig. 18 : 5-clique in a Larger Graph

A huge amount of work has been done in the field related to community detection hence defining community is a tedious task. A variety of perspectives to the community structure has varied the defination of community in different contexts. As a result there is no unique defination of community however their context is generally related to network structure. A community defines some sets of local entities depicting similarity or some sets of global entities. A community is constrained to the domain under which it is studied. At the most abstract level, given a Social Media network G = (V,E), a Social Media community can be defined as a subgraph of the network comprising a set VC⊆ V of Social Media entities that are associated with a common element of interest. The only variation that can be seen is in the element it can be a real-world person, a place, an event, an activity or a cause. In a network of blogs the element or node attributes are tags, articles, comments on related topics of interest**.** Similarly an application for sharing photos, node attributes are photos , tags and users and hence mutually associated constitutes a community. Social communities can be divided into two categories as explicit and implicit.

2.2.3.1 Explicit Communities

Explicit are those communities which are formulated as a result of mutual consent between community offerer and individuals. Facebook, Flickr are few examples of explicit communities.

2.2.3.2 Implicit communities

These are the communities those are already existed in nature but just need to be discovered over time.The most important property that makes them different from other communities is that there is no need to have a human effort to put in for its creation.

2.2.4 Attributes and structure of community

Although various community definations have been proposed but the most general defination based on sets is the relationship of vertices from vertex set (V), such that it is defined though boolean decisions.But in reality the defination of community concept is much more complicated. Other than above mentioned definations several local [8, 9], some based on percolation clique[10] and few

based on overlapping[11, 12] nature of communities. By overlaping we mean by the entities or attributes which are common to various communities. For instance same person can be a part of friends community and family community simultaneosly. Other than these attributes there are attributes related to vertices based on their centrality measures. Xu et al. [13] defines isolated vertices as (hubs or outliers). Hubs act as a communication intermediate between the communities and thus help to establish interaction. A special case of hubs are outliers and are connected to a single community through a single link. They are considered as the noise in the network and generally discarded. Scripps et al. [14] discussed the vertix role based on communities. Overall community structure can be identified and considered at various level and in varied application domains.

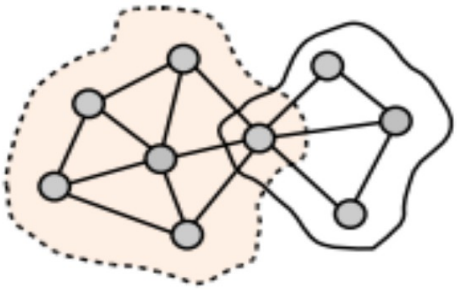


Fig. 19 : Overlapping Communities

Fig. 20 : Weighted Members

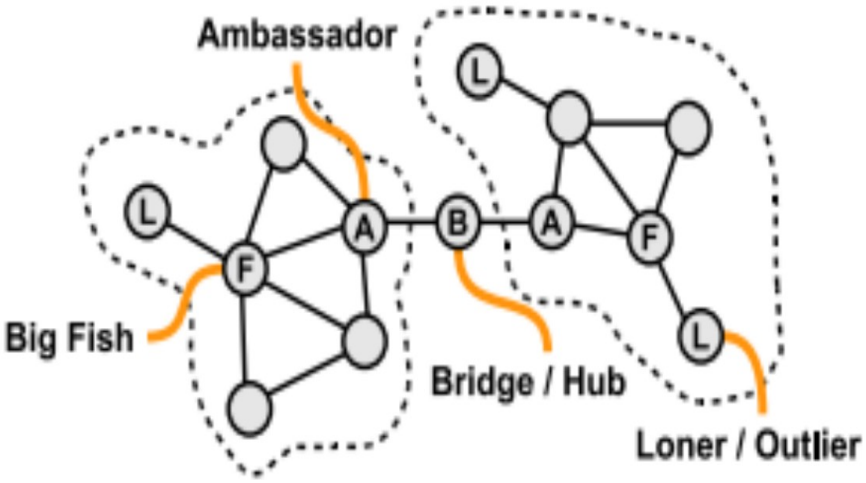


Fig. 21 : Roles of Different Nodes

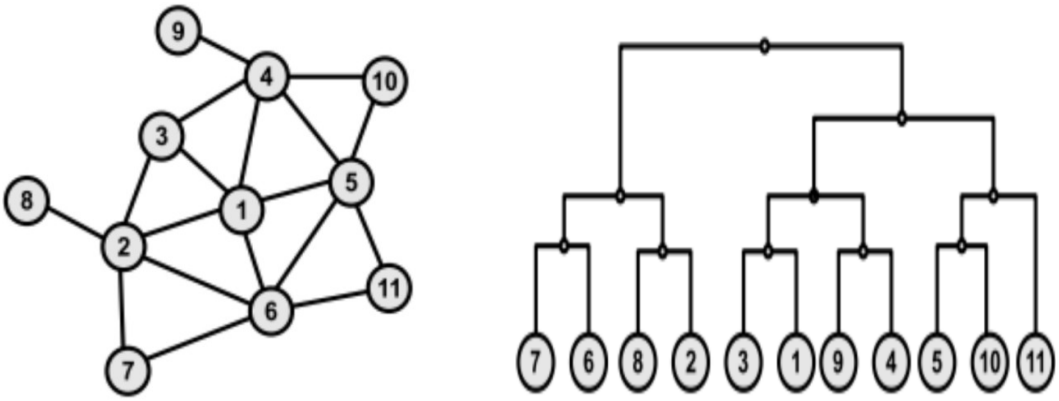


Fig. 22 : Hierarical Structure

2.2.5 Methods for Community Detection

A lot of research has been done on the very buzz topic "Community Detection". Various approaches and application perspectives have been suggested by researcher. A wide and quality discussion can be seen in the work done by Fortunato [15] and various other surveys conducted presently [16]. A varied amount of approaches for detecting communities have been proposed in [17]. Community-ness and cohesion in a community are the two main aspects to be considered for community detection algorithms. Community-ness can be classified into four major categories:

- Parameters such as Number of edges, Density within edges, Average degree [18] and Intensity [19] form an aggregate property of the edges in a community and hence called as internal score.
- Parameters such as betweenness centrality [20, 21], Expansion [18] and Cut Ratio [15] are based on the aggregate property of those edges which are medium of connection between community and the network.
- There are scores some of them are internal and some are external ones, such as Conductance [22] , Normalized 11 Cut [22] including the properties of internal as well as external edges.
- There are some scores based on network such as Modularity [23] and there are some of its variants such as local modularity [24], and sub-graph modularity [9].

So on the whole the concept of community-ness measures coherence and identifies the node in a network based on node's relative importance in a sub-network with respect to complete network. Palla et al. [10] defines the clique percolation algorithm

(CPM) for finding overlapping communities based on their density in a network. A further enhancement to CPM was given by Farkas et al. [27] but for graphs with weights. It excludes the cliques from the network that doesn't meet the threshold value at the time of percolation step. A concept of modularity was first introduced by [Newman] as a medium to dig deep into community structure. This gave a way to foster the research and developed the whole new class of modularity maximization methods for detecting communities [28, 29, 30, 31, 8]. But the most common problem that came across with these methods that they couldn't detect communities of smaller size [30]. As a results new methods based on label, propelled the further research and label propagation methods where devised. They can be further extended to overlapping communities as nodes with multiple labels came across as overlapping nodes. There are approaches such as EAGLE [31], LFM [32], GCE [33] and OSLOM [34] which grows a community form a single node to a complete set of community nodes till a defined fitness function holds. The quality of density of a community is characterised by the fitness functions defined. A clustering technique for community detection was proposed by Ahn et al. [35] using undirected graphs. It uses jaccard index of two nodes which are in neighbourhood for the analysis of links. Recently network models have been proposed for community affiliation [36, 37] claim that the community overlaps are more densely connected than the non-overlapping parts and build models on bipartite node-community affiliation networks**.** The methods we have discussed so far has focused on finding sub-graphs in a graphs. What about the situation if we are given a graph and we can run some kind of experiment on the graph, and the results that we will get from the experiment will give us insight about the communities that are existing. The experiment such as random walk could serve the purpose. What we mean by random walk is that say we have a vertex v in a graph G and we want to find vertices that are closest to v. We are actually finding vertices that are likely to be more communal. By doing random walk by starting from vertex v we can get the statistics about the visited vertices. The most visited vertices will actually form a community and are said to be in same community. Generally a walk is of 3 to 6 steps long.

Chapter 3 Research Problem

3.1 Problem Statement

Huge amount of digital data is being produced as a result of our routine activities. Information plays a vital role in modern world; it is influencing various complex business decisions, understanding customer's behaviour. Advancement in web2.0 has interconnected the world and thus the sources of information. Wide adaptation to mobile devices and social networking sites has led to explosion in data production. This data comprises of great potential in making business decision, analyzing customer behaviour and sentiment analysis etc..

Online social interaction results in massive explosion of data in the form of tweets, posts, reviews etc. One of the way in which web has boost the social data analytics is through "reviews". Different kinds of reviews are made on products, books, restaurants and on many more by their consumers. A review can be an appraisal or criticism but it adds a lot of value to the analytics, in making a big picture out of these reviews. Second thing that has propelled social data analytics is online posts and tweets about any new happening or event, about political personalities/celebrities or about their daily life.

Social analysis is a new paradigm of data analysis where information is used as a derived source to identify various trends, analyzing the mood variation and opinions that are popping up after a certain event. It gave the new dimension to social analysis, using which we can even predict the future trends and make recommendations based on prior analysis. We can identify the hidden association between the different online social groups. Use of graph mining in social analysis has emerged the new science of social analysis i.e. "Community Detection". In community detection the main focus is on determining the hidden association between the different conceptual social groups that has been formulated as a result of their interactions. Reviews, emoticons and their hidden context help in building opinions and identifying trends. Visualization of gained result is a tedious task. There are various set of mathematical methods that are used to analyze social aspect of sociology, psychology, anthropology and ethnology.

Most of the models of network analysis work on the assumption that the kind of communication between group members greatly affects important features of that group. Use of mathematical tools and concepts for network analysis can make our job easier. Social network analysis is focused on uncovering the patterning of people's interaction.

3.2 Research Gaps

A lot of research is going this new paradigm of analysis but no such case study has been discussed to focus upon the various community parameters and their comparisons. Statistical processes are compute intensive and hence with the leverage of parallel processing computational gains can be produced. A comparison of CPU vs GPU statistical tasks has been performed in determining communities. Specifically our research work focuses on the communal attributes for classification using the network populated by the extracted and processed social data.

3.3 Research Objectives

In the light of above discussed research gaps following objectives have been formulated.

- To study various techniques and tools available for data analytics and visualisation
- To develop a research based application using twitter and R-tool to analyze and visualize the social data.
- Execution of Community detection algorithm(s) on extracted Twitter data using R-tool.
- Performance Comparison of Community detection algorithm(s) on CPU and GPU technology.

3.4 Research Methodology

In our research work we are using language R. It is a language used for statistical analysis and visualizations. Various packages are available in R for analyzing and

visualizing the twitter data. We need R studio along with R framework. Different R packages can be downloaded and installed using CRAN-mirror.

- To get hands on experience with R and R tool.
- To extract and pre-process tweets from twitter for further analysis.
- To get a twitter developers account.
- To authorize connection with Twitter.
- To develop the user interface and backend of research application using Shiny package.
- To perform Hierarchical cluster analysis on the extracted data.
- To implement community detection algorithms on the populated graphs.
- To perform comparison of Community detection algorithm(s) on CPU and GPU technology.

4. Implementation

The following section discusses the implementation details of the project and the underlying frontend and backend functions.

4.1 Twitter Hashtag

A Twitter hashtag is simply a keyword phrase, spelled out without spaces, with a pound sign (#) in front of it. For example, #modi and #India are both hashtags. We can categorize a particular tweet by using a relevant hashtag in that tweet. It helps in easy search and categorization of that tweet or related ones. If we click on a hash tagged word it will show you all the tweets related to that keyword. We can place hashtag anywhere in the tweet. The hashtags which expels on popularity are mostly related to trending topics.

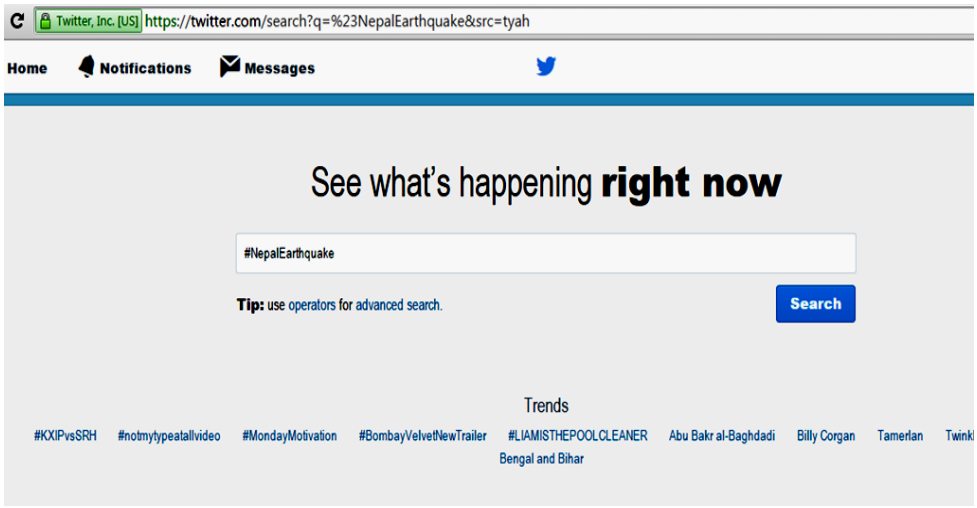


Fig. 23 : Twitter Search Bar

Using Twitter Hashtags

1. First decide on a keyword or a phrase, you want to search for and then visit https://twitter.com/search-home and enter your preferred hashtag in the search box. You will see a lot data being talked about related to your search query.
2. Pick Industry or Brand Keywords, so as to follow tweets of your favourite ones.

3. You can create your own hashtags too, the only thing to consider is to keep it short and simple.
4. Be Aware regarding the sentiments and the message you promote. It should not create a negative buzz and in return a messy situation for you.

4.2 Extraction of Data from Twitter

Data extraction is the most crucial task in any data analytics methodology. For extracting data from Twitter you have to first perform authentication for extracting tweets. For extracting data first you need to have a twitter account. If you don't have one sign up for new account.

- ❖ Getting a twitter developers account: first create a account at twitter developers page. Then you need to create a new application. For connecting with twitter account is needed. Details like your name, application description and your website are needed for creating an application.

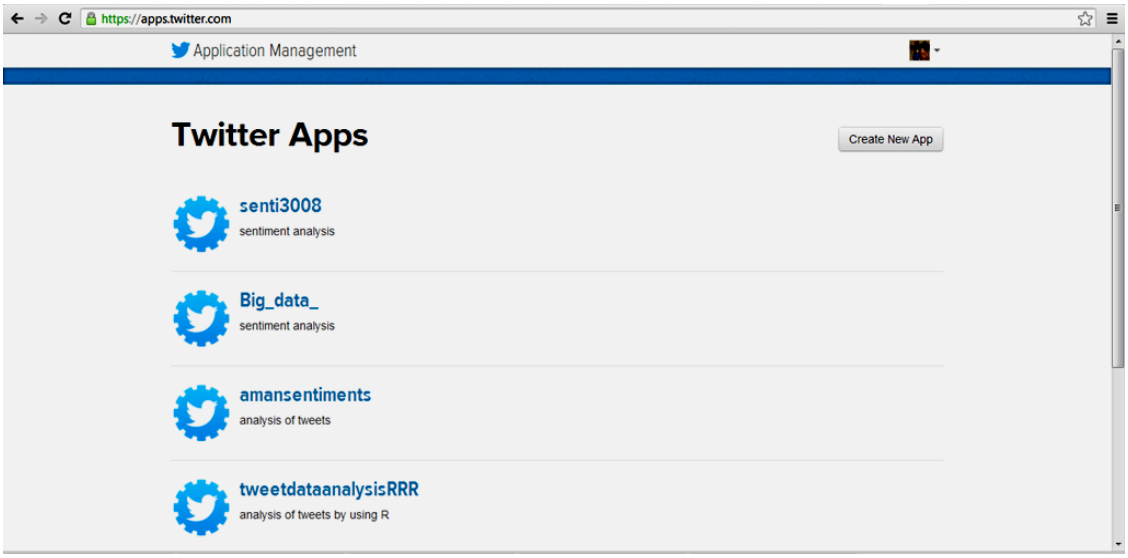


Fig. 24 : Click on Create New App

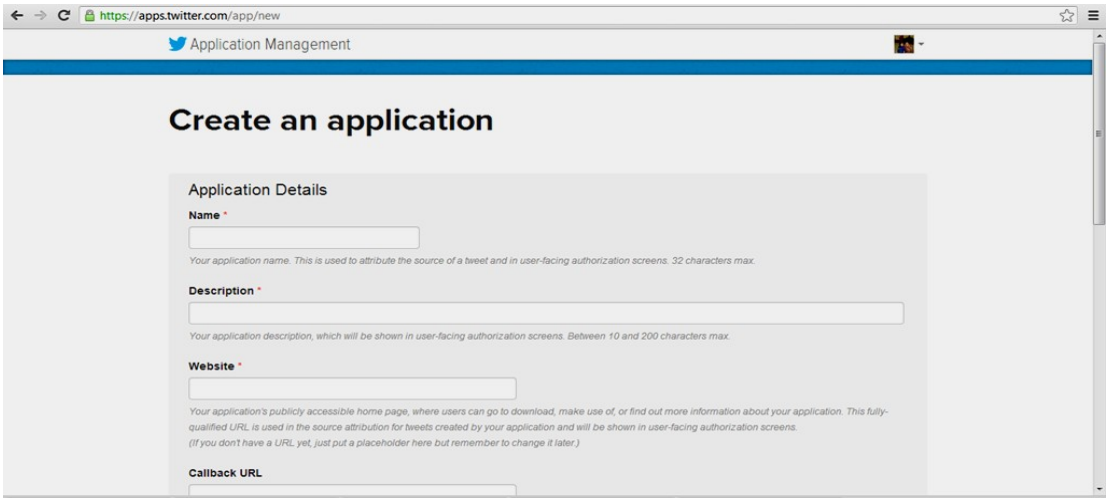


Fig. 25 : Application Details Form

You will be given consumer key and consumer token key. These keys are used for authentication purposes. Twitter make sure before providing access to its data that whether you are a registered user or not. Along with these two keys there are other two keys also known as access key and access token key. They also serve the same purpose as consumer keys. The only difference is in the Oauth call that is made through these keys. Some protocols use only consumer keys while some require all the four keys.

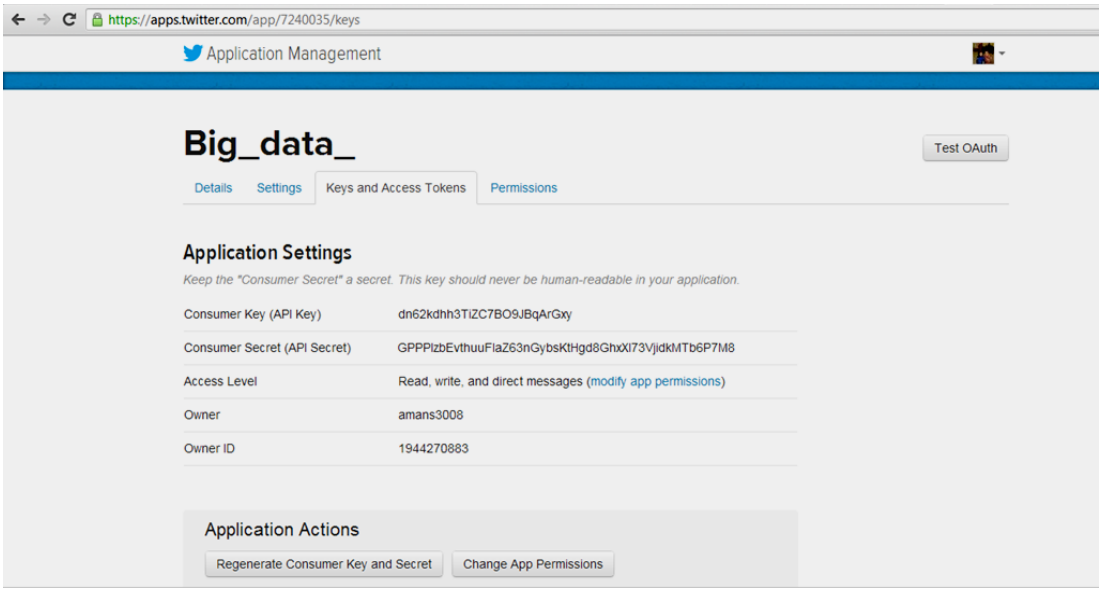


Fig. 26 : Consumer Keys

- Getting a curl certificate: Curl is a tool used to transfer to and from servers using any supported protocols like TELNET, HTTP, POP, SMTP, HTTPS, SMTPS and

many more. It also supports SSL connections, FTP transfer or upload etc. For downloading curl certificate open up your R console and start by loading the following libraries.

Getting a curl Certification

Open up your R console and start by loading the following libraries

```
rm(list=ls()) # Clear the previously used libraries

# Load the required R libraries
library(twitteR)
library(ROAuth)
library(RCurl)
```

Download the curl certificate and save it in the folder of your choice.

```
download.file(url="http://curl.haxx.se/ca/cacert.pem",destfile="cacert.pem")
```

- Setting up the certification for twitter: for this we need to provide the necessary parameters and call the oAuthFactory function for the authentication.

We will setup the necessary parameters and prepare the call to function `OAuthFactory` to get the authentication object.

```
# Set constant requestURL
requestURL <- "https://api.twitter.com/oauth/request_token"

# Set constant accessURL
accessURL <- "https://api.twitter.com/oauth/access_token"

# Set constant authURL
authURL <- "https://api.twitter.com/oauth/authorize"
```

Initialize consumer key and consumer secret variables with the values of access key and access token key you have generated and make a call to OAuthFactory function.

In the `consumerKey` field paste the access token you got for your twitter developer application.

```
consumerKey <- "xxxxxxxxxxxxxxxxx"
```

In the `consumerSecret` field paste the access token you got for your twitter developer application.

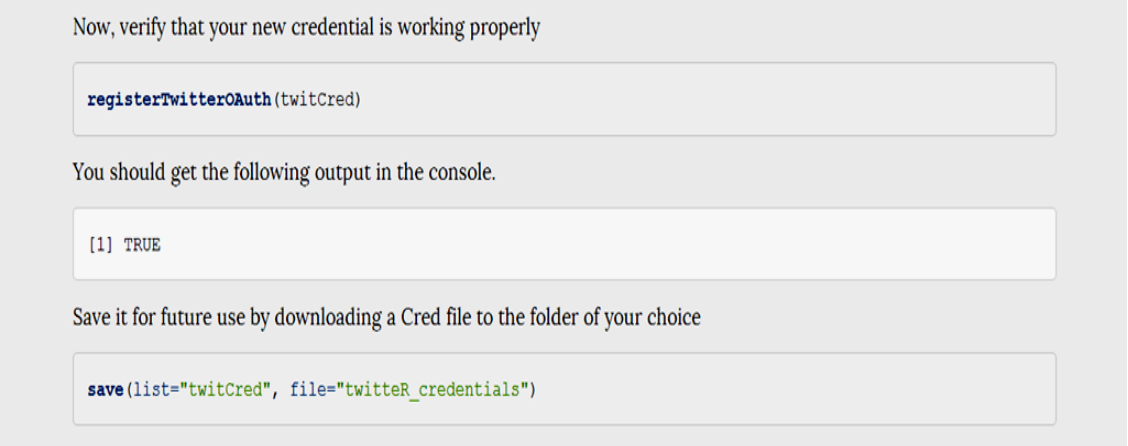
```
consumerSecret <- "xxxxxxxxxxxxxxxxx"
```

Now, create the authorization object by calling function `OAuthFactory`

```
twitCred <- OAuthFactory$new(consumerKey=consumerKey,
                             consumerSecret=consumerSecret,
                             requestURL=requestURL,
                             accessURL=accessURL,
                             authURL=authURL)
```

❖ Saving and using certification to connect to twitter: first ask for access using handshake function. Then fill the password provided by authentication of your application.

Now verify whether your connection is established or not and save it in a file.



4.3 To Design an Interface using Shiny Package

To visualize extracted data on the basis of specific hash tag (#tag). After establishing the connection the next step would be to design the frontend and backend of our application using shiny package. It is a package used for making web based applications using R. It provides easy application development. We have to create two files for any application, one is ui.R which contains the user interface design and the other one is server.R which contains all the backend functions of the application.For developing our application we have performed following steps:

- The first step would be to develop the UI of the application (Ui.R).User interface is kept quite simple. Four tabs have been provided which can give different view

to the data extracted from twitter. For search query two search boxes are provided to filter the relevant tweets. You can also scale up and down the number of tweets you wish to process.

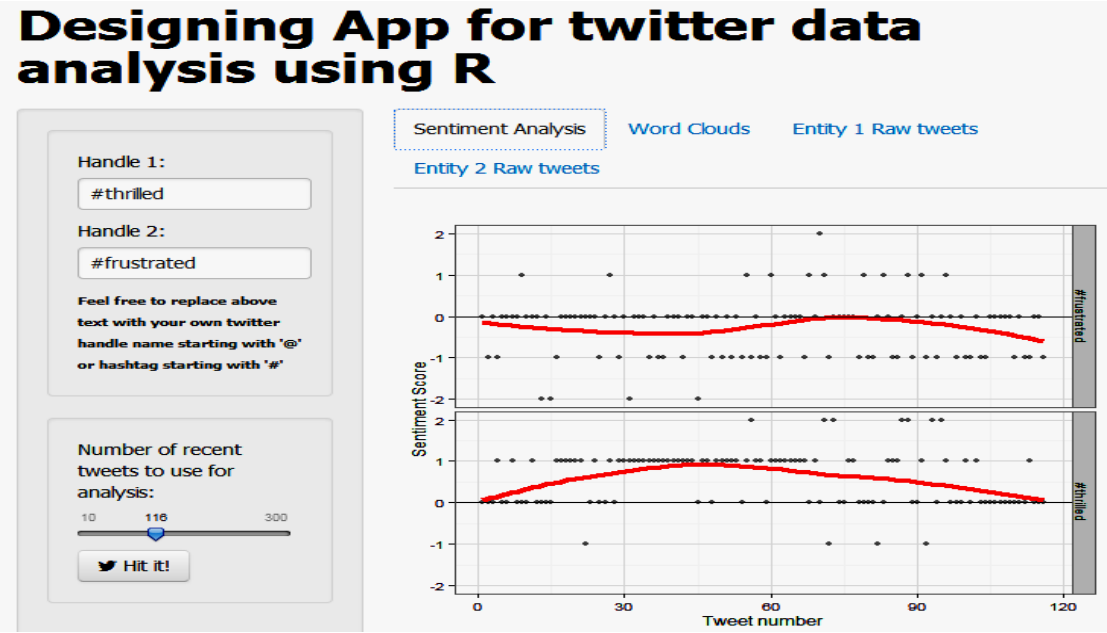


Fig. 27 : User Interface

- Second step will be to develop backend of application (server.R):for developing backend of the data we have created various functions that serve the various purposes for the application.
  - ❖ Tweet Frame function: This function is used to send request query to Twitter, and it fetches the tweets and cumulate them it into a data frame. Tweets are searched using this function. It includes call to built in function.

    searchTwitter (searchTerm,n=maxTweets,cainfo="cacert.pem",lang="en")

  - ❖ Clean Tweets function: This function pre-processes the fetched tweets and cleans them for further analysis. It makes a function call to build in function of TwitteR library.

    str_replace_all(tweets, "http://t.co/[a-z,A-Z,0-9]*{8}","")

  - ❖ Num of tweets function: It is used to calculate number of tweets fetched.

- Word cloud entity function: It is used to draw the word cloud of the fetched tweets. It uses following functions to create word cloud:

Fig. 28 : Wordcloud

```
tdMatrix <- as.matrix(tweetTDM) # creating a data matrix
sortedMatrix<-sort(rowSums(tdMatrix),decreasing=TRUE)#calculate row
sum of each term and sort in descending order (high freq to low)
```

- sentiment_analysis function: It is used to perform sentiment analysis of the fetched tweets and rate them according to the designed parameters. This is done with the help of Jeffrey Breen's algorithm.

- score.sentiment function: is a function used to calculate the sentiment score for the individual tweet based on the pre filtering performed in sentiment analysis function.

4.4 Streaming Data Visualization

Analysis (Mood variation) of continuous streaming data from twitter can be performed using streamR API of twitter. It extracts the tweets in a streaming mode, then further parses them and map them according to their sentiment points. Following steps are performed for carrying out real time analysis of tweets:

- Extracting data using twitter streaming API streamR: Continuous streams of data are fetched using streamR API. Parameters are provided to filter the tweets.

```
Tweets = filterStream(file.name = '', language = 'en', locations=c(-124, 23, -67, 50),
                      timeout = 5, oauth = my_oauth)
```

- Translating coordinates to states: It is carried out by using spacial mapping facilitated by sp package of R. It is based on algorithm that determines whether a point is in a complex polygon or not.

```
get_states = function (long, lati)
    {
     coordy = data.frame(cbind(long, lati))
    points_sp = SpatialPoints(coordy)
    proj4string (points_sp) = proj4string (states_sp)
    i = over (points_sp, states_sp)
    Names = sapply(states_sp@polygons, function(x)
      x@ID)
    Return (Names[i])
    }
```

- Calculating sentiment analysis and scores for the parsed tweets. This can be done using Jeffrey Breen algorithm. After calculating sentiment scores for individual tweet aggregation of sentiments based on individual states are also performed. After aggregation we also perform scaling of sentiment scores of states to map them properly. Scaling is done based on this formulae:

Sentiment of a overall state = sum of positive scores / sum of all absolute scores

Function to calculate sentiment:

```
get_sentiment = function (txt)
{
        Words = strsplit (txt, ' +')
        Words = unlist(words)
        pos_matches = match (words, pos)
        neg_matches = match (words, neg)
```

```
                    Score=sum (! is.na (pos_matches)) -
                          sum (!is.na(neg_matches))
                    Return (score)
}
```

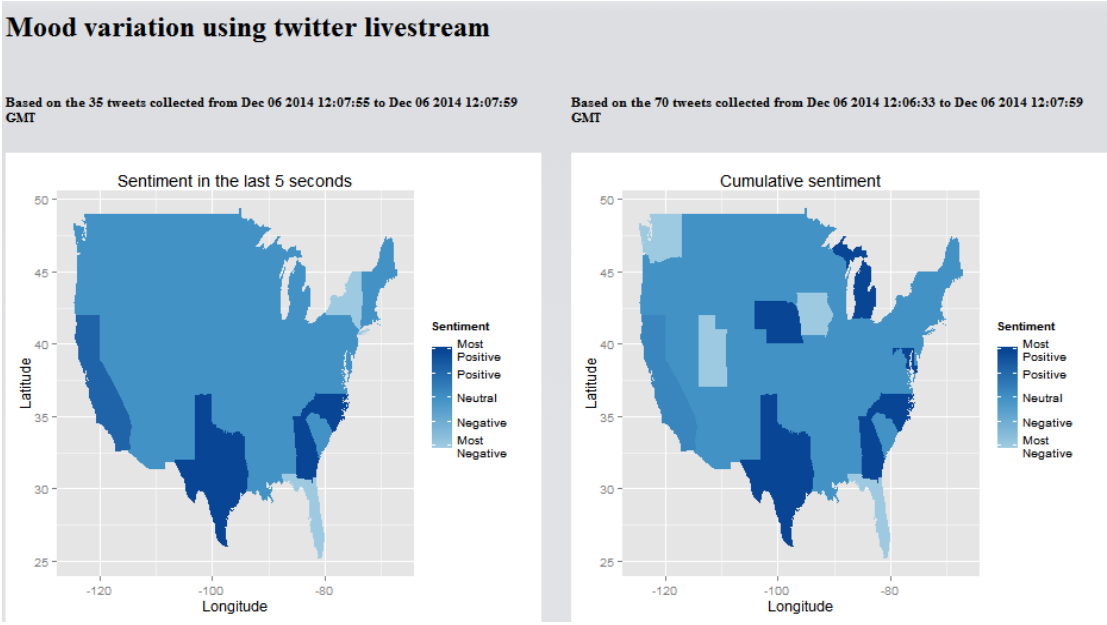


Fig. 29 : Mood variation Heat Map

4.5 Analyze #nepalearthquake Tweets (Twitter) with R

Here we are using R to explore the tweets with hashtag #Nepalearthquake. Following objectives are in consideration:

- Extract tweets with hashtag #Nepalearthquake
- Get some sense for themes using the following 2 methods:
  - ❖ Hierarchical cluster analysis
  - ❖ Community detection algorithms on graphs

4.5.1 Extract Data from Twitter

For extracting data from Twitter authentication has to be performed first performed. For extracting data first you need to have a twitter account. If you don't have one, sign Up for new account. For that following libraries are loaded:

- twitteR
- tm
- wordcloud

- dplyr
- topicmodels
- RColorBrewer
- igraph

4.5.2 Set up to Access Twitter Data

Before fetching data we need to establish connection using twitter OAuth library. Following steps are performed for the same:

- download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")
- cred <- OAuthFactory$new

```
(
    consumerKey=Key,
    consumerSecret=Secret,
    requestURL='https://api.twitter.com/oauth/request_token',
    accessURL='https://api.twitter.com/oauth/access_token',
    authURL='https://api.twitter.com/oauth/authorize'
)
```

- cred$handshake()
- registerTwitterOAuth(cred)

4.5.3 Get data for a Hashtag from Twitter

In order to pull the data from the twitter following steps have been used :

- First step is to get tweets with a hashtag, here specifically hashtag "Nepalearthquake" is used.

```
SearchTwitter ("#Nepalearthquake",n=5000,since="2015-04-26",until="2015-04-29",cainfo="cacert.pem")
```

- Text of each tweet is fetched to further analyze the extracted tweets.

```
sapply(getTw, function(x) x$getText())
```

- Get information regarding whether the tweet is a retweet or not.

```
sapply(getTw,function(x) x$getIsRetweet())
```

- After that get information regarding date of creation of tweet.

```
do.call(c,lapply(getTw,function(x) x$created))
```

4.5.4 Analyze Data

At this time we have raw tweets that are needed to be processed and cleaned before any further analysis. Following steps explain the complete pre-processing of raw tweets:

- First step is to remove data redundancy. Duplicate tweets are removed as they may occur due to occurrence of various retweets.

```
getTw_df_rmRT <- getTw_df%>%filter(!duplicated(txt)
```

- Remove non alphanumeric characters

```
getTw_txt_cln <- gsub("[^a-zA-Z0-9 ]","",getTw_df_rmRT$txt)
```

- Remove words starting with @

```
getTw_txt_cln <- gsub("@\\w+","",getTw_txt_cln)
```

- Remove words containing http or https.

```
getTw_txt_cln <- gsub("\\bhttp[a-zA-Z0-9]*\\b","",getTw_txt_cln)
```

After the pre-processing we will be left with the tweets with text part only which is easier to analyze. The below Fig. 30. gives the glimpse of few of such tweets.

```
[[751]]
[1] "PCIGlobal: We are partnering with organizations to deliver medical supplies and begin
the #NepalEarthquake rebuilding effort: http://t.co/obcCabf7T0"

[[752]]
[1] "IvanaMaGarcia: RT @Louis_Tomlinson: Just donated @savechildrenuk #NepalEarthquake app
eal. If you can too, go here http://t.co/NXBJftNl9m or text: DONATE5 â€¦"

[[753]]
[1] "eiyato_madao: RT @UN_News_Centre: UN humanitarian chief announces $15 million emergen
cy fund for #NepalEarthquake response http://t.co/Hru0cS72Fc http://â€¦"

[[754]]
[1] "Daniel_Brand: RT @iTunesTrailers: Hereâ€™s how you can help those affected by the #Ne
palEarthquake. Donate now: http://t.co/LfX5SuD1NO http://t.co/rFwTJa0Vâ€¦"

[[755]]
[1] "fcksmaliker: RT @Louis_Tomlinson: Just donated @savechildrenuk #NepalEarthquake appea
l. If you can too, go here http://t.co/NXBJftNl9m or text: DONATE5 â€¦"

[[756]]
[1] "NewCovenant_Net: RT @kathmandupost: Thousands of people have left the Capital and hea
ded for their homes after the quake. #NepalEarthquake #Nepal http://t.câ€¦"
```

Fig. 30 : Snapshot of Extracted Tweets

4.5.5 Hierarchical Clustering

Clustering techniques are most widely used statistical approach used for analysis and community detection purposes. Here discussion on hierarchical clustering technique is done with social data collected and pre-processed. Fig. 31. describes the cluster Dendrogram.

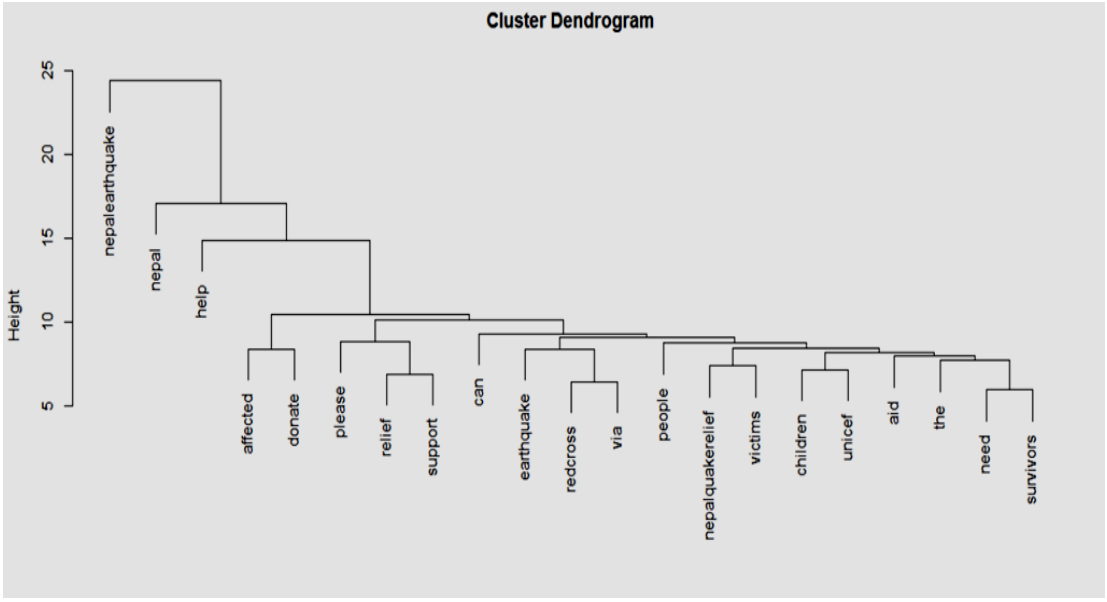


Fig. 31 : Cluster Dendrogram for Hash-Tag(#Nepalearthquake)

4.5.6 Graph Community Detection

Following steps are performed for finding communities using graph based approach:

- Create a graph for the data being analysed.

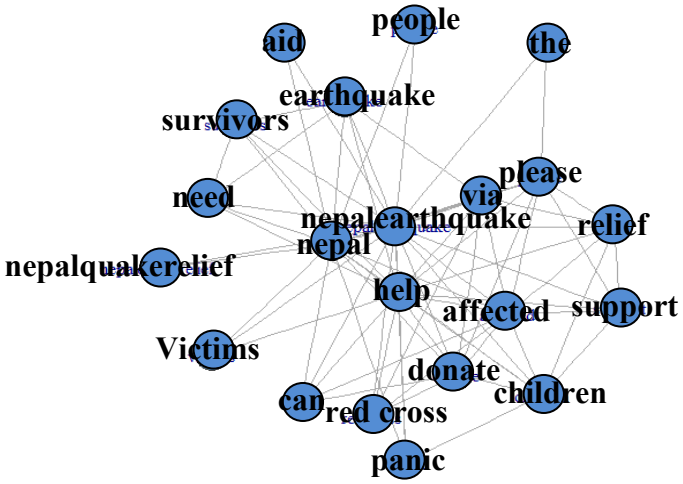


Fig. 32 : Social Graph for Tweets Hash-Tag(#Nepalearthquake)

- With R we are using igraph package to analyze the graph and find out the communities. Various community detection algorithms are available with igraph package. Following are the few of the algorithms:
    a. label.propagation community
    b. leading.eigenvector community
    c. edge.betweenness community
    d. fastgreedy community
    e. walktrap community

4.6 Comparative Analysis CPU vs. GPU using Social Data Set

Statistical tasks are computational intensive. All the routine statistical tasks such as data extraction, graphical summary and technical interpretation use of modern computing machinery. Parallel computing environment can greatly help these tasks to be performed simultaneously. A recent advance in computer hardware makes parallel computing capability widely available to most users. Video graphics cards are available nowadays to support parallel computing operations besides the routine graphical functions. Applications or statistical processes that use these graphics processing units (GPU) have reportedly shown performance gains. It is evident that parallel computing will be of tremendous importance in the near future. Earlier many tasks seem to be impossible due to resource constraint but now it seems to be a one click away task using GPU technology. We have selected one of the most affordable

options available is NVIDIA's CUDA. We will not deal with CUDA directly rather we will be exploiting rpud package for studying GPU computing. We will be comparing performance of GPU statistical functions with regular r statistics.

4.6.1 Distance Metric

Measuring distance for statistical analysis is a routine activity but we hardly bother upon its computational speed. This statistical quantity describes the dissimilarity between the data to be analysed. The Euclidean distance is most popularly used distance metric for populating the sample data set. It is the square root of sum of squares of attribute differences. For two data points x and y with n numerical attributes, the Euclidean distance between them is given by:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Query 1. We have computed the distance using the data matrix generated using twitter data using regular R function and rpud function. We are using vector length to be 1500.

```
system.time(dist(m)) #using simple r function
```

User system elapsed
42.2 0.03 48.14

```
Library (rpud)      #using GPU function
system.time(rpuDist(m))
```

User system elapsed
1.98 0.494 3.24

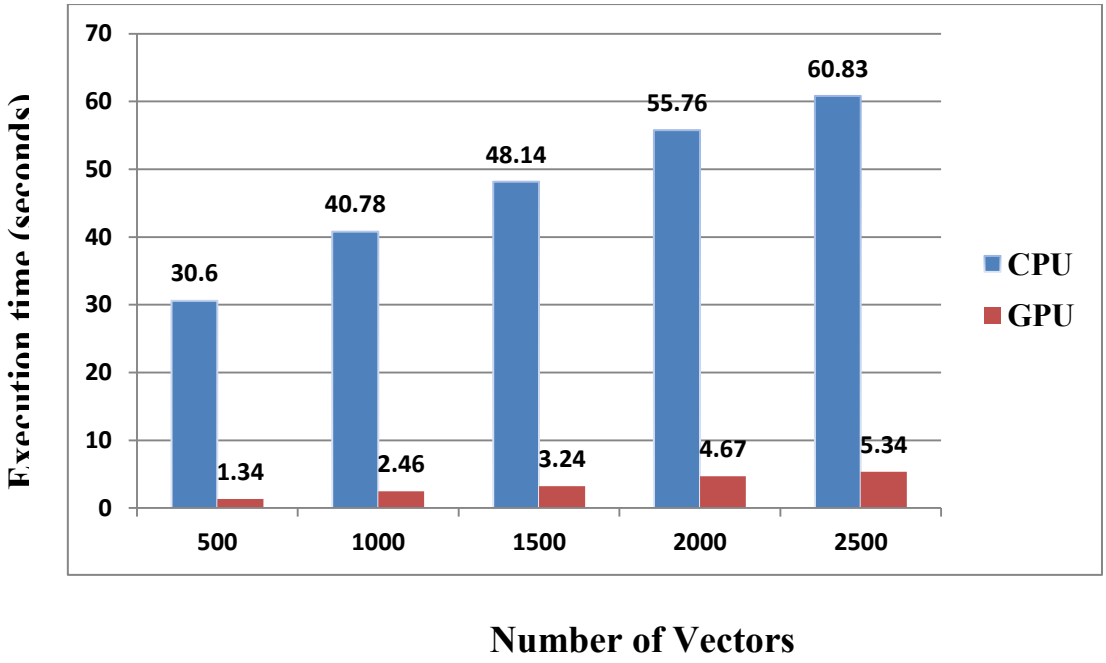


Fig. 33 : Comparison Graph for Distance Metric GPU vs. CPU

4.5.2 Hierarchical Cluster Analysis

As we have already discussed distance as a statistical metric for data analysis, one of its prominent uses is in cluster analysis for relationship discovery. For example in the data set of social data, like tweets, posts we can calculate the distance matrix with hclust, and a dendrogram can be plotted as described above in Fig. 31. that displays a hierarchical relationship among the tweets or posts.

```
d <- dist (as.matrix(m))   # find distance matrix

Var=hclust(d)     #apply hierarchical clustering

Plot (Var)          # plot the dendrogram

system.time(hclust(d))  #using simple r function
```

User system elapsed		
115.76	0.087	115.914

Query 1. We have computed the distance using the data matrix generated using twitter data using regular R function and rpud function. We are using vector length to be 1500.

library (rpud)

system.time(rpuHclust(d))

User	system	elapsed
0.792	0.104	0.896

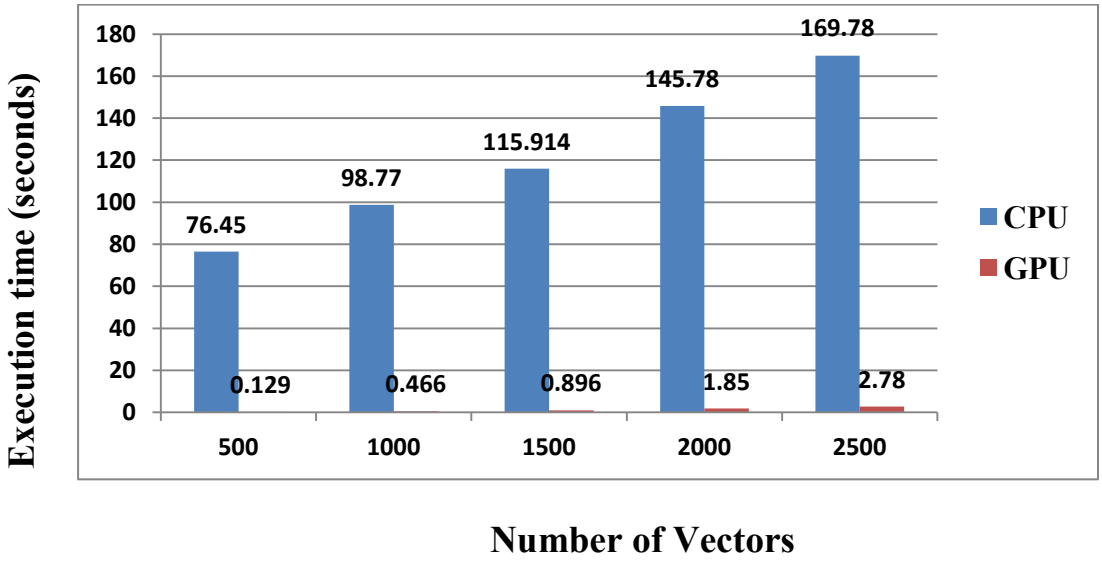


Fig. 33 : Comparison Graph for Hierarchal Clustering GPU vs. CPU

Chapter 5 Conclusion and Future Scope

Conclusion

In this thesis an attempt has been made to design a web based application for analysis and visualisation of social data. For this analysis twitter REST APIs are used to extract the data from Twitter site. The language used for designing the app is R. As R is a powerful tool for visualisation and analysis of data. Major features of R for analysis as well as visualization has been exploited. As an outcome an integration of both the aspects of analysis is evolved, as a web application which can further be upgraded to cloud. Moreover a survey on various IoT solutions available for 3-D visualisation of Twitter data is conducted. The various arduino boards available in the market for providing IoT solutions has been outlined.

Data visualisation is mapping of numbers or text to visual forms, something that is hardcore need of SNA. 3-D data visualization extends visual possibilities for data analysis and data science. It is an attempt to enrich IoT market with smart devices and even help the social data analyst to dig the new visualisation paradigm. The main purpose of this survey is to boost new and innovative IoT solutions using social analysis as a key research background. Finally the comparative analysis of GPU vs. CPU techniques for some of the statistical metrics and algorithms used for community detection is conducted over the extracted and populated social data set.

Future Scope

- As a future work this analysis can be extended to streaming data analysis using twitter stream API.
- The benefits of GPU technology can be leveraged to improve the analysis experience to many folds.
- GPU and R combination can be extended using CUDA programming in statistical tasks.

References

[1] R. Cuevas, R. Gonzalez, A. Cuevas, C. Guerrero, " Understanding the Locality effect in Twitter: Measurement and Analysis", Personal and Ubiquitous Computing Journal, Vol. 18, No. 2, pp. 397-411, 2014.

[2] Twitter, blog. [Online]. Available: http://blog.twitter.com/2012/03/twitter-turns-six.html

[3] Twitter [Online]. Available: http://www.twitter.com

[4] Techcrunch. [Online]. Available: http://techcrunch.com/2012/07/30/analyst-twitterpassed-500m-user-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweetingcity/

[5] M. Cha., H. Haddadi and Gummadi P. K., " Measuring user Influence in Twitter: the million follower fallacy", in Proceedings of International Conference on Web and Social Media, pp. 1010-1017,2012

[6] Twitter, Stream API. [Online]. Available: https://dev.twitter.com/docs/streaming-apis.

[7] A. Clauset, M. E .J. Newman and C. Moore, "Finding Community Structure in very large Networks", Physical Review E, No. 70, 2004.

[8] L. Feng, James, Z. Wang and E. Promislow, "Exploring Local Community Structures in Large Networks", Web Intelligence and Agent Systems, Vol. 6, No. 4, pp. 387-400, 2008.

[9] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society", Nature, Vol. 435, No. 7043, pp. 814–818, 2005.

[10] J. Chen, O. R. Zaiane and R. Goebel, “Overlapping Community Detection in Networks", ACM Computing Surveys (CSUR), Vol. 45, No. 43, pp. 338–343, 2013.

[11] S. Gregory, "Finding Overlapping Communities in Networks by Label Propagation", New Journal of Physics, Vol. 12, No. 10, pp. 1-26, 2010.

[12] X. Xu, N. Y. Z. Feng, T. A. J. Schweiger, "Scan: A Structural Clustering Algorithm for Networks", in Proceedings of SIGKDD, pp. 824–833, 2007.

[13] J. Scripps, P. N. Tan, and A. H. Esfahanian, "Node Roles and Community Structure in Networks ", in Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social network Analysis, pp. 26–35, 2007.

[14] S. Fortunato "Community Detection in Graphs", Physics Reports, Vol. 486, No. 3, pp. 75-174, 2010.

[15] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping Community Detection in Networks: the state of the art and comparative study ", ACM Computing Surveys, Vol. 45, pp. 45-49, 2013.

[16] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community Detection in social media", Data Mining and Knowledge Discovery, Vol. 24, No. 3, pp. 515–554, 2012.

[17] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Dening and Identifying Communities in Networks", in Proceedings of the National Academy of Sciences, Vol. 101, No. 9, pp. 2658-2663, 2004.

[18] J. P. Onnela, J. Saramki, J. Kertsz, and K. Kaski, "Intensity and Coherence of motifs in Weighted Complex Networks", Physical Review, Vol. 71, No. 6, pp. 065103, 2005.

[19] M. Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks", PNAS, Vol. 99, No.12, pp. 7821–7826, 2002.

[20] M. E. J. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks", Physical Review E, Vol. 69, No. 2, pp. 1-16, 2004.

[21] J. Shi and J. Malik, "Normalized cuts and Img. Segmentation," IEEE Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp. 888-905, 2000.

[22] M. E. Newman, "Modularity and Community Structure in Networks," Proceedings of the National Academy of Sciences, Vol. 103, No. 23, pp. 8577–8582, 2006.

[23] A. Clauset., "Finding Local Community Structure in Networks," Physical Review, Vol. 72, No. 2, pp. 026132, 2005.

[24] I. J. Farkas, D. Abel, G. Palla and T. Vicsek, "Weighted Network modules," New Journal of Physics, Vol. 9, No. 6, pp. 180, 2007.

[25] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for Image Annotation", International Journal of Computer Vision, Vol. 90, No. 1, pp. 88–105, 2010.

[26] C. P. Massen and J. P. K. Doye, "Identifying Communities within Energy Landscapes", Physical Review, Vol. 71, No. 4, pp. 1-13, 2005.

[27] M. E. J. Newman, "Fast algorithm for Detecting Community structure in Networks", Physical Review, vol. 69, No. 6, pp. 1-23. 2004.

[28] M. E. J. Newman "Finding Community Structure in Networks using the Eigen Vectors of Matrices", Physical Review, vol. 74, No. 3, pp. 1-14, 2006.

[29] E. Giannakidou, V. A. Koutsonikola, A. Vakali and Y. Kompatsiaris, "Co-clustering tags and Social data sources", in Proceedings of Web-Age Information Management, pp. 1-23, 2008.

[30] H. Shen, X. Cheng, K. Cai, and M. B. Hu, "Detect Overlapping and Hierarchical community structure in Networks", Physical A: Statistical Mechanics and its Applications, Vol. 388, No. 8, pp. 1706-1712, 2008.

[31] A. Lancichinetti, S. Fortunato, and J. Kertsz, "Detecting the overlapping and hierarchical community structure in complex networks", New Journal of Physics, Vol. 11, No. 3, pp. 1-15, 2009.

[32] C. Lee, F. Reid, A. McDaid and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion", arXiv preprint arXiv, 2010.

[33] A. Lancichinetti, F. Radicchi, J.J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks", PLoS ONE, Vol. 6, No. 4, pp. 1-18, 2011.

[34] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann "Link communities reveal multiscale complexity in networks", Nature, Vol. 466, No. 7307, pp. 761–764, 2010.

[35] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection", Data Mining (ICDM), IEEE 12th International Conference, pp. 1170–1175, 2012.

[36] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach", in Proceedings of the sixth ACM international conference on Web search and data mining, pp. 587–596, 2013.

[37] Twitter Mood Light, [Online]. Available: http://www.instructables.com/id/Twitter-Mood-Light-The-Worlds-Mood-in-a-Box/

[38] Tasty Tweets, [Online]. Available http://duino4projects.com/tasty-tweets/

[39] Twitter Enabled Coffee Pot, [Online]. Available: http://www.instructables.com/id/Tweet-a-Pot-Twitter-Enabled-Coffee-Pot/

[40] Twitter theft control, [Online]. Available: http://www.instructables.com/id/Touch-Me-NotWithout-Ethernet-Sheild/

List of Publications

[1] Aman Sharma and Rinkle Rani, "A 3-level Model for implementing MOOC in India", in Proceedings of IEEE 2nd International Conference on MOOCs, Innovation and Technology in Education, pp. 132-137, 19-20 December, 2014.

[2] Aman Sharma and Rinkle Rani, "Analysis and Visualization of Twitter Data using R", in Proceedings of IEEE sponsored International_Conference ON Technology and Management INBUSH ERA , 25-27 February, 2015.

[3] Aman Sharma and Rinkle Rani, "IoT Solutions for 3-D visualization of Twitter Data", in Proceedings of 5th IEEE International Advanced Computing Conference (IACC 2015), 12-13 June, 2015.

Video Presentation

Video link : https://www.youtube.com/watch?v=VSvHXKnq_mE

Reflective Diary

December,2014

Day after the final exams of EST I had the presentation for the Research paper on MOOC's. It was my first experience and first paper presentation for any research conference. But at last presentation went pretty well and it was a worth remembering experience. After that I realised that how difficult it was for me to write my first research paper. There was a huge difference between the final and the first version of my paper. After many iterations and modifications by my guide it got its final worth as a research paper for submission. But thanks to her for immense patience and devoting her time with me. After a gap of two or three days I started exploring further about my research topic i.e. (SNA) and hence came across various visualisation techniques for data analysis. I was keen to know about "3-D visualisation" and read about it in details.

January,2015

In the first week of the month I completed my 2nd research paper for the conference in Noida. Now writing this paper was not that difficult. At last ! Yippee !! I got the paper writing skill. I have read many research papers on "Big Data" and hence got to know about "IOT" (Internet of Things). So I was fascinated about it and hence registered for the workshop in India on IoT. It was fun and informative experience. they made us work on arduino boards, hence I got to know about the IoT solutions for the twitter and social analysis.

February,2015

In this month I explored more about social data, various platforms for data extraction. I read specifically about community structure and its attributes. Any "Data Science" process undergoes following sub tasks majorly:

1. Data Acquisition(Extraction)
2. Data Processing(Filtering)
3. Data Analysis
4. Data Visualisation

So it is a complete chain of processes that are needed to be carried out for any Data Science process. Even I read book of O' Riley " Mining the Social Web" Mean while I also started Writing my 3rd research paper on ("IoT Solutions for 3-D visualisation") for a IEEE conference at Bangalore.

March,2015

This month starts with the presentation at the IEEE Conference, Noida. I went there for presentation. It went good and even conference chair applauded me for the presentation. I read few research papers. Few of them are:

" Big Graph Mining for web and Social Media: Algorithms, Anomaly Detection and Application"

"Community detection: effective evaluation on large social networks"
and even introduced to new term GPU for data Science using R

R+GPU=HPC

April,2015

NVIDIA supports GPU technology hence helps to parallelize and speed up the statistical tasks. I earlier used R for making my project in previous semester but I was not aware about the R and GPU combination. R itself doesn't support parallelism for statistical tasks. Other problem with R is that it implements different functions as a separate instance on multi-core on clustered hardware.
So R and GPU are quite interesting match and NVIDIA supports it and there is still to explore and innovate. I implemented Community detection algorithm and metric on GPU to compare it with CPU over the extracted data.

May,2015

This is the Thesis summation month, we have to sum up all the research on implementation tasks for final thesis submission. So I fulfilled the following objectives for my thesis and hence started with my thesis writing.
1. To study various techniques and tools available for data analytics and visualisation.

2. To develop a research based application using Twitter and R-tool to analyze and visualize social data.

3. Execution of community detection algorithms on extracted Twitter data using R-tool.

4. Performance Comparison of these algorithms on GPU and CPU.